



October 28 - November 1, 2016



The Twelfth Conference of
The Association for Machine Translation
in the Americas

<http://www.amtaweb.org/amta-2016-in-austin-tx>

**VOLUME 2:
MT Users' Track**

Editors:

**Olga Beregovaya, Jennifer Doyon,
Lucie Langlois & Steve Richardson**

AMTA 2016

October 28 – November 1, 2016 -- Austin, TX, USA

Proceedings of
AMTA 2016,
Vol. 2:
Commercial MT Users and Translators Track
Government MT Users Track

Olga Beregovaya, Jennifer Doyon,
Lucie Langlois & Steve Richardson, Eds.



Association for Machine Translation in the Americas

<http://www.amtaweb.org>

©2016 The Authors. These articles are licensed under a Creative Commons 3.0 license, no derivative works, attribution, CC-BY-ND.

Introduction - Commercial MT Users and Translators Track

The Commercial MT Users and Translators Track at AMTA 2016 features presentations from organizations, enterprises, and individuals in the translation and language technology industry worldwide, including Language Service Providers, commercial machine translation technology and service providers, and a wide range of machine translation practitioners. This year's presentations reflect the evolving variety of MT applications beyond traditional post-editing scenarios, such as Internet search, pseudo-localization, and knowledge base article translation. Industry experts and MT enthusiasts will cover topics relevant to everyone involved in the rapidly growing adoption of MT across the industry spectrum, including:

- techniques and tools for more effective post-editing,
- automatic and immediate adaptation of MT systems based on user feedback,
- quality assessment of post-edited content using industry-standard metrics,
- analysis of MT acceptance by professional translators,
- development of machine translation engines for less common languages,
- effective integration of translation memories with MT technology, and
- applications of speech translation.

Presentations on these and other topics will document the rapidly expanding acceptance and integration of machine translation technology, demonstrating its essential role in successful localization and translation efforts around the world.

The Commercial MT Users and Translators Track Co-Chairs

Steve Richardson

Olga Beregovaya

Contents

Commercial MT Users and Translators

- 1 MT crowdsource at Yandex
Irina Galinskaya and Farhat Aminov
- 3 MT Post-Editing in a cloud based environment
Jean-Luc Saillard
- 19 MT Adaptation from TMs in ModernMT
Marcello Federico
- 58 Web App UI Layout Sniffer
Raymond Peng and Xin Jing Hu
- 65 Multilingual Search with Machine Translation in the Intel Communities
Ryan Martin
- 72 MT Thresholding: Achieving a defined quality bar with a mix of human and machine translation
Dag Schmidtke
- 82 Machine Translation Acceptance Among Professional Linguists: Are We Nearing the Tipping Point?
Yves Champollion
- 86 What Can We Really Learn from Post-editing?
Marcis Pinnis, Rihards Kalnins, Raivis Skadins and Inguna Skadina
- 92 An Empirical Study: Post-editing Effort for English to Arabic Hybrid Machine Translation
Hassan Sajjad, Francisco Guzman and Stephan Vogel

- 114 Divide and Conquer Strategy for Large Data MT
Dimitar Shterionov
- 123 The Reasonable Effectiveness of Data
Achim Ruopp
- 143 MT for Uralic Languages: Yandex Approach
Irina Galinskaya and Alexey Baytin
- 145 Seamlessly integrating machine translation into existing translation processes (STAR MT and Transit NXT)
Nadira Hofmann
- 170 Building a Translation Memory to Improve Machine Translation Coverage and Quality
Duncan Gillespie and Benjamin Russell
- 179 Enhancing a Production TM-MT Environment Using a Quotation TM
Hitokazu Matsushita and Steve Richardson
- 193 Improving Machine Translation for Post-Editing via Real Time Adaptation
Dragos Munteanu
- 222 Improving KantanMT Training Efficiency with fast_align
Dimitar Shterionov, Jinhua Du, Marc Anthony Palminteri, Laura Casanellas, Tony O'Dowd and Andy Way
- 232 Speech translation user experience in practice
Chris Wendt, Will Lewis and Tanvi Surti
- 240 Evaluation of machine translation quality in e-commerce environment
Maxim Khalilov

263 I Ate Too Much Cake: Beyond Domain-Specific MT Engines

Alex Yanishevsky

286 What? Why? How? - Factors that impact the success of commercial MT projects

John Tinsley

Government MT Users

304 Assessing Translation Quality Metrics

Jennifer DeCamp

322 Machine Translation for English Retrieval of Information in Any Language (Machine translation for English-based domain-appropriate triage of information in any language)

Carl Rubino

355 A Taxonomy of Weeds: A Field Guide for Corpus Curators to Winnowing the Parallel Text Harvest

Katherine M. Young, Jeremy Gwinnup and Lane O.B. Schwartz

371 Toward Temporally-aware MT: Can Information Extraction Help Preserve Temporal Interpretation?

Taylor Cassidy, Jamal Laoudi and Clare Voss

385 "Did You Mean...?" and Dictionary Repair: from Science to Engineering

Michael Maxwell and Petra Bradley

412 Principle-Based Preparation of Authentic Bilingual Text Resources

Michelle Vanni

422 Machine Translation of Canadian Court Decisions

Lucie Langlois, Michel Simard and Elliott Macklovitch

- 453 Putting the "human" back in HLT: The importance of human evaluation in assessing the quality and potential uses of translation technology
Erica Michael, Petra Bradley, Paul McNamee and Matt Post
- 551 Proto-MT Evaluation for Humanitarian Assistance Disaster Response Scenarios
Douglas Jones
- 575 Wearable Devices to Enable Communication via ASL (Sign Language Translation)
Patricia O'Neill-Brown and Nicolas Malyska
- 595 Tuning for Neural Machine Translation
Guido Zarrella
- 622 Invisible MT
Patricia O'Neill-Brown
- 654 MoJo: Bringing Hybrid MT to the Center for Applied Machine Translation
Marianna Martindale
- 715 Building Renewable Language Assets in Government Domains
Beth Flaherty and Joshua Johanson

MT crowdsource at Yandex

Irina Galinskaya
Farhat Aminov
Yandex, LLC

galinskaya@yandex-team.ru
aminov@yandex-team.ru

1. Abstract

Yandex.Translate is a popular online service with a various translation scenarios and a great opportunity for crowdsourcing. Some crowdsource systems invent artificial tasks and enforce volunteers to accomplish them. We instead encourage ordinary users to run their own tasks and thus collect valuable data from natural activities in different translation scenarios. In this presentation we are going to outline the following cases:

Scenario 1: users of Yandex.Translate correct machine translation of their own text.

It is widely known, that users are not always satisfied with the results provided by machine translation systems, especially when translating texts into foreign languages. One of such cases is correspondence translation. As a result, the users unsatisfied with the translation quality copy the machine translated texts into mail clients or messengers and manually edit those translations. Only after these steps are accomplished the users would be ready to send their message to partners, clients or friends. Thus, the user edits, which represent very valuable information about incorrect translations, are carried out in third party apps and not on the translation service itself. A while back, we implemented a feature that lets our users edit translations in place, without leaving Yandex.Translate service. The feature turned out to be highly demanded and we were able to collect about 100K edits in a span of three months.

We conducted a quantitative and qualitative analysis of the gathered data and explored a possibility of using this data to improve machine translation quality.

Scenario 2: users of Yandex.Translate report errors in machine dictionaries.

In addition to the full text machine translation system, we develop machine dictionaries, both translation and monolingual. They help our users better understand complicated translations or find an appropriate replacement (e.g. synonyms) for a typed word. Evidently, information from the machine dictionaries, as well as machine translated texts, contains errors. Based on the assumption that our users would be actively engaged in improving the dictionaries, we developed a feature to easily report a wide range of dictionary related issues, such as wrong translations, capitalization errors, wrong linguistic attributes, etc. Further, we created a moderation section for professional linguists to conveniently review the reported issues. Such a two-step error reporting and verification system proved to be very effective and allowed us to significantly improve the quality of the machine dictionaries. We believe the described feature will evolve in the nearest future and will be used not only to report issues, but also to add new information to the dictionaries. To demonstrate how machine dictionaries could benefit from crowdsourcing, we analyzed the received data and accepted user contributions in various ways.

Scenario 3: Wikimedia contributors are making new articles for Russian Wikipedia by translating articles from English Wikipedia with the help of Yandex.Translate API.

Wikipedia's content translation tool has been available since the beginning of 2014 and allows contributors to translate articles into variety of languages. A year after the tool was released Yandex's machine translation technology was integrated to help contributors create new articles. Initially the feature was introduced to Russian speaking users only, but eventually it proved to be very useful and became available for numerous other languages. Recently, Wikimedia Foundation published an API that lets anyone access the post-editing data, which was produced by Wikipedia users. Based on this data, we built distributions of languages, lengths of edited fragments, edit distances between machine translated sentences and and final results published by human editors (post-edited texts). We also analyzed a possibility of using this data for training and testing of machine translation systems.

MT Post-Editing in a cloud based environment

PEMT Landscape

- PEMT Consumers
- Challenges
- Standard Solutions
- The “cloud-based” advantages
- Pushing the envelope!

PEMT Consumers

- eCommerce
- SaaS Providers
- Hardware Manufacturers
- Software Developers
- News
- Governmental Organizations
- And more...

eCommerce Challenge

Translation has an impact!

Studies have shown that customers are 6x times more like to shop from sites that are in their native languages*



* CRACKER and LT_Observatory

Opening New Opportunities **SmartCAT™**

eCommerce Challenge

Volume of content makes the standard translation process unaffordable!

- High number of languages needed
- The leading eCommerce reseller lists over 300 million products
- Goods and services have a limited shelf life
- Translation is another overhead whose cost has to be limited

Standard Solutions

Decentralized model based around desktop applications

- Files are processed with MT in batches / no possibility for feedback.
- Linguistic assets such as TM and glossaries are not shared and must be updated manually.
- Communication between translators is nearly impossible.

And it gets worse!

**Project management is the biggest challenge.
The main issues are:**

- File management
- Progress reports
- Communication and collaboration
- Accounting

The cloud to the rescue

The main advantages are:

- Centralized language assets
- Real-time progress reports
- Real-time communication and improved collaboration features
- Deep integration with MT engines with the possibility of feedback for retraining purposes

The cloud to the rescue

Centralized language assets

- Translation memories are shared between all project participants and updated in real-time
- Glossaries can be enriched as needed and their content shared with the MT provider
- LQA rules are applied uniformly

The cloud to the rescue

Project management made easy

- Translation progress is updated in real-time without the need to exchange emails
- Editing tasks can be assigned easily to all participants without a single file being exchanged
- Communication is greatly simplified with chat modules and real-time commenting in the editor

The cloud to the rescue

A greatly improved MT process

- Post-editing corrections can be fed back to the MT provider in real-time
- Terminology additions can also be used for engine improvements
- Detailed productivity statistics can be generated at any time to measure MT quality and the results of training

Pushing the envelope!

How does SmartCAT improve on the model:

- Our marketplace gives project manager access to more than 70,000 possible post-editors
- Multi-mode task assignment features – splitting documents and tasks is automatic
- Payment process is also automated, accounting tasks are greatly simplified

Pushing the envelope!

A better post-editing environment:

- Our multi-user editor allows large numbers of post-editors to work in the same document at different stages of the process
- A sophisticated locking mechanism keeps everybody in their own section while promoting communication
- Change tracking can be used to help improve MT output

Pushing the envelope!

Sample cases:

- Coursera classes translated into other languages using 100's of translators working together at post-editing
- A review and grading system was put in place to help editors select the best possible proposed translation
- The interface was customized to display video classes matching the subtitles being translated

Pushing the envelope!

Sample cases:

- eCommerce website requiring large volume of translation
- Up to 40 translators working at the same time in the same document to post-edit over 100,000 words in a day without interfering with each other's work
- Post-editors can work at their own pace on their own schedule in the section assigned to them

SmartCAT

Thanks for Watching!

2016

SmartCAT gives business wings for global expansion. The first Translation Automation platform powered by over 60 000 language professionals worldwide.

MT adaptation from TMs in ModernMT

Marcello Federico - FBK, Italy

AMTA, Oct 29 2016 - Austin, Texas

ModernMT Next Generation
Machine Translation



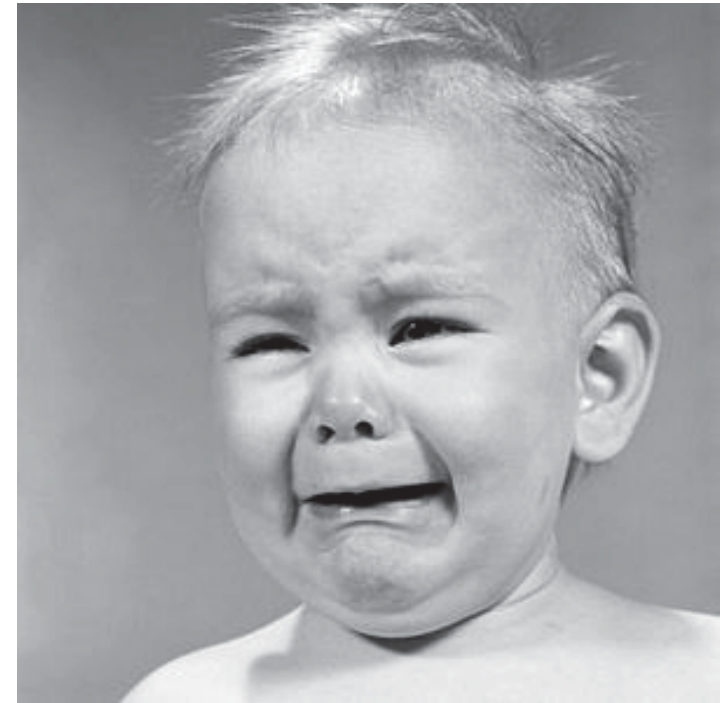
Translators' pains with MT

output is often poor or contextually wrong



LSP engineers don't laugh either

cumbersome setup of MT
lack of training data
online MT is too generic



The Modern MT way

- (1) connect your CAT with a key
- (2) drag & drop your private TMs
- (3) start translating!



Modern MT in a nutshell

zero training time
manages context
learns from users
scales with data and users



Team

Business



Research



ModernMT Next Generation
Machine Translation

Roadmap

2015 Q1



development
started

2016 Q2



first **alpha**
release.

10 langs,
fast training,
context aware,
distributed

2016 Q4



first **beta**
release

45 langs,
Incremental
learning

2017 Q4



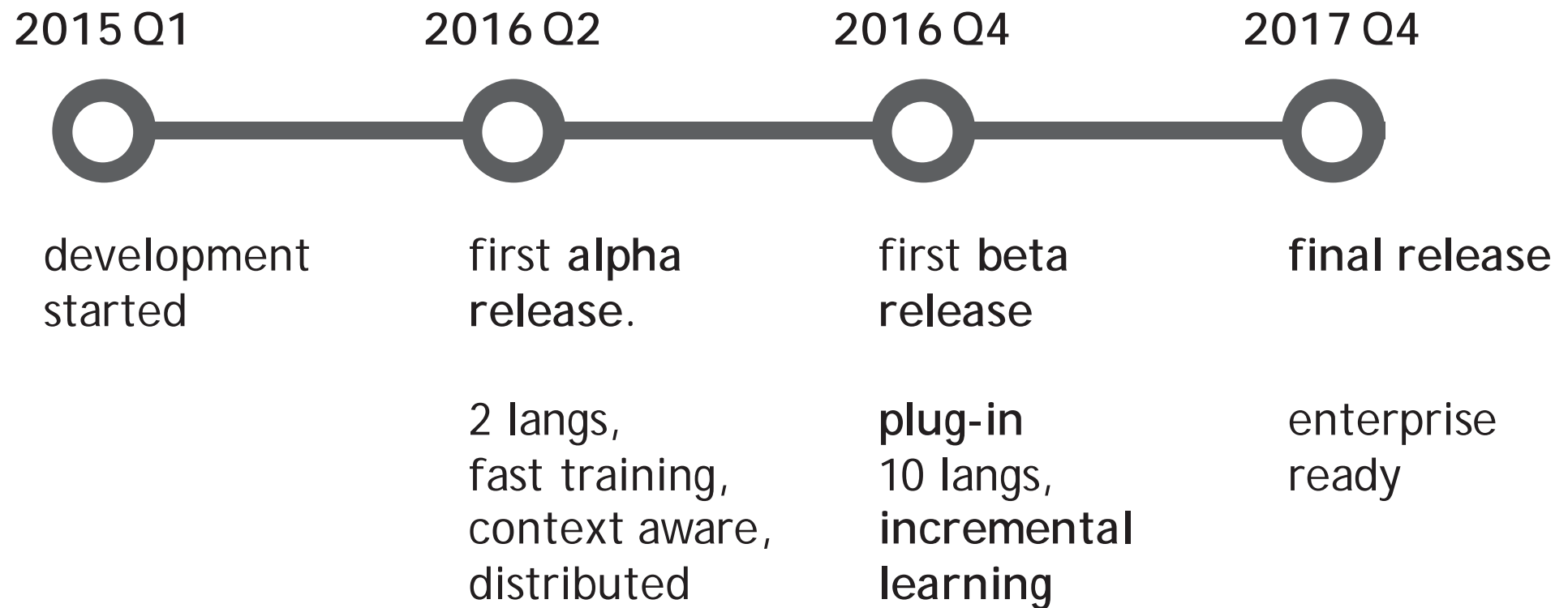
final release

enterprise
ready

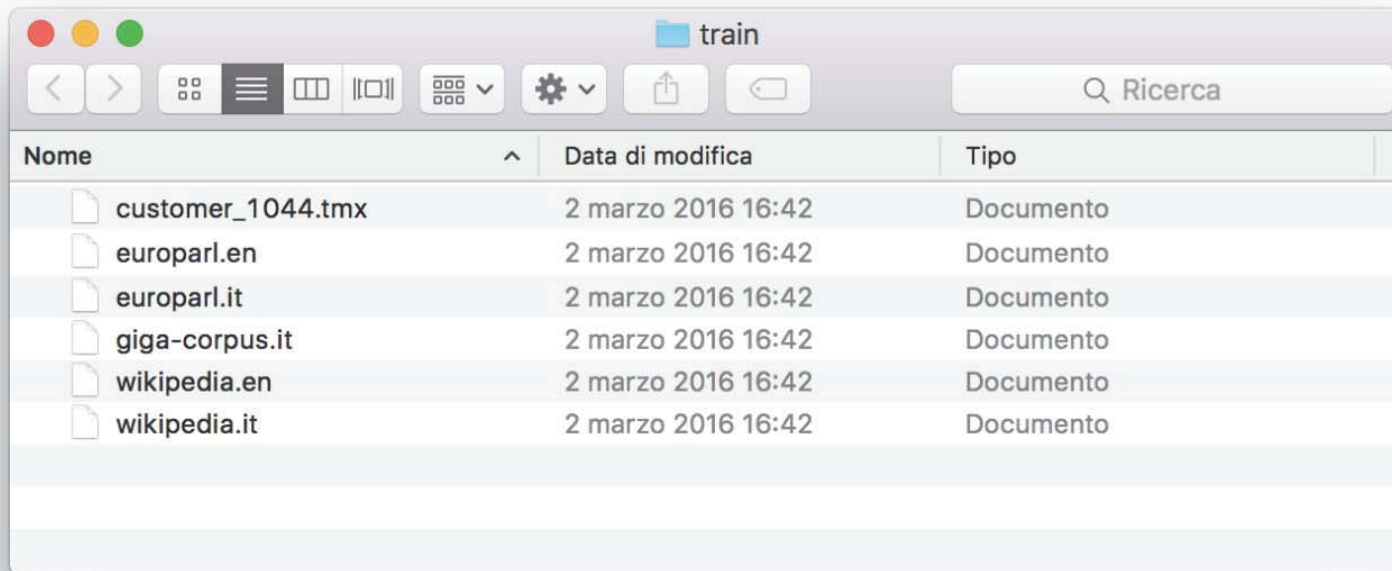
ModernMT

Next Generation
Machine Translation

Roadmap



Prototype - Easy training



```
> mmt create en it path/to/data
```

Prototype (April 2016) - Fast training

Training takes **30s** for a
1M word TM

MMT is **12x time faster**
than std Moses



Context aware translation

TEXT 1

**We're going out.
party**

TRANSLATION

**Nous sortons.
fête**

Context aware translation

TEXT 1

We're going out.
party

TRANSLATION

Nous sortons.
fête

TEXT 2

We approved the law.
party

TRANSLATION

Nous avons approuvé la loi.
parti

Context aware translation

SENTENCE

party

CONTEXT

We are going out.

TRANSLATION

fête

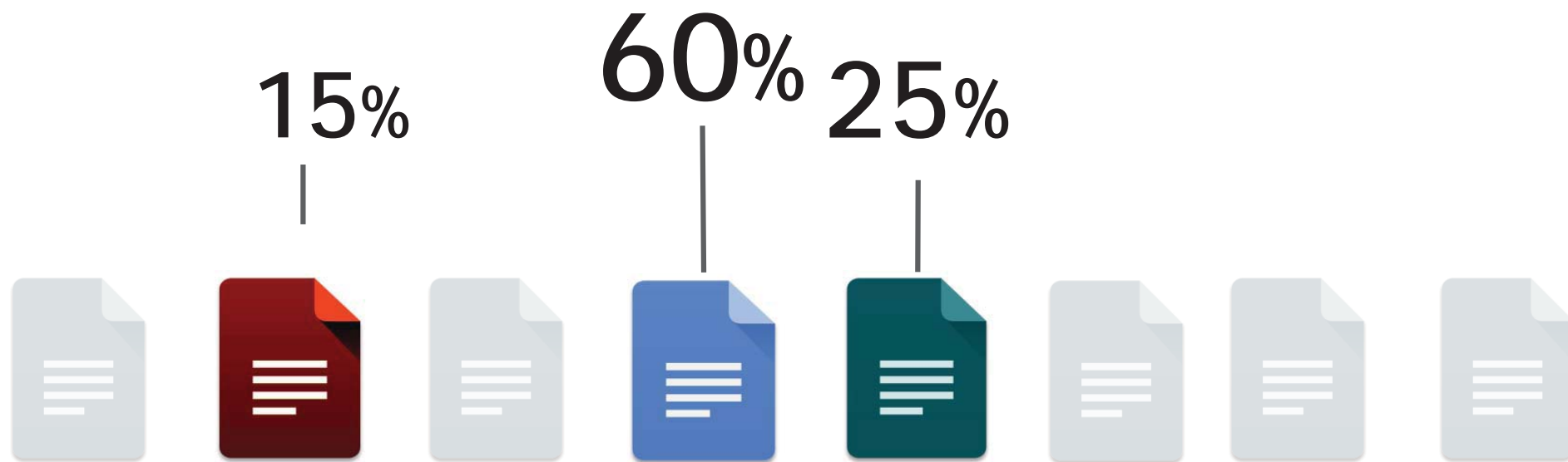
CONTEXT

We approved the law

TRANSLATION

parti

Context Analyzer



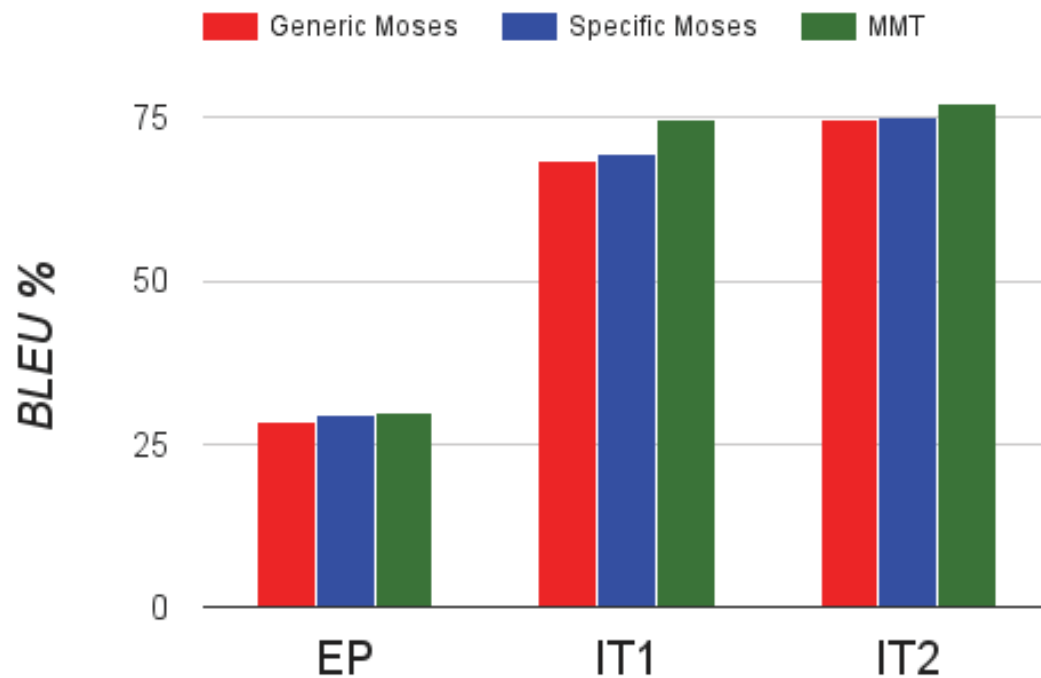
REST API

GET /translate?q=party&context=We+approved+the+law

```
"translation": "parti",  
"context": [  
  { "id": "europarl",  
    "score": 0.10343984  
  }, ...  
]
```

```
> mmt start
```

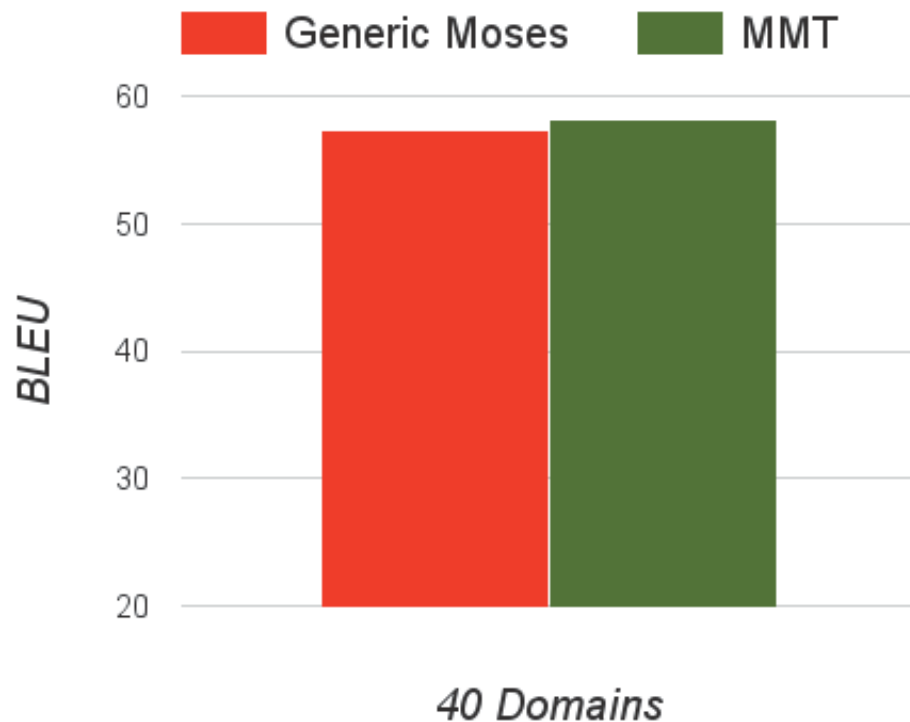

Minimum Viable Product (June 2015)



**6x faster training
than std Moses!**

**MMT outperformed
specific and
generic Moses
 $0.5 \leq \Delta \leq 5$**

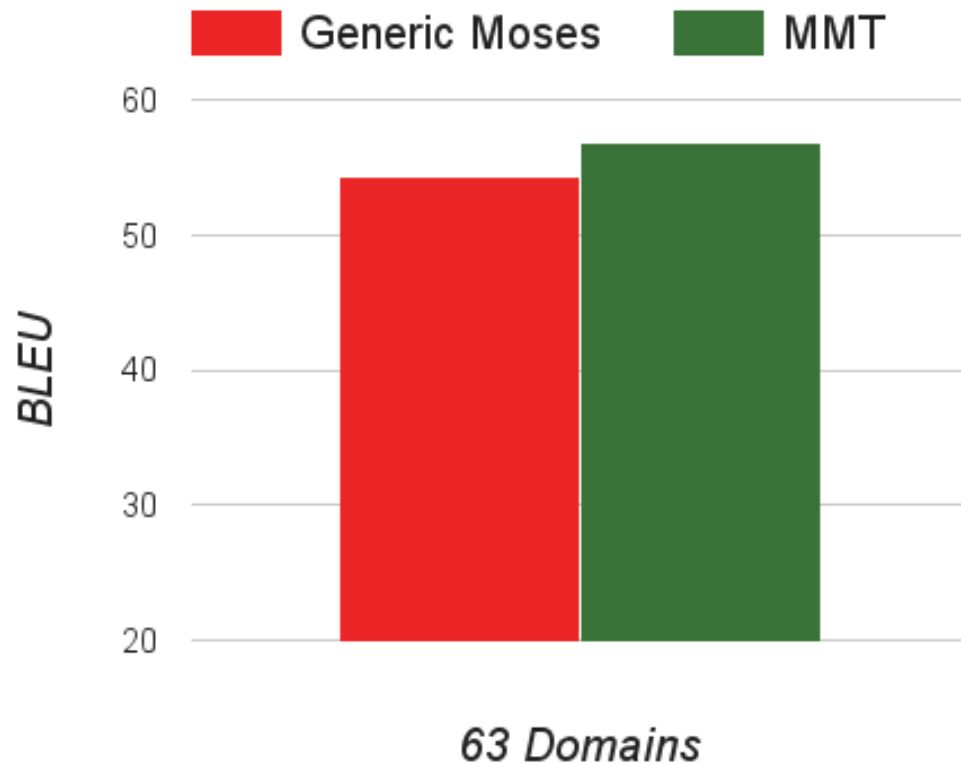
Prototype (January 2016)



11x faster training
than std Moses!

MMT outperformed
specific and generic
std Moses
(Delta=0.8)

Prototype (March 2016)

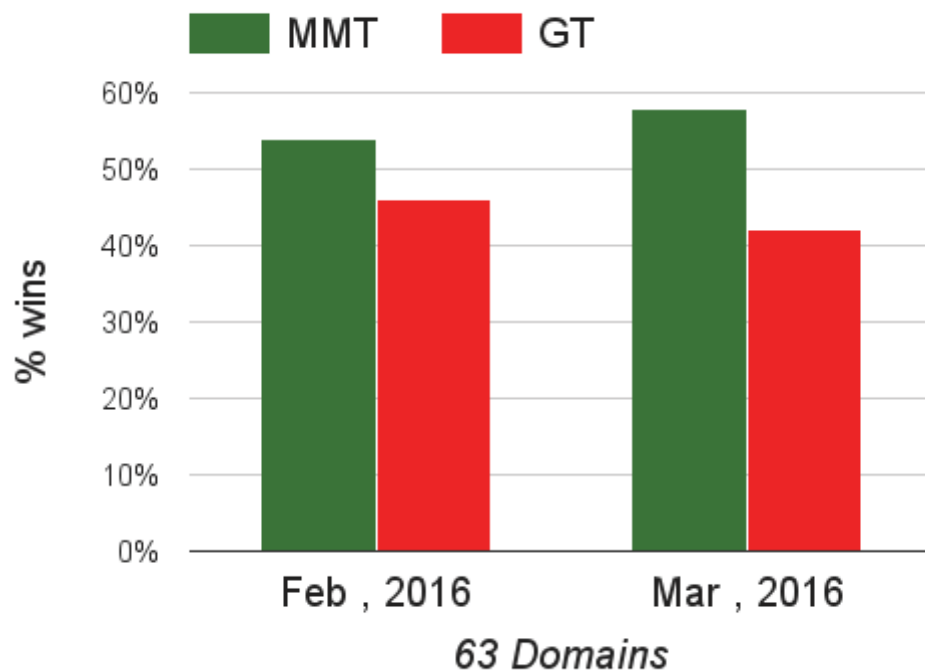


Benchmark 1.1

- No tags
- No xml

MMT outperformed
generic Moses by
>2 BLEU points
12x faster training

Prototype (March 2016)

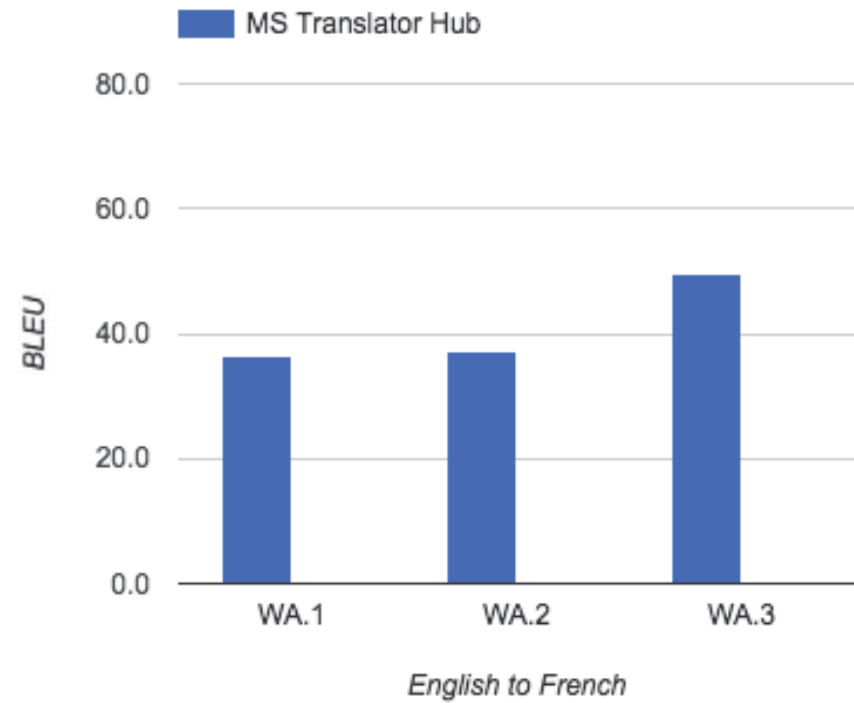
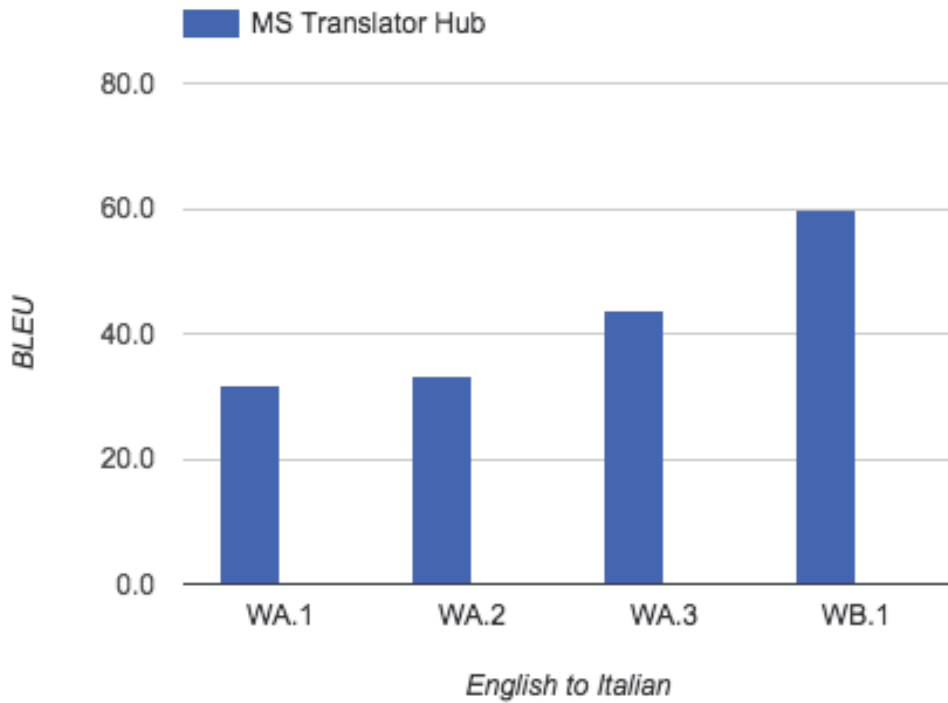


Benchmark 1.1

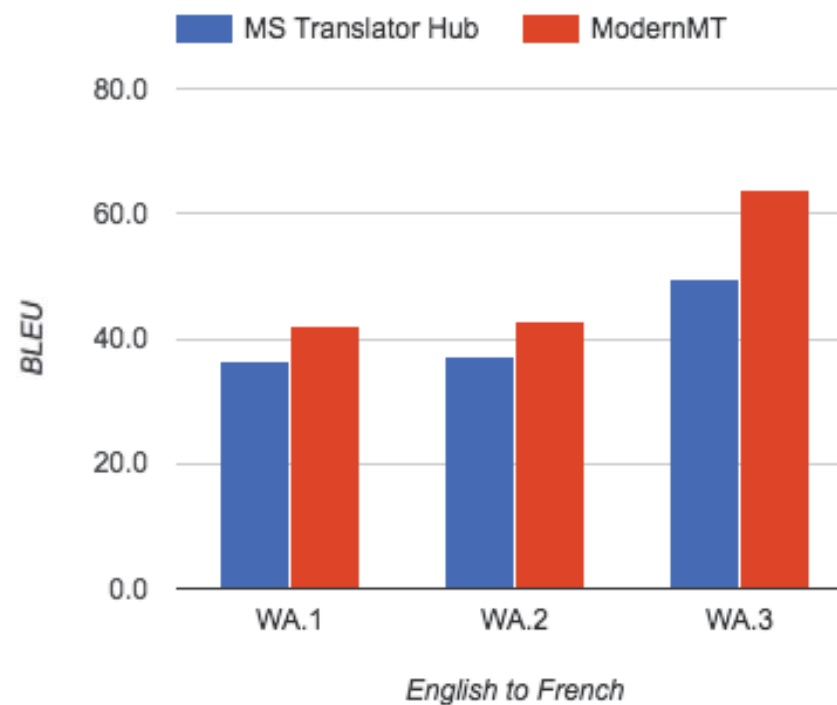
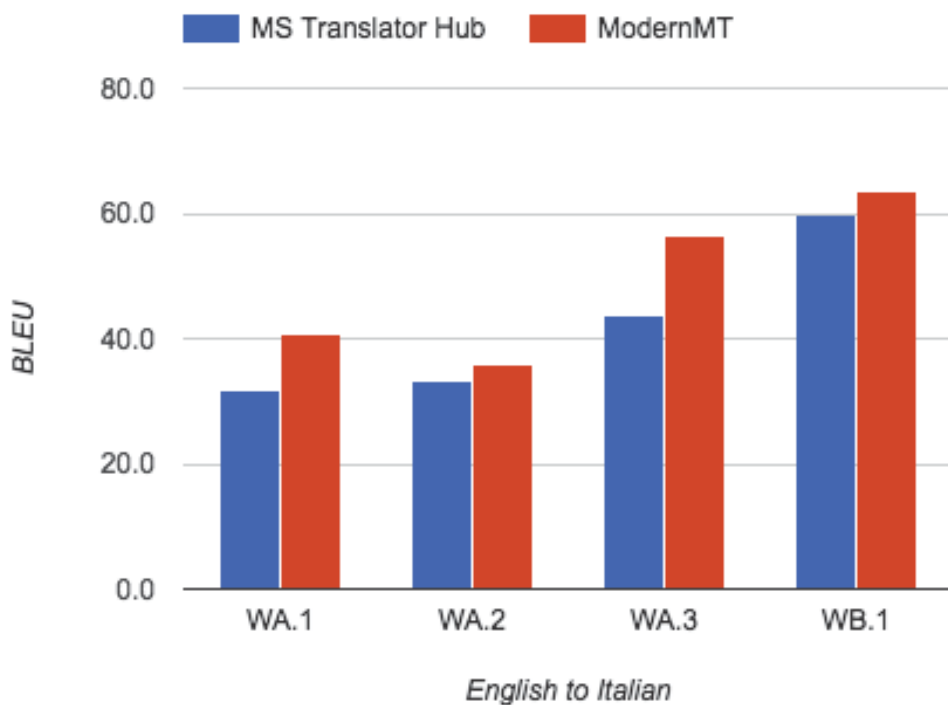
A/B testing vs GT:
~ 300 rnd segments
~ 3 judges

**Distance doubled,
from 8% to 16%!**

MS Translator Hub vs Modern MT



MS Translator Hub vs Modern MT

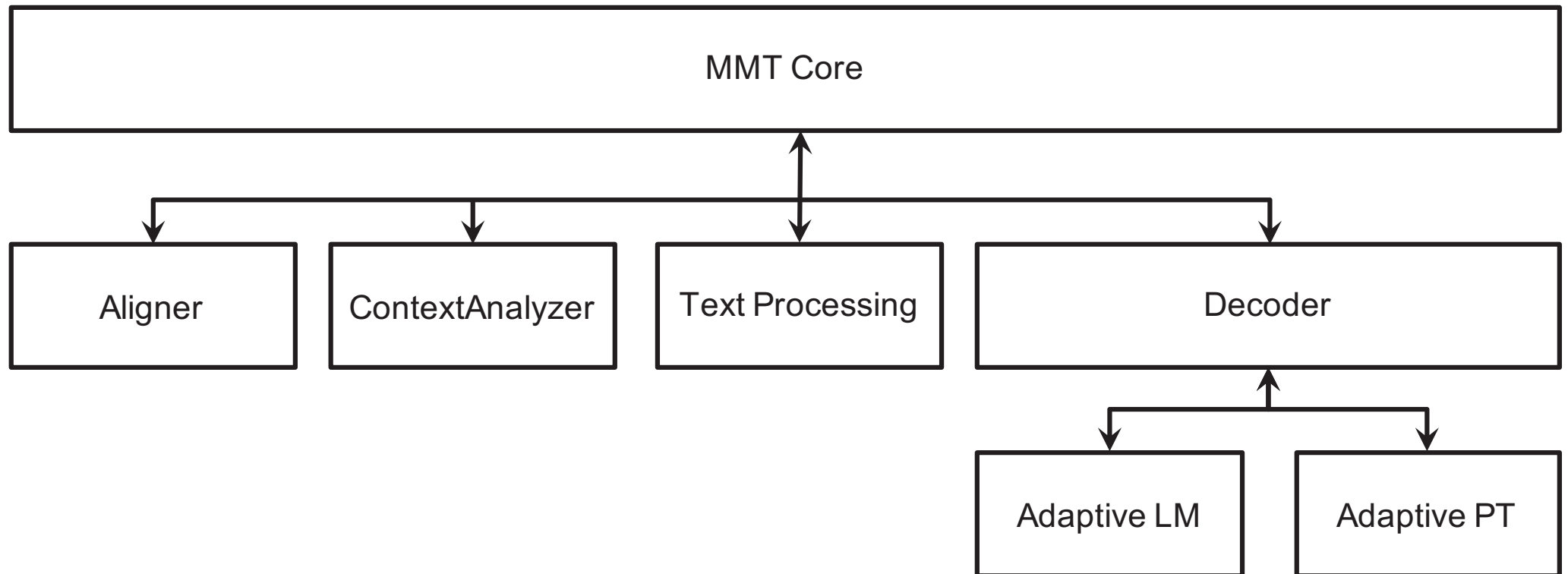


Modern MT core technology

context adaptive
incremental learning



Overall Architecture



Word Alignment

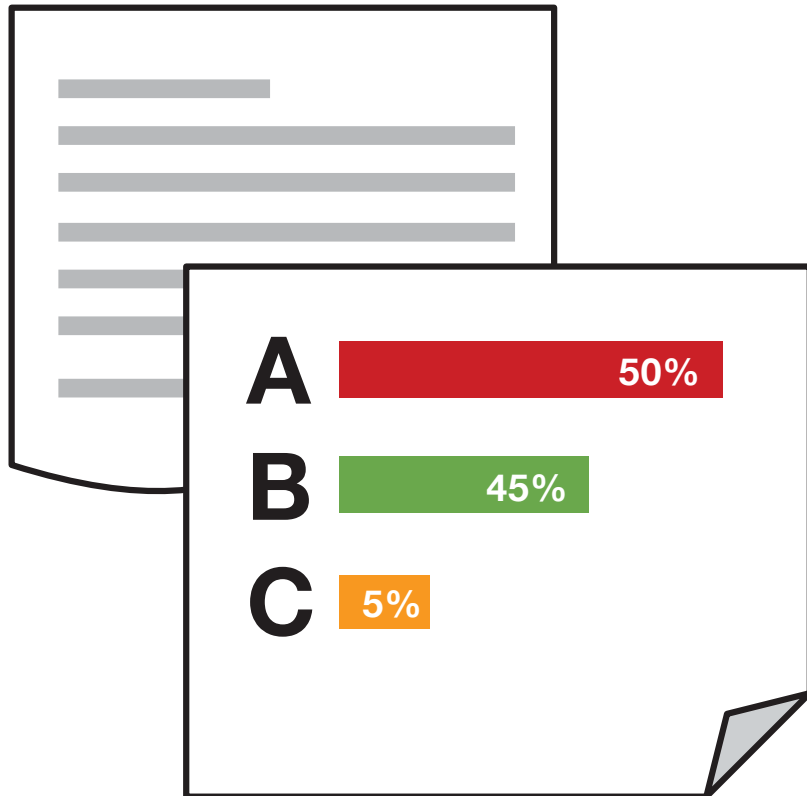
Object oriented **re-implementation** of FastAlign

Multithreading

Incremental training

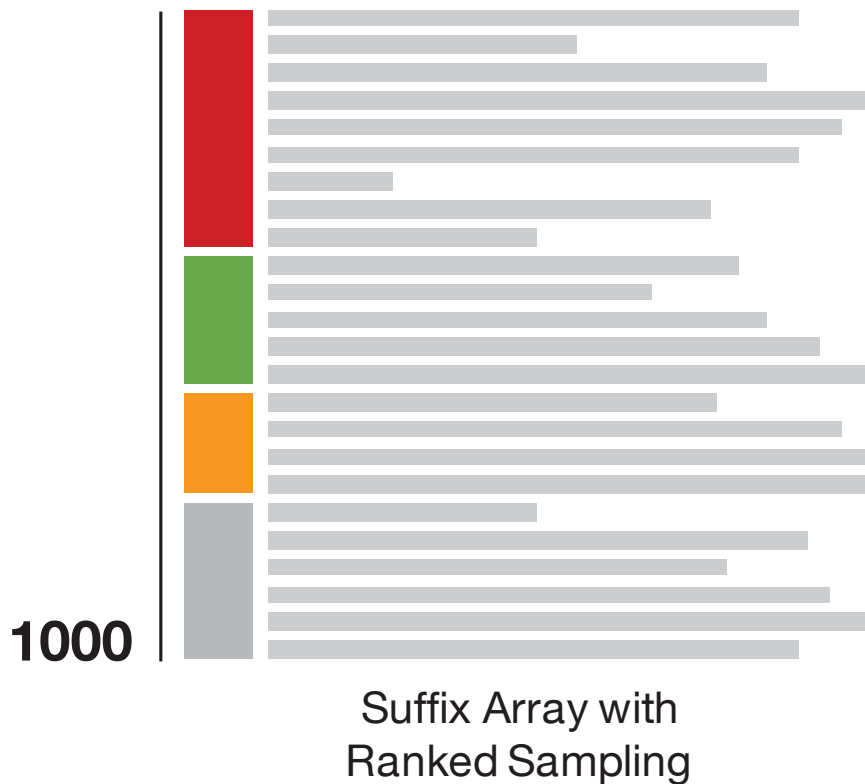
Giza++	FastAlign++
48,000 sec	2,800 sec (17x speed up)
19.3 BLEU	18.9 BLEU (-0.4 loss)

Context Analyzer



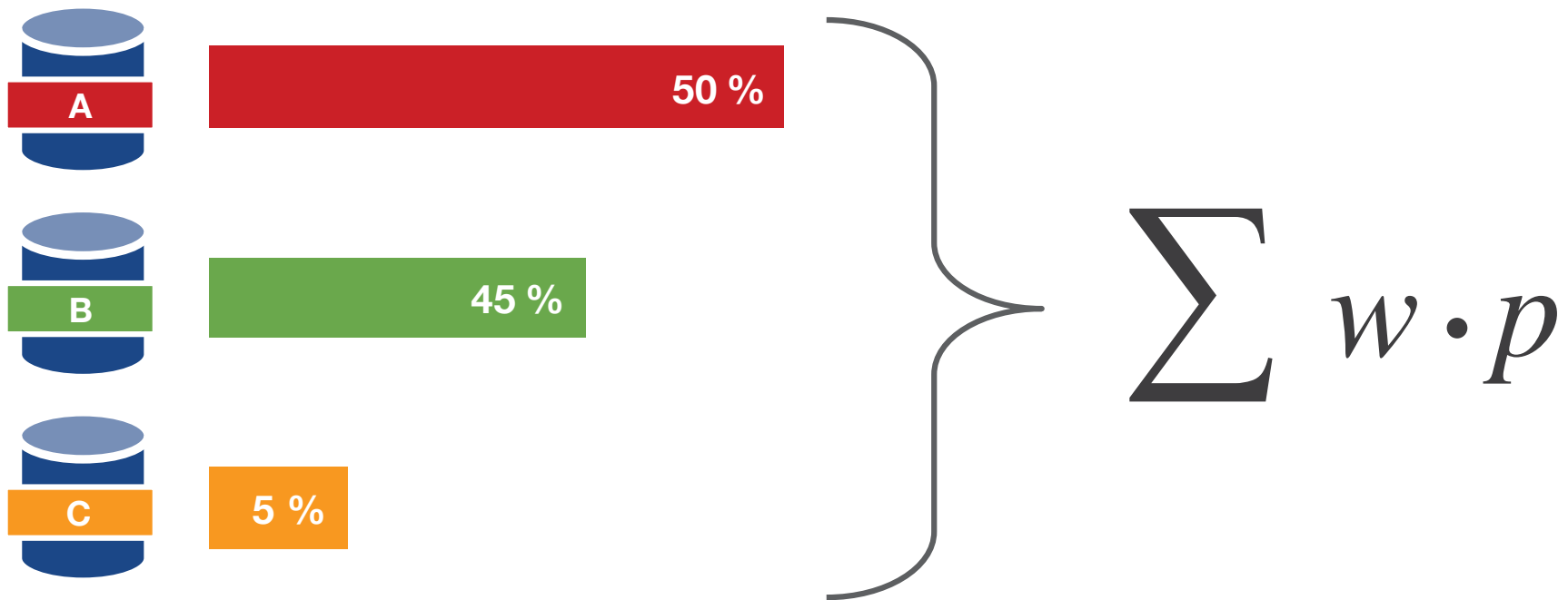
- Analyze the input text (tokenization, stop words)
- Retrieves best matching TMs
- Computes matching score

Adaptive Phrase Table



- Suffix array indexed with TMs
- Phrase table is built on the fly by sampling from the SA
- Phrases of TMs with highest weights sampled first

Adaptive Language Model



ModernMT vs. Moses text processing

- **More** supported languages
- **Faster** processing
- **Simpler** to use
- **Tags** and **XML** management
- Localization of **expressions**
- **TM cleaning**



TM Cleaning

- Multiple versions of segments -> keep most recent only
- Xml expressions or tags -> clean
- Wrong language pairs -> filter out (*)
- Wrong translations -> filter out (*)
- Poor translation quality -> filter out (*)

(*) TMOP - Translation Memory Open-source Purifier

Word Tokenizer

One interface to 8 open-source tokenizers

including **re-implementation** of Moses tokenizer

	Moses Perl Tokenizer	MMT Tokenizer
Languages	21	45 (+24)
Speed*	17k w/s	340k w/s (x20)

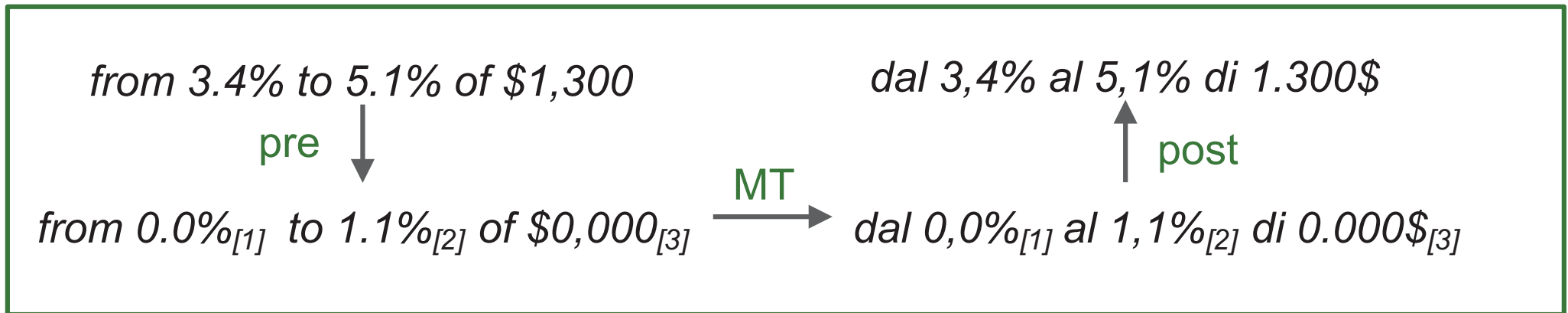
* 4 CPU, 83M word English corpus

Numeric Expressions

Convert digits into placeholders

Translate with placeholders

Apply transformation and heuristics (for unaligned expr)



Numeric Expressions



Subset of Benchmark 1.1

65 segments

135/134 expressions

MMT better than GT

(rel. delta 25%-21%)

detoken. is problematic

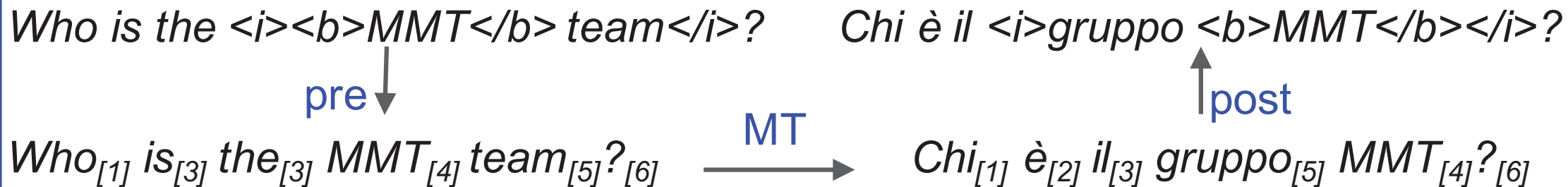
Tag Manager

Identify, classify, and remove tags

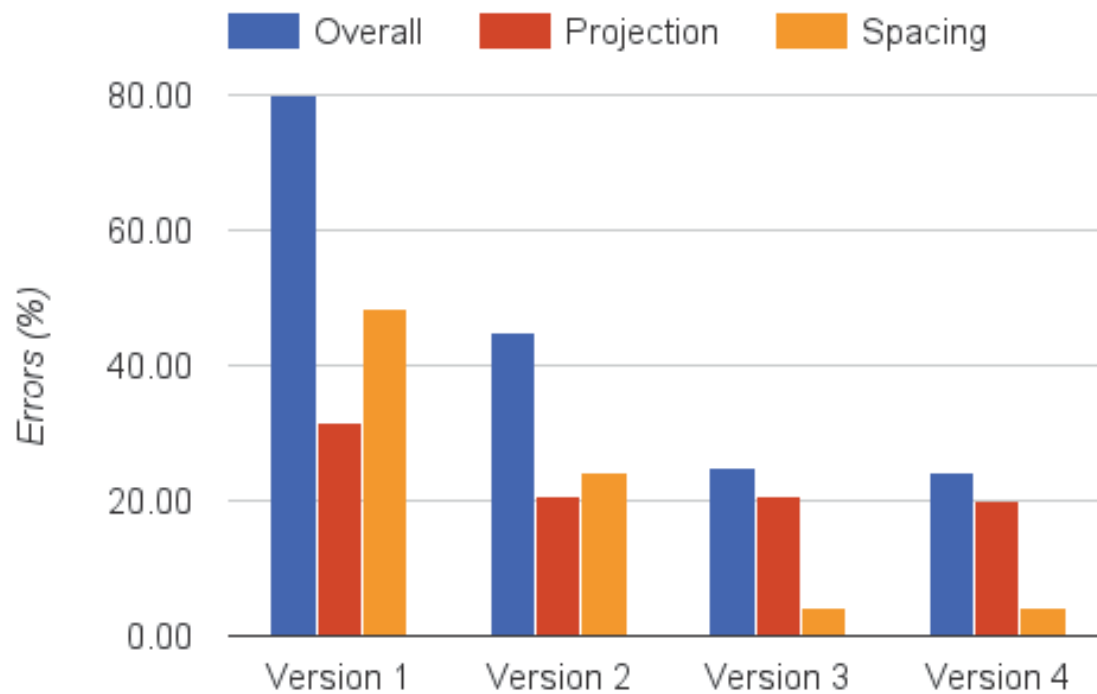
Translate w/o tags

Search insertion points using alignments and heuristics

Handle opening/closing, self-closing, nested, malformed tags



Tag Manager



Tag projection
error < 20%

Spacing errors
around tags \leq 4.2%

Modern MT

big data, context aware, enterprise

Did I mention that MMT will be free?

LGPL/Apache licences
new core technology
no licensing



github.com/ModernMT/MMT

Thank You

Project website:

www.ModernMT.eu



github.com/ModernMT/MMT

Acknowledgment

program: H2020
type: innovation action
funding: 3M €
duration: 2015-2017
grant: 645487



Team

Davide Caroselli

Alessandro Cattelan

Luca Matrostefano

Marco Trombetti

Jaap van der Meer

Achim Ruopp

Anna Siamotou

Uli Germann

David Madl

Luisa Bentivogli

Nicola Bertoldi

Mauro Cettolo

Roldano Cattoni

Marcello Federico

Matteo Negri

Marco Turchi

Web App UI Layout Sniffer

First Author

Raymond Peng pengr@vmware.com

Second Author

Xin Jing Hu xhu@vmware.com

Product Globalization, VMware, Beijing, 100190, China

Abstract

The problem: the UI layout looks great in the English version of VMware web applications but when the product is localized for our global markets, the UI layout doesn't work.

UI layout bugs found during product localization can be as high as 15–35 percent. Testing, triaging and fixing so many UI layout bugs involves multiple cross-functional teams, and requires extra time and budget. Taking one layout bug as example: QE tests and reports it, assigns it to Dev for fix, and QE then verifies the fix to sign off. If the bug is not fixed correctly, QE re-opens the bug and repeats the whole process. This process is complex and efficiency is painfully low. Consequently, it dramatically slows down product delivery in today's fast-paced and competitive market.

These bugs happen for various reasons but we rarely have time to wait until the testing team finds and reports them in Bugzilla. We must find a better way to identify these bugs.

We believe that UI designers or developers can and should avoid these bugs, and in an automatic smart way as they work from mockup to layout coding.

We propose to help the designer sniff potential layout problems early in the software development cycle to kill the problem from the beginning.

Our idea is to develop and adopt a browser extension to call and refresh UI pages by replacing English UI strings with machine translation (MT) that is close to human translation. UI designers or developers can identify potential layout problems at any time, just as an actual user would see in their production environment.

1. Background

As below sample in Figure 1 shows, estimated an 15%-35 % globalization (g11n) bugs belong to the UI layout category for a typical VMware web application, especially for products or features released for the first time.

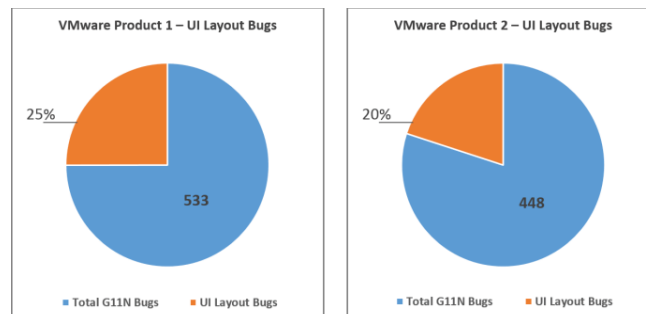
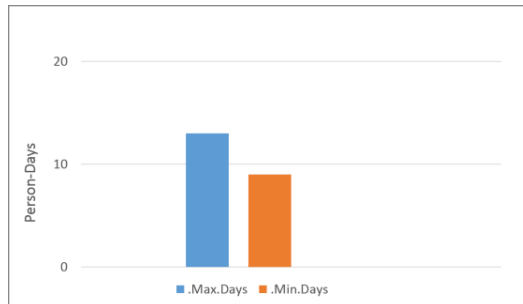


Figure 1. Layout Bug Percentages for Three VMware Products

Testing, reporting, triaging and fixing so many layout issues involves multiple parties and requires a high investment in time and cost, and dramatically slows down time-to-market.



For example, 132 UI layout bugs were reported for product 1 and 91 layout bugs for product 2, as illustrated in Figure 1. One engineer can fix 10–15 layout bugs in one working day. To fix the 132 bugs for Product 1 then would need 9–13 person-days, as shown in Figure 2.

Figure 2. Bug-Fixing Workload for Product 1

Besides the time and cost, another challenge is that some layout bugs can be hard to resolve because the fix does not just involve UI resizing but depends on a localization vendor. Usually there is no time for a re-translation cycle given the schedule pressure.

These problems obviously drag down VMware’s ability to deliver better value to global markets quickly.

Our idea, the Web Application UI Layout Sniffer, is designed for VMware web applications, to prevent and solve layout problems at early phases of the software development cycle. With our technology, the expected result is that there will be no layout bugs during i18n and l10n testing.

2. Overview

The web applications mainly render the UI layout within browser in two ways:

Client rendering: The initial request loads the page layout, CSS and JavaScript. All common contents like static pages will be returned from backend server except that some or all of the content isn’t included. Instead, the JavaScript makes another request (Ajax), gets a response and generates the appropriate HTML.

Server rendering: The initial request loads the page, layout, CSS, JavaScript and content. For subsequent updates to the page, the client-side rendering repeats the steps used to get the initial content. Namely, JavaScript is used to get some JSON data, and template is used to create the HTML.

In standard processes, QA has to wait and use a pseudo build until all UI freeze to start a series of testing. Once layout bug is found, the whole process has to repeat again. Nevertheless, it is too late and risky to fix global CSS settings for developers at this late stage in product release cycle.

Our idea eliminates this dilemma and streamlines the process from the beginning because we kill the possibility of layout bugs at the design and development stage without needing to wait for a UI freeze to verify and report such bugs. We developed a browser plug-in to support both major types of UI layout rendering. Contents script that can read details of the web pages whenever the browser visits, and analyze contents in text node.

We utilize Machine Translation from a web portal to represent the target language just like

being translated by human translators, so developers can check and verify potential layout issues. The plug-in also supports customized pseudo translation by extending English string length by 20–80 percent longer, as chosen by the user, for a static HTML check. Obviously, using machine translation is more accurate and suitable for detecting layout issues since machine translation results are much closer to the real translations than the pseudo strings. This could increase the bug hit rate and provide much more valuable references for UI designers to adjust the layout related parameters.

Typical testing and user scenarios. UI designers and developers using our browser extension can detect potential layout issues at any time on a daily basis.

3. Advantages

Facing with a real problem world of UI layout bugs, we have following goals in mind:

- **Lightweight:** Most automation tools are hard to install, given the heavy, dependent environment needed, but a browser extension is easy to install, without requiring that extra assets be pre-installed.
- **Easy to use:** No training is needed to use this tool to check potential layout issues. It’s a “what you see is what you get” (WYSIWYG) application for web UI designers.
- **MUI supported:** The tool supports a multi-language interface so UI designers from different countries have no language obstacle using it on a daily basis.
- **Fast performance:** Adopting a DOM tree to read each node one time, it stores the text information and uses the xPath to locate the nodes, so users will not feel a response delay.
- **Easy to extend:** For now, the tool supports two types of MT engines, but it can be configured to connect to all kinds of MT engines.

4. Architecture

Figures 3 show the architecture of the Web App Sniffer. It consists of three core components: the DOM Tree Parser, Pseudo Module, and MT Adapter.

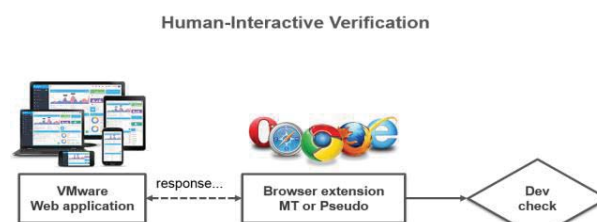


Figure 3. Human-Interactive Verification

4.1. DOM Tree Parser

This component is used to parse the terminal DOM tree and extract the text nodes and natural language processing words. To handle text in a DOM structure is to get intimate with the text

nodes themselves. The application needs to distinguish different kind of nodes. As a W3C standard, the key differences are summarized as follows:

- Text nodes don't have descendant (or child nodes), instead only text content without any HTML or XML markup.
- Almost all contents about a text node are contained in its data (or node value) property, which contains whatever text the node encapsulates.
- Text nodes don't trigger events and can't have any styles applied.

These text nodes will appear in the DOM structure just like element nodes. For example, if we consider this paragraph:

```
<p>
  <a href="/ dashboard ">Go back to dashboard! </a><br/>
  This dashboard reference ID is 12345678.
</p>
```

Its DOM structure is as follows:

```
-> P ELEMENT-> TEXT NODE (data: "n ")
-> A ELEMENT (href: "/dashboard")
-> TEXT NODE (data: "Go back to dashboard!")
-> BR ELEMENT
-> TEXT NODE (data: "n This dashboard reference ID is 12345678.n")
// temporary. innerHTML is now:
// "n <a href="/ dashboard ">Go back to dashboard! </a><br/>This dashboard reference ID is 12345678."
// |-----|-----|-----|-----|
// |
// | ELEMENT NODE      TEXT NODE      ELEMENT NODE      TEXT NODE
```

The parser module supports two types of replacing text sources: using pseudo strings to replace the original text and using machine translation.

Pseudo: The tool supports using pseudo tags, traversing all text nodes according to the node type property, including the nested nodes. At the same time, it does not touch the contents in some special tags as below:

```
var arr = ["STYLE", "IMG", "NOSCRIPT", "SCRIPT"];
```

In special HTML structure, before traversing the nodes, it also needs to traverse the Iframe node.

By current design, in pseudo model the tool will just add pseudo string and pseudo mark simply to expand English string length by 20 – 80 percent. Annotations and semantic are not considered.

Machine translation (MT): The tool supports using MT to pull translation to replace English text. The HTML contains all possible markups that a typical VMware web application has. This means the MT engine translates the whole HTML content including its markup. How-

ever, the MT engine does not support HTML markup well, e.g. translating UI text while keeping HTML markup intact. If we don't do any extra processing of the machine translated HTML files, we will have following errors in the translated HTML:

- 1) Disrupted tags — If the MT engine doesn't consider HTML structure, they can potentially move the HTML tags randomly, leading to disrupt tags in the MT result.
- 2) Wrongly placed embedded tags — The example given below illustrate this. It is more serious if content includes links and link targets were swapped or randomly given in the MT output.

```
$ echo 'Hello <b>World</b>' | apterium en-es -f html
<b>Hola</b> Mundo
```

- 3) Missing embedded tags - Sometimes the MT engine loses embedded tags in the translation process.
- 4) Split embedded tags - During translation a single word can be translated to more than one word. If the source word has a markup, for example, an <a> tag. Will the MT engine apply the <a> tag wrapping both words or apply to each word?

All of the above issues can impact the accuracy of layout issues detection. To avoid these issues, our application adopts embedded tags mapping, using the concept of fuzzy matches. Essentially the algorithm does a fuzzy match to find the target locations in translated text to apply embedded tag and content given to MT engine is plain text only.

- 1) For the text to be translated, the tool finds the text of inline embedded tags like bold, italics, links etc. We call it sub-units.
- 2) The tool passes the full text and sub-units to the MT engine. Use some delimiter so that we can do the array mapping between source items (full text and sub-units) and translated items.
- 3) The translated full text will have the sub-units somewhere in the text. To locate the sub-unit translation in full text translation, use a fuzzy search algorithm.
- 4) The fuzzy search algorithm will return the start position of match and length of match. We map the embedded tag from the source HTML to that range.
- 5) The fuzzy match involves calculating the edit distance between words in translated full text and translated sub-unit. It is not strings being searched, but n-grams¹ with n=number of words in sub-unit. Each word in n-gram will be matched independently.

To understand this, let's try this algorithm in a UI string. Translating the English sentence “<p> Please click Start button to run the command sequence</p>” to French: The plain text version is “Please click Start button to run the command sequence” and the subsequence with annotation is “Start”. We give both the full text and subsequence to MT. The full text translation is “S'il vous plaît cliquer sur Démarrer pour exécuter la séquence de commande” and the word Start is translated as “Démarrer”. We do a search for “Démarrer” in the full text translation. The search will be successful and the tag will be applied,

¹ An n-gram model is a type of probabilistic language model for predicting the next item in such a sequence in the form of a (n - 1) - order Markov model. [2]

resulting `<p> S'il vous plaît cliquer sur < span> Démarrer pour exécuter la séquence de commande < /p>`. The search performed in this example is a plain text exact search.

4.2. Pseudo Module

This module provides dynamic Pseudo tag with content of a different length, from 20–80 percent, as chosen by the user. One type of pseudo is ASCII characters, the other one is super string

with some typical Unicode characters.

```
superString: ["表","ホ","あ","A","中","E","é","鷗","停","B","道","ü","ß","à","ù","ª","ñ"]
```

4.3. MT Adapter

The MT adapter wraps the MT service, for example if the user configures MT as Bing Translator, then MT adapter will automatically translate each piece of source contents by Bing Translator engine after source DOM tree was processed by parser.

About call mode, the MT adapter uses Ajax to asynchronously call Machine Translation service. Then depending on the different types of Machine Translation engines, the tool uses DOM parser to do corresponding post-editing on the result from the MT engine.

5. Experiments

We compared results from two different methods : One running with a pseudo tag and the other with machine translation. We utilized an extensive series of languages in the pilot project to compare the hit rate (showed in Table 1, Table 2). For the MT engine, in the first environment we used Microsoft Translator Hub. We also ran on the other MT engine like Google translate. We compare results with different scenarios on VMware Product 3.

5.1. Result and Discussion

Table 1 shows the result of using a pseudo tag to increase the string length by 30% to detect potential layout issues. Table 2 show the result of using MT. Results show that using MT to detect potential layout issues results in a higher hit rate than using a pseudo tag.

Language	Detected By Pseudo	Real	Effective	Hit Rate (%)
German	50	41	34	68
French	58	40	33	56
Russian	63	52	29	46
Spanish	59	53	37	62

Table 1. Hit Rate in Pseudo Model Analysis

Language	Detected By MT	Real	Effective	Hit Rate (%)
German	45	42	41	91
French	43	40	37	86
Russian	50	52	40	80
Spanish	49	53	40	82

Table 2. Hit Rate in MT Model Analysis

Note that :

$$Hit\ Rate = Effective\ number / Detected\ number$$

6. Conclusion

In this paper, we discussed the purposes of developing our browser extension, the Web App UI Layout Sniffer. It fundamentally prevents layout bugs in early UI design phases, saves time and testing costs, and reduces layout bug fixing during the i18n and l10n testing process. Ultimately it helps accelerate product time-to-market. This extension adopts Machine Translation to replace an English string so the UI designer can identify potential layout issues from the user perspective. It dramatically streamlines product globalization. Per our testing and pilot, it is effective in detecting UI layout issues for VMware web applications and it can be considered for production use for designer and developers.

References

- [1] Fighting Layout Bugs: <https://code.google.com/p/fighting-layout-bugs/>
- [2] n-gram: <https://en.wikipedia.org/wiki/N-gram>
- [3] Pseudo Localization: <https://en.wikipedia.org/wiki/Pseudolocalization>

Multilingual Search with Machine Translation in the Intel Communities

Ryan Martin

ryan.c.martin@intel.com

Abstract

This paper describes an experiment performed at Intel to assess the viability of using machine translation for cross-language information retrieval within the Intel Communities (public user forums). Many of the Intel Communities are mixed-language, with a large majority of content being posted in English. In order to make this information available to non-English speakers, our team researched the effectiveness of using machine translation to translate search queries to English using general domain machine translation systems.

1 Introduction

Many of the Intel Communities are multilingual by necessity; a single forum supports users worldwide. Although it is not uncommon for visitors to post content in other languages, the majority of content in these forums is English. In order to improve the experience for non-English speaking visitors, a real-time translation feature was added to the forums in 2012. This feature gives site visitors the ability to translate individual posts to one of 10 languages. While this feature is used often and gets positive feedback from site visitors, a notable shortcoming has been the lack of cross-language search.

It is believed that users would be more likely to find information relevant to their visit if the search results from non-English search terms also included English content within the forums. In order to test this hypothesis, it was first necessary to understand if using machine translation for cross-language search¹ could be used with good results in the selected user forum.

Previous studies have looked at using dictionary-based techniques in cross-language information access (Levow et al., 2005). For this experiment, we evaluated the use of machine translation (MT) without any special query *pre-* or *post-*processing. This decision was largely influenced by the fact that a third-party collaboration platform is used to host the user forums, and we are limited to the amount of customization that can be implemented.

To better understand the viability of using machine translation for this purpose, the Intel team used machine translation to translate non-English search queries written in Spanish and Simplified Chinese², and compared the results with searches performed using roughly equivalent English queries.

The following sections describe the experiment set up, scoring methods, and results.

2 Platform and User Description

The *Intel Support Community* was used as the test platform. This forum is hosted using a third-party collaboration platform. The Support Community serves a large number of active users

¹Or *cross-language information retrieval* (CLIR), using the more common description.

²<https://communities.intel.com/community/tech>

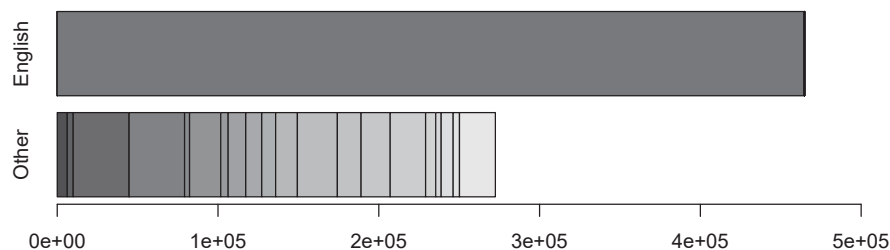


Figure 1: Unique sessions by browser language (based on 1 month of data)

and contains discussions on a diverse set of Intel products and services. There are currently over 70K searchable items (blog posts, discussions, documents), focused on computer hardware and software. Of this content, approximately 82% is from discussion posts³. The variety and quantity of data made this particular community a good choice for the study.

2.0.1 User Language Profile

Based on one month of data from 2016, approximately 37 percent of sessions have a browser language other than English. This ratio of non-English browsers is illustrated in Figure 1.

While this is a relatively high percentage of users with non-English browser settings, an informal review of actual user searches shows that most visitors are searching in English⁴.

2.0.2 Search Engine

The experiment was carried out using the collaboration platform's native search engine. The search engine supports the following features:

- Basic word search (ignores case and order)
- Exact phrase search using quotation marks to delimit phrases
- A simple one-or-more wildcard (*)
- Compound expressions using AND, OR, NOT, and grouping ()
- No removal of stop words
- Content must contain *all* search terms

All of the searches performed for this experiment were basic word searches, and didn't make use of any special query syntax.

³Data collected on August 29, 2016

⁴User data does not record search language. The actual ratio of non-English search queries requires further investigation.

LANG	N	TER	EQUIV.
ES	59	0.300	45.8%
ZH	59	0.318	44.1%
ES*	37	0.324	45.9%
ZH*	37	0.117	82.9%

Table 1: Search Data (* represent actual user data)

2.1 Experiment

Spanish and Simplified Chinese were selected as the source query languages. Both languages are spoken in locales with a relatively high percentage of site visitors. Additionally, these languages both had sufficient examples of non-English queries collected from real user sessions.

2.1.1 Search Data

The search queries were a combination of both real-world searches collected from the site, and artificial searches that were developed by the team based on existing site content. It was necessary to develop our own search examples for a couple of reasons. For one, many searches fail to return results regardless of the language used. It was desirable to have a reference set that was guaranteed to return some meaningful results – there is little reason to translate a search that is known to return few or no results. Additionally, the number of non-English searches collected from the site was relatively small after being filtered for quality.

A total of 59 samples were developed in English by the team as a reference set. These samples were then human translated into Spanish and Simplified Chinese in order to achieve a close approximation of the English source query. The translators were given instructions that each line of the source text was composed of independent search terms – this was done so that translators did not attempt to fix the input by producing more grammatical output.

Additionally, 37 Spanish and 37 Simplified Chinese searches were collected from real-world sessions and human translated to English.

The median length of the searches was 4 terms, although the real user searches for Chinese were somewhat shorter with a median of 2 (See Figure 2).

2.1.2 MT systems

Generally available commercial MT systems were used for the translation of the non-English queries to English. In this experiment, we chose to focus on the use of general domain MT alone. This decision was largely influenced by the fact that we have limited control over the collaboration platform and search engine; the later being more-or-less a black box. Although the domain is relatively well-defined (hardware and software), we have typically relied upon general domain systems for user-generated content since the quality and style of the source content is considered to be less predictable. Improving the system using domain-specific MT systems, or by including in-domain dictionaries in the query translation process (Jones et al., 2008) certainly deserves further investigation.

The selected MT systems are also used to support the real-time translation feature that has been integrated into the platform *Discussions*. TER scores were calculated between the English reference searches, and the English output from machine translating the non-English source (See Table 1).

The EQUIV column of Table 1 shows the percentage of translations that were functionally equivalent to the reference query. In other words, the percentage of searches where the the translation and reference contained the same terms when ignoring both letter case and order. These

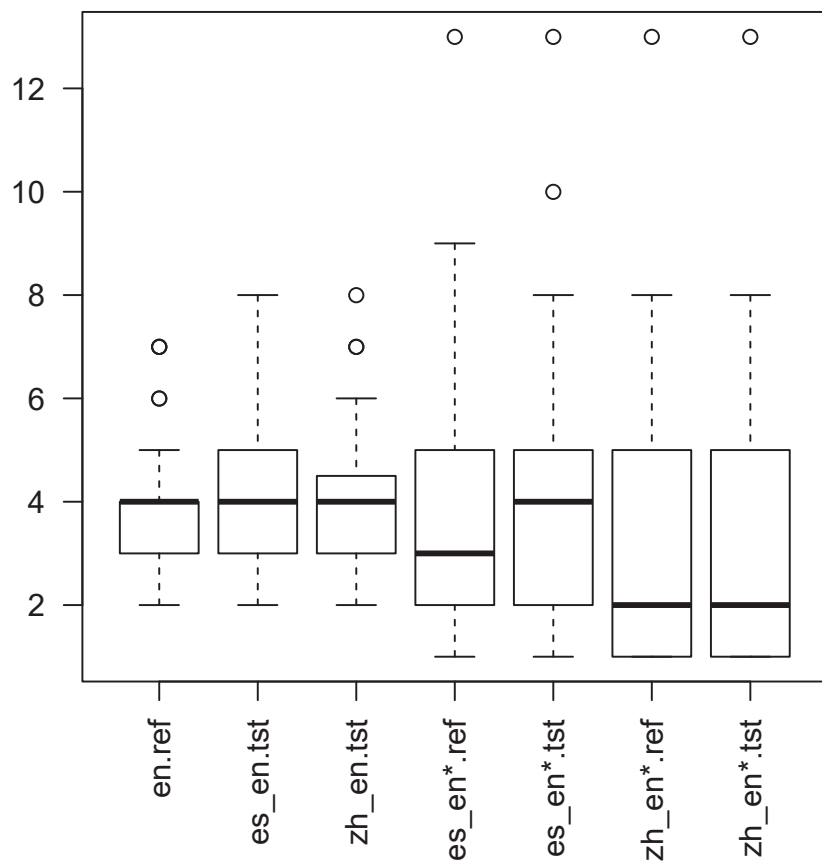


Figure 2: Search query length (words)

translated queries will return the same search results as the reference based on the behavior of the search engine (Section 2.0.2).

2.1.3 Data Collection

An automated script was used to perform searches of each of the English reference queries. The same process was used to collect results from the equivalent machine translated versions of each search (the ‘test’ set).

For each search performed, the top 5 search result URLs were parsed from the result page⁵. When fewer than 5 results were returned, then all URLs from the results page were collected.

3 Scoring Methods and Results

Each search query was given a point for each search result that matched between the reference (English) and the translated search. The maximum score for any search was 5 points given that only the top five results for any search were recorded. Note that the total number of possible points for a particular search may be less than 5 if the reference search returned less than 5 results.

A final calculation was made by summing the points, and then dividing by the total number of search results in the reference set. These results are displayed in the SCORE column of Table 2.

In addition to calculating the simple ratios shown in Table 2, we also calculated adjusted scores that scaled the points for each search to the range 0 - 5. In the resulting frequency distribution, 0 represents no matches between the reference and test search, while 5 means that the search results were identical between the reference and test (e.g. This was done so that 3/3 and 5/5 would both get a score of 5). The results are shown in Figure 3.

Finally, the TOP RESULT column of Table 2 shows the percentage of searches where the top (first) search result matched between the test and reference searches.

3.1 Analysis

The scores for the four test sets displayed in Figure 3 all share a very similar distribution. There was also very similar distributions between the test sets developed by the team and those harvested from site users.

It is important not to interpret the scores displayed in Table 2 as a measure of search effectiveness since the experiment used a diverse set of search queries that were known to return reasonable results. To actually measure the total search accuracy, one would also need to include failed or low quality searches. In other words, the final score does not imply that actual users would have a similar success rate in finding relevant information.

These results show that given a *good* search query where the equivalent English would return meaningful search results, that the machine translated queries provided identical results most of the time. In our experiment, a 4 or 5 could be expected at least 70% of the time.

When the translated search fails, it usually fails completely. This is shown by the spikes at 0 in Figure 3. This particular feature of the distributions could be explained a number of ways. One likely explanation is that the basic search fails if there are any out of vocabulary (OOV) terms. This is true regardless of language, or whether MT was used to produce the search terms. For the basic search to succeed, all of the terms in the search query must appear in a target document. It is reasonable to conclude that a translation error in a single term could cause a search to completely fail – this hypothesis would need to be confirmed by a complete review of the translated search terms.

⁵Each URL contains a unique numeric identifier for the target content.

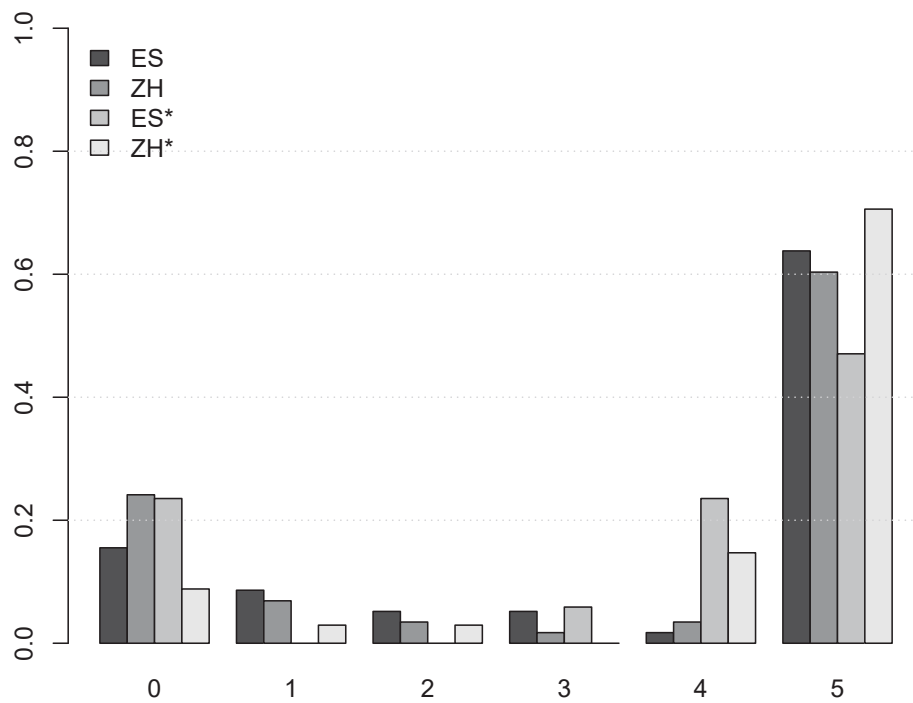


Figure 3: Distribution of scores

LANG	N	SCORE	TOP RESULT
ES	37	71.3%	71.2%
ZH	37	67.9%	64.4%
ES*	59	71.1%	64.9%
ZH*	59	79.3%	78.3%

Table 2: Total ratio of successful (intersecting) results.

Although Spanish and Simplified Chinese had very similar results, it should be noted that these are some of our best performing MT languages, especially when the target language is English. Similar results should not be assumed for other language pairs. Additionally, the relevance of the search results was not considered in the scoring process; only the intersection count of reference and test search results are used in the scoring.

4 Conclusions

Our study looked at the viability of using machine translation to translate non-English search terms to English within the Intel Communities. The ability for users to search cross-language across discussion forums is one possible method to connect users with relevant content in multi-language user forums. Once users are able to find relevant content, they can use the real-time translation features already available on the site. The median score of the four test sets (based on the scoring method described in Section 3) was approximately 71%, and we consider this to be a good baseline when evaluating methods to improve the system. Further research may include evaluating custom MT system developed for the Intel domain.

Acknowledgments

I would like to thank Julie Chang for her assistance coordinating translations and providing actual user search queries from the Communities.

References

- Jones, G. J. F., Fantino, F., Newman, E., and Zhang, Y. (2008). Domain-specific query translation for multilingual information access using machine translation augmented with dictionaries mined from wikipedia. In *Proceedings of the 2nd International Workshop on Cross Lingual Information Access Addressing the Information Need of Multilingual Societies*, pages 34–41.
- Levov, G.-A., Oard, D. W., and Resnik, P. (2005). Dictionary-based techniques for cross-language information retrieval. *Information Processing and Management*, 41:523–547.



MT Thresholding: Achieving a defined quality bar with a mix of human and machine translation

Dag Schmidtke

Senior Program Manager, Office Global Services & Experiences

Microsoft Ireland



Recycling & Machine Translation in Office

Goal: Maximise use of Recycling and Machine Translation, while protecting Customer Satisfaction

- Focus spend: Human translate high priority and most popular content
- Recycle as much as possible and machine translate the rest, publish and upgrade based on quality and traffic
- Increase velocity & reach with added coverage

Recycling

Reuse of existing high quality translations
Automated in production process
Typically reduces wordcount & cost by 60 to 70%

Human Translation with MT Post-Editing (MTPE)

Improve MT output with human translators
No quality degradation
Applied after recycling
Part of production process for UA and UI
In use for 35+ languages

MT Publishing

Machine translation published without human editing (raw-MT)
Applied after recycling
Used for long tail content, speed
In use for 38 languages

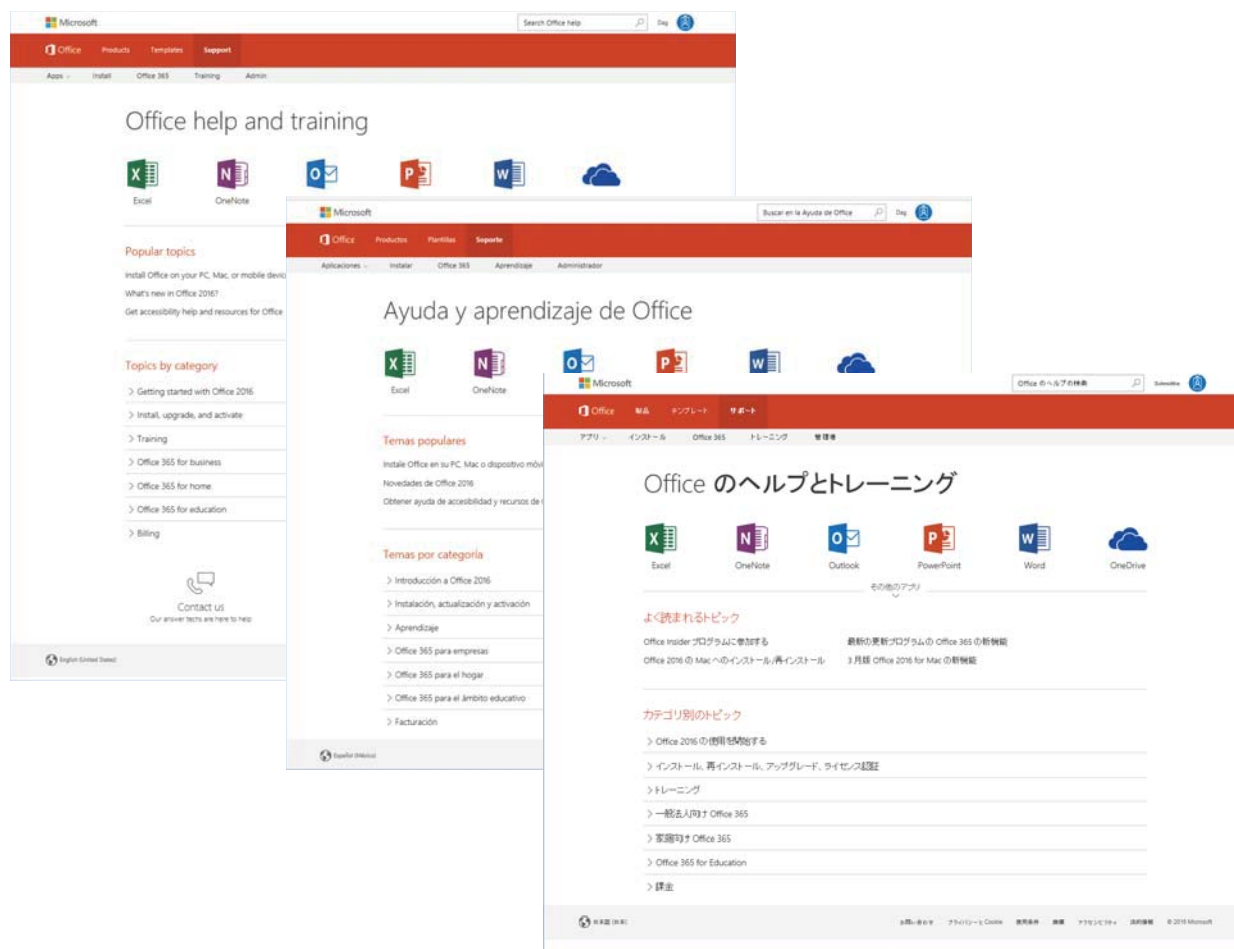
Human Translation (HT) Workflow



MT Publishing Workflow



Use-case: MT Publishing for Office Help



- Support.Office.COM (SOC)
- End-user help & training
- Large scope: 40 languages x 15k articles
- About 2.9 bn PVs /year, 45% non-English
- Significant translation effort
- Can MT help?



Challenges

- Is MT good enough for end-user help?
 - MT quality unpredictable and hard to measure
 - What is the right metric for 'MT quality'?
 - Will the Office end user audience accept MT?
- How to prioritize human versus machine translation?
- How to achieve scale?
 - Office Translation requirement: 100s of millions of words/year
- How to listen to customers and respond?

MT Publishing: Our Approach

1. Plan: Establish KPIs

- Quality bar: 'Acceptable'
- Speed: <24 hours
- Scope: Low PV topics

2. Prepare: Engineering

- Benchmark MT quality
- Automation in platform
- Internal telemetry
- Business Intelligence: traffic and ratings

3. Deploy: Optimised MT

- Custom MT domains with Microsoft Translator Hub
- Recycling: re-use of high quality translations
- Quality gating with thresholding

4. Iterate and adjust

- Active monitoring of usage and ratings
- Traffic-based Upgrades
- Adjust thresholds
- Increase MT scope



Quality model

Quality bar – ‘Acceptable’

- MT Publishing needs to reach a minimum bar to be usable
- Starting metric: 2.5 /4 for human evaluation
- Ongoing metric: within 10% of Human Translation User Rating (CSAT), for each language

Thresholding - based on initial human evaluation and recycle rate per article

- Good quality MT ($\geq 2.5/4$): article published without restriction
- Medium quality MT (≥ 2.5 with recycling): article recycle rate of $\geq 50\%$ needed to publish
- Lower Quality MT (< 2.5 with recycling): article recycle rate of $\geq 80\%$ needed to publish

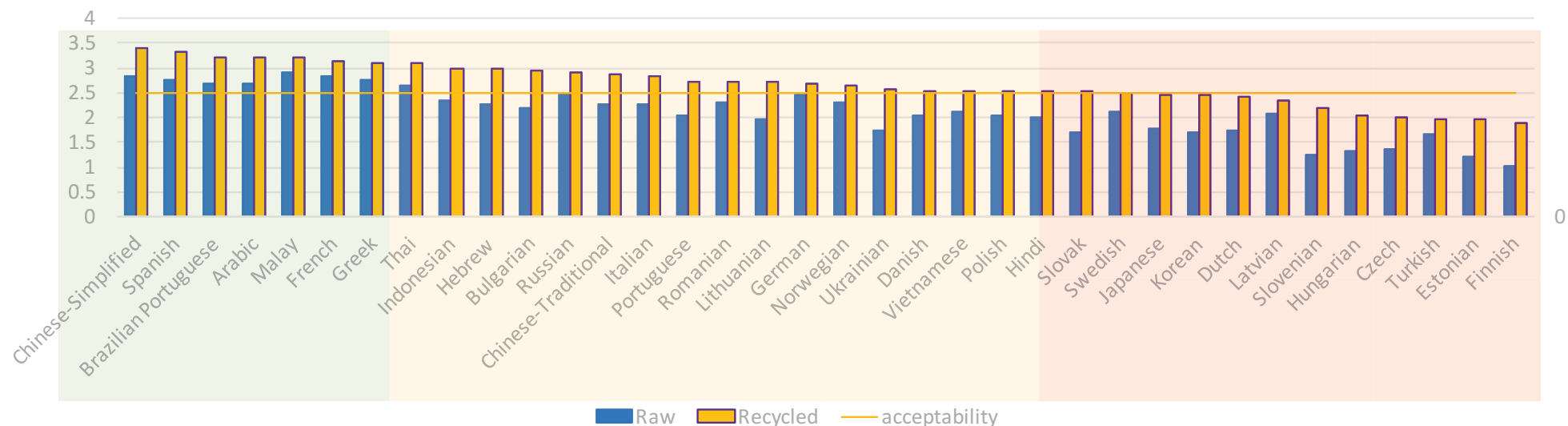
Iterate and adjust

- High traffic MT assets (within top 70%) upgraded to HT, to ensure optimal customer experience
- HT used for high priority articles based on source meta-data



Initial MT Human Quality Evaluation

Machine translation quality, Support.Office.COM evaluation, Oct 2014



Methodology

- Human evaluation, 3 reviewers per language
- Judged on scale of 1-4, with 2.5 set as acceptability threshold for production use
- 10 help articles x 36 languages: 5 with 50% recycling, 5 with low or no recycling

Results: Variable MT quality

- 8 languages have good enough, 'acceptable' MT quality
- 16 additional languages reach quality bar only with use of recycling, medium quality
- 12 final languages have lower quality, did not meet the quality bar even with recycling



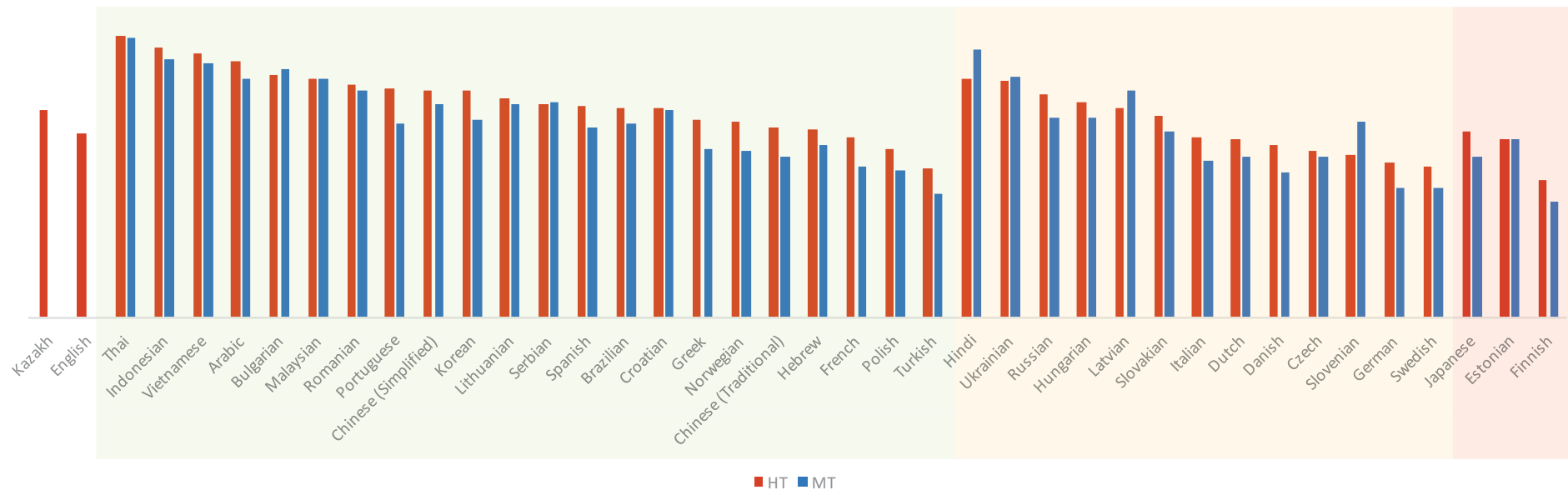
Progress

- June 2015
 - MT Publishing in use for 38 languages, 20% of monthly translation volume
 - Initial thresholding: 8 languages with 0 threshold
- November 2015
 - MT Publishing used for >50% of monthly volumes
 - Thresholding adjusted: 18 languages with 0 threshold, 15 with 50% threshold
- August 2016
 - MT Publishing used for >70% of monthly volumes
 - 47% of live articles published through MT pipe, generating 15% of traffic
 - Thresholding adjusted: 22 languages with 0 threshold, 13 with 50% threshold



Support.Office.COM Customer Satisfaction

Support.Office.COM Customer Satisfaction, August 2016, HT and MT articles

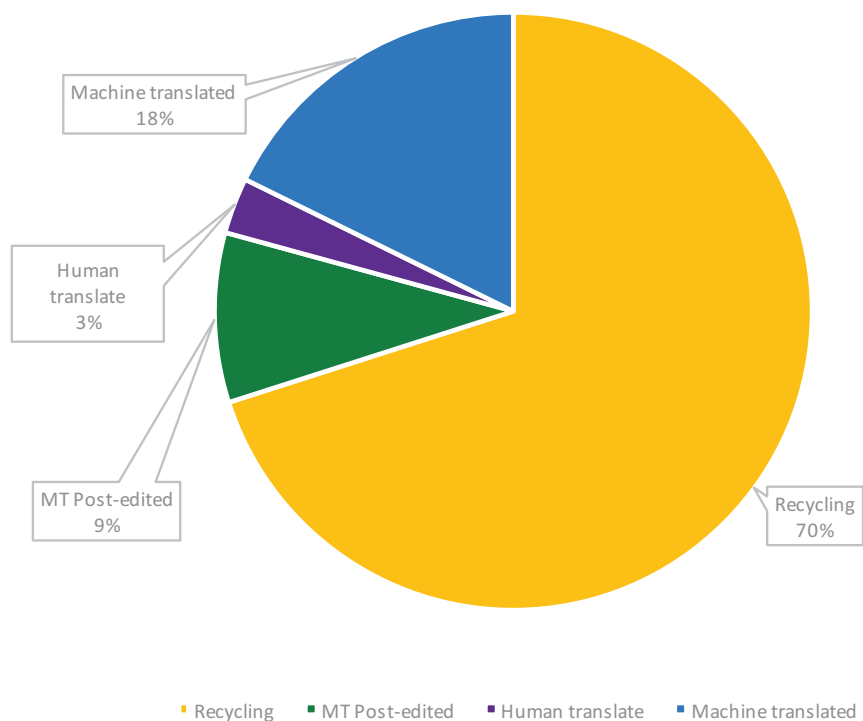


- CSAT based on Article ratings, question: 'Was this information helpful to you?' (yes/no)
- Grouped by thresholding level: none vs 50% vs 80% recycle rate
- MT within 10% of HT for all languages, except Portuguese (11%)



Summary, Lessons & What's next

Support.Office.COM word-count distribution by translation type, for FY16



- MT Publishing can be used at scale, for end-user content
- Recycling helps extend the scope for lower quality MT languages
- Thresholding lets us control unpredictability in MT quality
- User acceptance of MT is greater than offline MT evaluations suggest

Still to do

- Some languages still need work

Coming

- Neural MT
- Experiment: MT Publishing for



Machine Translation Acceptance Among Professional Linguists: Are We Nearing the Tipping Point?

Author

yves@champollion.net
ceo, Wordfast LLC, Paris, France

Abstract

This paper evaluates the level of acceptance of Machine Translation (MT) among translators as of Summer, 2016. Translators are notoriously shy about MT, and their level of acceptance has a lot to say about the level reached by MT.

1. Credits

This text format is derived from the AMTA guidelines for documents submitted for inclusion in AMTA proceedings.

2. MT Among Translators: the Early Days

Translators have been aware of Machine Translation (MT) ever since it first began: the era of automated weather bulletin translation by robots after World War II. But they remained outsiders for a few decades. The Nineties can be seen as the first foray of MT in the world of translators, when the first off-the-shelf solutions (Systran, PowerTranslator, ProMT, etc.) appeared at affordable prices. Still, computer savviness was a prerequisite, which greatly limited the spread of MT use among translators.

3. The Real Start

The nineties witnessed a few earth-shaking changes that redesigned the life of translators for ever, like the widespread acceptance and use of wordprocessors. The spread of the internet is another key factor. While the Internet (and online sales) are credited for the downfall of middlemen like travel agencies, the same factor can be credited for the opposite in translation: the sharp increase in the number of translation agencies. With the ability to team up hundreds of translators and quickly form ad hoc translation task forces for large translation projects, the translation agency, as we know it today, was born.

One collateral effect of agencies attracting the bulk of translation orders was the relentless push toward lower costs. Human translation was considered a relative luxury before the nineties. With the age of luxury for the masses (from smoked salmon to ocean cruises to oversea vacations) came translation for the masses, at least in the business world. Businesses and companies, however small they are, can today call an agency and get 100 pages translated into 20 languages in a week: that was a surhuman feat in the past, available only to very large organizations.

4. Was MT Here to Stay?

The nineties taught us that one low-costing any human activity is how addictive that particular activity becomes. Spend one vacation abroad on an exotic island and presto, you must get back there every year. The addiction of the masses to luxury inexorably leads to a price war, where every trick is authorized. The translation industry is not immune to that effect. Much to the contrary: it is a textbook example of that trend.

As of 2016, we see the maturity of the translation-for-the-masses model. Little more can be added to the existing array of tricks and techniques that agencies use to satisfy a translation-addicted business world.

However, just as air transport or ocean cruises, the translation industry has reached a pricing floor below which it cannot plunge.

Or can it?

5. MT's Age of Reason

Machine Translation is one factor that can prolong the downward spiral of prices in the translation industry. With a twist.

In other industries, when a price war is declared and low-cost is the order of the day, quality is sacrificed first. Summer fruits and vegetables are now available at dirt-cheap prices even in Winter - but they taste like wet cardboard. Senior members of the audience remember a glorious and distant past when tomatoes where had only in season, but Lord did they taste good.

On the contrary, MT's salvation lies in the increase of its quality - because unlike tomatoes, quantity is no issue at all.

As hinted earlier, MT is one of the factors, perhaps the only one, that the translation industry can use to prolong the price war, to keep lowering costs. The human factor cannot be compressed any further - the price paid to translators cannot be significantly lowered at this stage.

The acceptance of MT by linguists lies at the heart of the equation.

6. MT Acceptance Among Translators: Where we Come From

In the 2000's, ten years ago, MT usage was still between low and non-existent among translators, remaining at the level it was in the nineties. A comprehensive research made at London's Imperial College (Elina Lagoudakis, 2006¹) surveyed Translation Memory usage (MT) among translators, in short, technology used in the translation process, as TM was the prevalent technology back then. The expression "Machine Translation" only appears *once* in 36 pages - and actually does not refer to MT as we discuss it. Those were the days! Machine... what?

That does not mean MT was not used before 2010. But it was rare among independent freelance translators. However, translation agencies would sometimes inject MT into pre-processed translation jobs, but that was not a widespread practice, and remained undercover. Linguists raised eyebrows at MT usage, and many would outright refuse; there was still as strong stigma associated with artificial translation. Another survey on translation technology for individual translators (Luciano Monteiro, 2009²) was published by

¹ Elina Lagoudakis, in "*Translation Memories Survey*" published by the Imperial College of London, 2006.

² Luciano Monteiro, in "*Translators Now And Then - How Technology Has Changed Their Trade*", www.proz.com, 2009.

www.proz.com. The article has a belated entry for Machine Translation (stuck just behind "Other useful software"), which mentions that "*when coupled with terminology management, and post-editing services, MT can provide an attractive cost/benefit solution*". MT was still, as late as 2009, seen by translators as a minor, last-resort crutch for those who needed speed.

7. MT Acceptance Among Translators: Today

Hard statistics about translators' habits are hard to come by. My focus here is on translators defined as *individual* practitioners: freelance translators, and employed translators who have a say in their workflow. The reason statistics are hard to collect is that translators are very scattered. The profession is atomized into individual, isolated, practitioners.

To make things more difficult, agencies are shy about revealing their real practices, the technology they use, their prices. Prying reliable information out of translators and agencies is not easy, and will certainly be obsolete in a short few years.

One category of translation tools is the online Computer-Assisted Translation (CAT) tool: a browser-based alternative to the classical, installed CAT tool that translators love to hate. The online CAT tool is on the upswing, especially among two classes of translators: the younger generation, and translators in emergent markets. Whence a precaution about the following figures: the surveyed population is not characteristic of the entire population of translators, as of 2016. But biology and economics being what they are, that young and emergent population will inevitably become mainstream.

8. Today in Figures

Statistics are a difficult to handle properly, and can mean just about anything. Stats on the acceptance of MT by translators are difficult to form. We can only formally poll the use of MT among translators; as for acceptance, which is an attitude toward MT, we can only get clues.

The stats below are derived from two formal sources and one informal source. The two formal sources are an online translation tool (Wordfast Anywhere) with a community of 25,000 registered users, and over 3,000 regular users translating for over ten hours every month. In that situation, figures are reliable, as the tool provides detailed stats on the setup, as well as MT consumption, for each connected translator. The other source is derived from an installed tool (Wordfast Classic and Wordfast PRO), and the associated hotline, which registers the nature of hotline calls, and therefore has a good overview on MT usage. The last source, an informal one, is the speaker's personal experience as a former translator and project manager, a trainer, a CEO in the translation industry, and a CAT evangelist. While not incorporated in the figures, that experience was used to perform sanity checks on the figures, and to offer an interpretation of the figures.

A. MT usage from paid sources

Paid sources are basically subscription-based MT providers, the ubiquitous ones being Microsoft Translator and Google Translate, but there are others, like iTranslate4.eu. We should note that most paid sources cost literally nothing per month for a typical freelancer's consumption: about the price of a good beer. Still, the need to fill a form and provide credit card details ensures that users are 1. indeed professional translators, and 2. deliberately opt for MT.

The stats here is: 15% of translators use a paid source. The statistics in Wordfast Anywhere use IP numbers to track the approximate location of translators, and it appears that most of those using paid MT are in Europe (45% of the grand total), followed by North America (30%). The rest is evenly distributed around the world.

Stats in installed tools use email addresses, language code, and hotline call records to estimate location, and they concur with the above figures.

B. MT usage from free sources

Free MT sources are numerous; we may quote WordLingo³ and MyMemory.

The stats here is a whopping 45%. The figure is evenly distributed among all regions. The figure means that nearly half of all translators regularly use a free MT provider.

I will let everyone decide on the figures above, but here are my observations:

- Paid MT versus free MT. 15% translators using paid MT may seem low as of 2016. Note, however, that paid MT is an opt-in (adhering to paid MT is a deliberate act), while free MT is an opt-out (it is active by default, and can be opted out). Many translators run their tool stock. Like most car owners, they rarely open the hood, if ever. Also note that the 15% figure for paid MT usage was under 10% just 18 months ago, which reveals a fast growth rate: it projects into almost 50% by 2020. Free MT, however, remains relatively flat.

- The 15% figure includes an optical illusion, which is typical in statistics, and I will explain it here. Translators upload documents in different formats. Native formats (like DOC, PDF) are markers of an independent translator, dealing direct with clients; while pre-processed formats (XLIFF, TXML, and generally speaking, XML-based formats) indicate that the document was processed ahead of the translator, by a translation agency or corporate translation department. In that case, it is common, almost a rule, that those formats had MT injected at pre-processing time, in which case MT does not appear in the figures above. If high-tech is used ahead of translation, it is likely that artificial translation was used. Well over 75% of pre-processed formats are injected with a mix of Translation Memory and Machine Translation, with MT being more frequent than TM. With pre-processed formats making up nearly one half of the documents today, the real figure of MT use among translators, thus corrected, is above 20%.

9. Conclusion

We can safely say that Machine translation is now mainstream among translators. Concerning fully independent translators, the trend is still modest, but really present, and it is growing fast. As the younger generation steps in, and the emergent economies further develop, that trend can only intensify.

Translation used to be a luxury at prohibitive costs. It is now used at all levels in business and institutions. Two curves are predicting the advent of widespread use and acceptance of MT at all levels of translation: one is the curve drawn by the need to lower costs in mass translation, the other is the slowly but steady rise in MT quality.

³ WordLingo is not a free MT provider, but it is free for Wordfast users due to a special deal.

What Can We Really Learn from Post-editing?

Marcis Pinnis
Tilde, Riga, Latvia
Rihards Kalnins
Tilde, Riga, Latvia
Raivis Skadins
Tilde, Riga, Latvia
Inguna Skadina
Tilde, Riga, Latvia

marcis.pinnis@tilde.com
rihards.kalnins@tilde.com
raivis.skadins@tilde.com
inguna.skadina@tilde.com

Abstract

This paper describes the findings of a large post-editing project in the medical domain carried out by Tilde. It analyzes the efficacy of post-editing highly technical texts in a specialized domain, and provides answers to questions important to localization service providers that are considering the introduction of post-editing in their translation workflows. The results show that by carefully analyzing post-editing projects, machine translation providers and language service providers can learn how to boost productivity in localization, save time and optimize resources in the language editing process, as well as leverage post-edits to improve machine translation engines through dynamic learning.

1. Introduction

In order to analyze the efficacy of post-editing highly technical texts in a specialized domain, Tilde embarked on a project to analyze a large post-editing effort in the medical domain. During the project, Tilde had the unique opportunity to take detailed logs of each activity performed by post-editors. Tilde then analyzed the post-editing results, allowing us to answer important questions like: (1) How effectively do post-editors really work with machine translation (MT)? (2) Do post-editors expend their efforts usefully on editing MT results? (3) How can MT be improved to meet the needs of localization companies that utilize post-editing to boost translation productivity? (4) How does the MT quality affect post-editing performance?

2. MT System

During the course of the project, post-editors used a statistical MT (SMT) system that was based on the phrase-based Moses SMT system (Koehn et al., 2007). The system was trained on the European Medicines Agency (EMA) parallel corpus from OPUS corpus (Tiedemann, 2009) and latest documents from EMA website (years 2009-2014) collected by Tilde on the Tilde MT platform (Vasiljevs et al., 2012). The statistics of the training corpus before and after filtering are given in Table 1. The system's automatic evaluation results are given in Table 2.

Corpus	Sentences before filtering	Sentences after filtering
Parallel	378,869	325,332
Monolingual	378,869	332,652

Table 1: Statistics of the training corpora used to train the SMT system

Evaluation scenario	BLEU	NIST	METEOR	TER
Case sensitive	47.42 (45.82-48.88)	9.5300 (9.3469-9.7027)	0.3637	0.3952
Case insensitive	45.79 (44.23-47.26)	9.2735 (9.1036-9.4539)	0.2575	0.4105

Table 2: Automatic evaluation results of the SMT system

3. Post-editing Task

The post-editing task was performed using the tool PET (Aziz et al., 2012), which is able to precisely track the time spent on each segment and all keystrokes that a post-editor performs while post-editing each segment. An example of the graphical user interface of PET as used in the post-editing task is given in Figure 1. The whole post-editing task, which contained 22,500 (360,000 words) sentences, was split into jobs that consisted of 100 sentences. All jobs contained consecutive sentences from the latest documents. All jobs were pre-translated with the SMT system prior to giving the jobs to post-editors, so that translators would not have to wait for the SMT suggestions to appear.

While post-editing texts, translators were asked to evaluate the quality of each MT suggestion, marking it as one of the following: “near perfect,” “very good,” “poor,” and “very poor.” If the translator did not apply changes to the MT suggestion, the post-editing tool automatically rated it as “unchanged”, which means that the MT suggestion was perfect and did not require any post-editing.



Figure 1: Example of the user interface of the PET post-editing tool showing: 1) the MT suggestion; 2) previous context; 3) the source text of the segment that is being post-edited; 4) the target editing field (showing the MT suggestion); 5) the further context; and 6) the entries of the term collection that are found in the source text of the current segment

A total of five professional translators worked on the post-editing project full-time for approximately five weeks in total. The post-editors were also asked not to spend excessive amounts of time on each segment, as the quality expectations were not “human translation quality” but rather “post-editing quality.” To assist post-editing, post-editors were provided

with an in-domain term collection that was integrated in the post-editing tool and automatically showed translation suggestions for known terms.

The detailed logs of each translator’s work recorded the timing of each keystroke, measuring the time spent on post-editing in three distinct intervals: the amount of time that elapsed between the appearance of a MT segment and the first click, or “reading time”; the amount of time between the first edit and approving the segment, or “editing time”; and the amount of time spent between approving the segment and completing the assessment of the quality by clicking the “Finish” button, referred to as “assessment time.”

4. Preliminary Results

The results showed that the use of custom MT resulted in a considerable boost in overall translation productivity (see Table 3). The translators’ average translation speed for human translation of medical domain texts is approximately 800-900 tokens per hour (pure translation time, not counting pauses between sentences). But MT succeeded in boosting the average translation productivity to 2,694 tokens per hour – an approximately 200% increase.

This strong boost in productivity came about thanks to the high translation quality of the MT system used by post-editors (BLEU score of 47.42). The analysis showed that the MT system produced a majority of MT segments – over 37% – that were marked “unchanged,” demanding no editing time at all from the post-editors.

MT suggestion assessment	Total editing time	Total source length in tokens	Segment count	Productivity (tokens post-edited in one hour)
0. Unchanged	14:15:43	83,661	5,488	5,865
1. Near perfect	12:12:01	46,108	2,458	3,779
2. Very good	44:11:31	102,309	4,962	2,315
3. Poor	26:40:50	37,956	1,717	1,422
4. Very poor	04:13:39	3,582	175	847
Grand Total	101:33:46	273,616	14,800	2,694

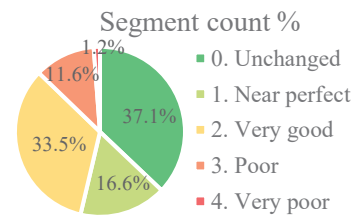


Table 3: Sum of editing time and productivity gains

Not surprisingly, the average amount of post-editing time for each segment rose incrementally as the quality of the MT result (as marked by the post-editor) declined (see Figure 2). A “near perfect” segment had an average of 19.23 seconds of post-editing; a “very good” segment had an average of 34.80 seconds; a “poor” segment demanded 59.55 seconds of the post-editor’s time; and a “very poor” segment clocked in at 90.98 seconds of work.

Though the editing time grew incrementally, the total “reading time” on the part of post-editors grew more gradually. Even an “unchanged” segment demanded an average of 10 seconds of reading time from post-editors. This figure should be kept in mind by localization companies that want to increase their MT use: even an unchanged segment, with perfect MT quality, demands 10 seconds of a post-editor’s time for review.

The most surprising results of the study came when graphing the relationship between the quality of an MT suggestion and the PE quality of the segment *vi-à-vis* a reference human translation (see Figure 3). We found that segments with “near perfect” or “very good” MT quality led to the creation of post-edited texts with Translation Edit Rate (TER) scores that ranked fairly consistently in relation to the reference translation. For instance, an MT segment with “near perfect” quality produced, on average, a post-edit that had TER scores of 0.22-0.30 in relation to the reference translation.

However, when MT produced “poor” and “very poor” segments, the post-editing quality was furthest from the reference translation. Post-edits of “poor” and “very poor” segments could have TER scores that ranged from 0.50 to 1.00. One possible reason for the wide range in scoring was that the target language in the project was a morphologically rich, highly inflected language with relatively free word order. Therefore, thanks to the liberal syntax, post-editors had a wider range of options for constructing post-edited sentences. This led to grave inconsistency in results for post-edits that demanded the most efforts, namely, edits of “poor” and “very poor” segments.

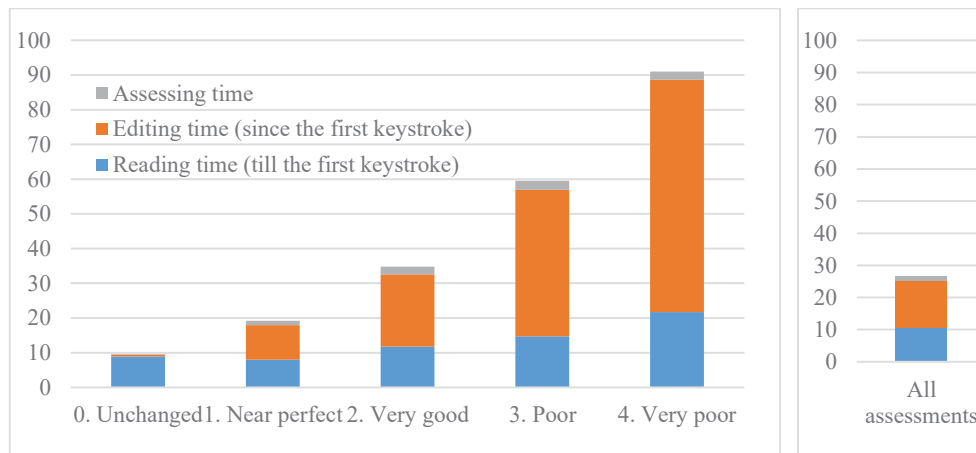


Figure 2: Average of reading, editing, and assessment times for segments with different MT suggestion quality assessments (left) and all segments (right)

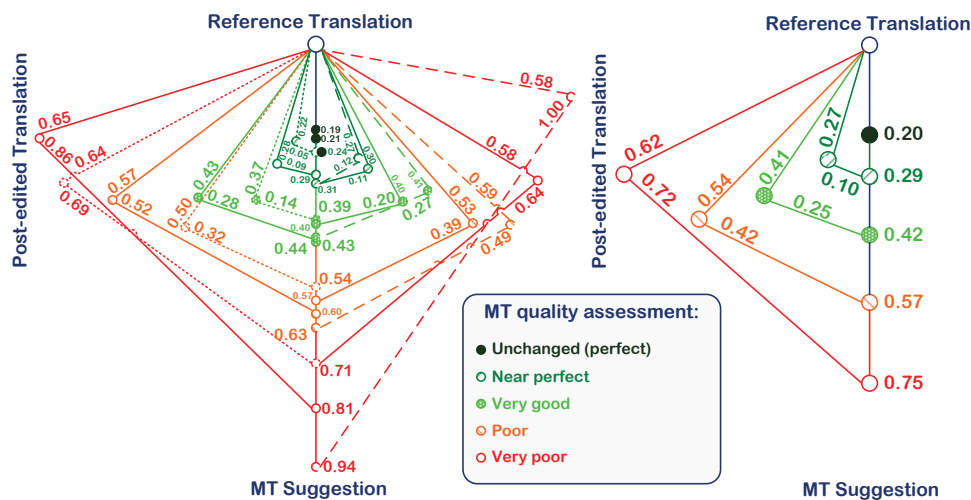


Figure 3: Translation edit rates between reference translations, MT suggestions, and post-edited translations (left – for 4 individual translators, right – for all translators) depending on different MT suggestion quality assessments

Inconsistent post-editing quality poses a serious problem for language editors and linguists at localization companies, who must expend extra effort in establishing linguistic uniformity in a text. This problem is further compounded for post-editing projects that utilize multiple post-editors.

5. Key Findings

By carefully tracking the work of post-editors during this project, we were able to make four concrete findings:

- Post-editing resulted in a huge increase in translation productivity (200%)
- The amount of time spent on post-editing a segment is, on average, directly proportional to the quality of the MT
- High quality MT results lead to relatively consistent post-editing quality
- Poor quality MT leads to a high degree of inconsistency between post-editing quality and a perfect human translation

6. Conclusion: What Does This Mean for Language Service Providers?

Most importantly, however, these results have allowed us to offer several recommendations for localization service providers utilizing MT in the post-editing process.

First, post-editing projects must be carefully tracked in the CAT tool environment. By asking post-editors to rank the quality of a segment – an activity that takes up only a fractional

interval of editing time, approximately 3% – much insight can be gained and then applied to the final language editing process.

As many segments in post-editing will remain “unchanged” or just slightly changed – that is, if the quality of the MT system is high – language editors who are alerted to the quality of segments can safely accept these post-edits without expending any additional language editing efforts. However, taking into account the great inconsistency in post-editing of “poor” and “very poor” segments, language editors should expend extra effort on editing these segments in order to ensure that they conform to the overall stylistic quality of the text.

Second, the results of the finding also illuminate the ways in which Dynamic Learning can improve MT quality. MT results that are marked “near perfect” and “very good” produce relatively high quality post-edits, therefore these post-edits can safely be used to dynamically improve MT engines through the Dynamic Learning function. However, inconsistent post-editing quality, as produced from “poor” quality MT can severely pollute the quality of a MT system and should be removed from dynamic improvement to the MT engine.

Therefore, logging a post-editor’s quality assessment or editing time of MT suggestions can also help improve the quality of engines with Dynamic Learning. LSPs and their MT vendors should only allow post-edited segments of “near perfect” and “very good” quality to be used to dynamically improve the underlying MT engine.

By carefully analyzing post-editing projects, MT providers and LSPs can learn how to boost productivity in localization, save time and optimize resources in the language editing process, as well as to leverage quality post-edits to improve MT engines through Dynamic Learning. Only in this way will LSPs be enabled to meet the booming volumes of translation – up to a 67% increase (Lommel, 2016) – that they can expect from enterprises in the next few years.

Acknowledgements

The work within the QT21 project has received funding from the European Union under grant agreement n° 645452.

References

- Aziz, W.; Sousa, S. C. M.; Specia, L. (2012). PET: A Tool for Post-editing and Assessing Machine Translation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey. May 2012.
- Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., ... Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions* (pp. 177--180), Stroudsburg, PA, USA: Association for Computational Linguistics.
- Tiedemann, J. (2009). News from OPUS-A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing* (Vol. 5, pp. 237-248).
- Vasiljevs, A., Skadiņš, R., & Tiedemann, J. (2012). LetsMT!: A Cloud-Based Platform for Do-It-Yourself Machine Translation. In *Proceedings of the ACL 2012 System Demonstrations* (pp. 43–48), Jeju Island, Korea: Association for Computational Linguistics.
- Lommel, A., DePalma, D. (2016). “Europe’s Leading Role in Machine Translation: How Europe Is Driving the Shift to MT.” Common Sense Advisory Report. <http://cracker-project.eu/csa-mt-report/>

An Empirical Study:

Post-editing Effort for **English to Arabic** Hybrid Machine Translation

Hassan Sajjad, Francisco Guzman, Stephan Vogel
Qatar Computing Research Institute, HBKU

Introduction

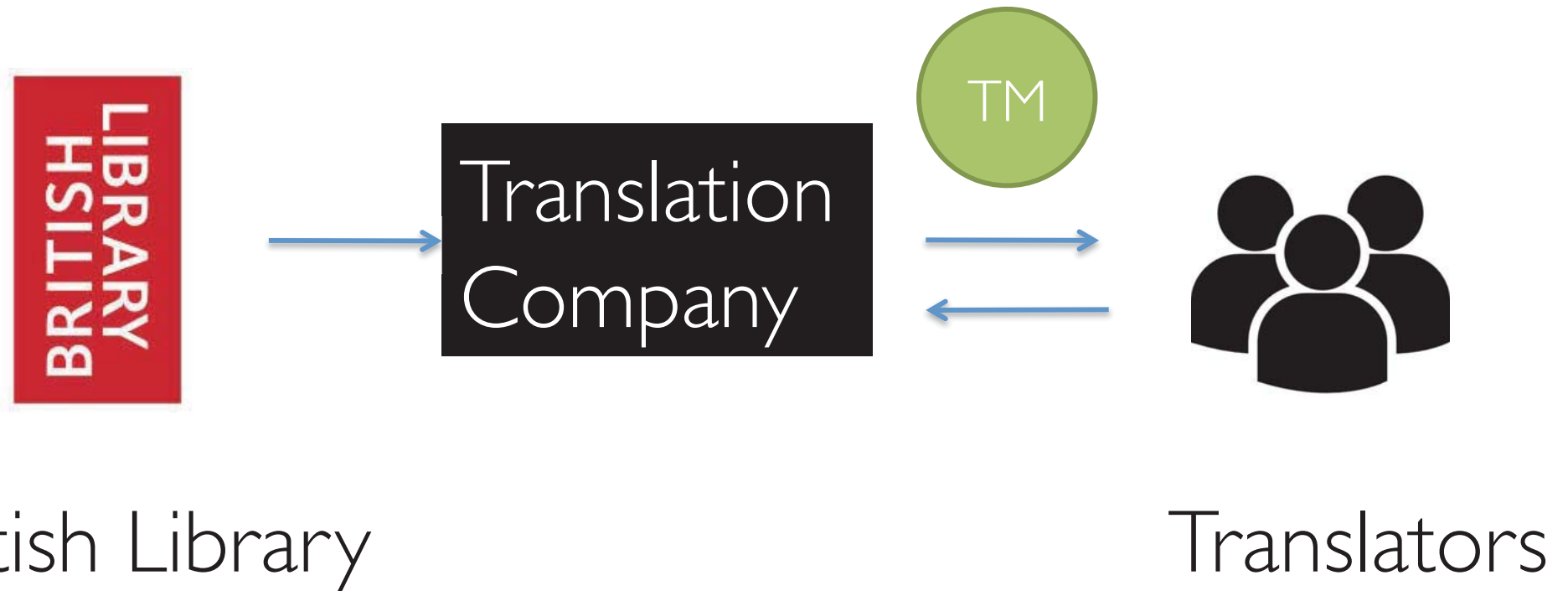
- Old Arabic documents



- Translation of metadata from English to Arabic



Traditional Translation Process

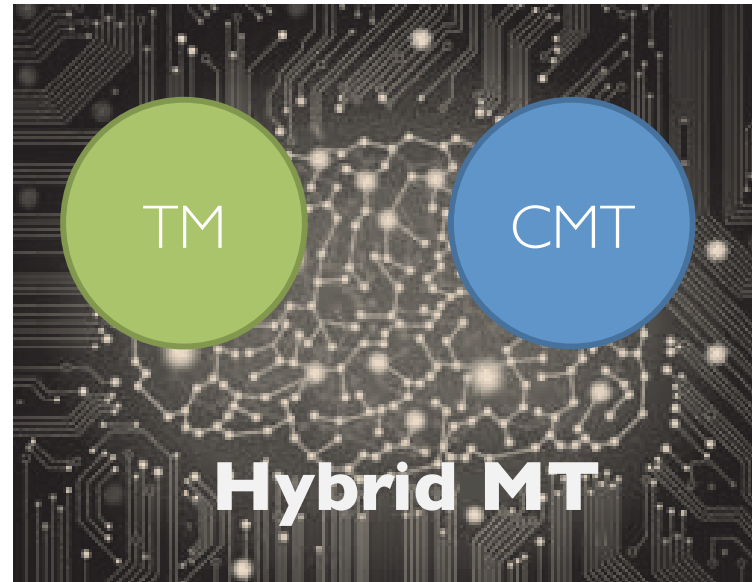


Problem

- Few translation memory matches
 - A lot needs to be translated from scratch
- Time and cost inefficient

Solution: Hybrid Machine Translation

High precision translations



100% recall – readily available translations

Hybrid MT: Combines the benefits of both!
Translation Memory and Customized MT

Hybrid MT System



Translation Memory

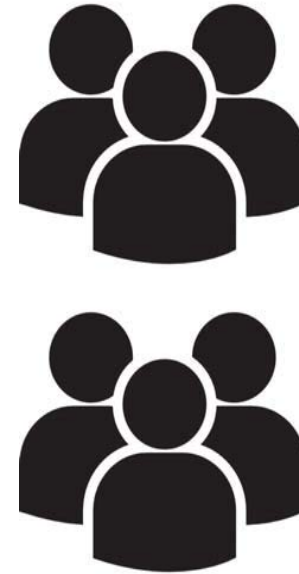
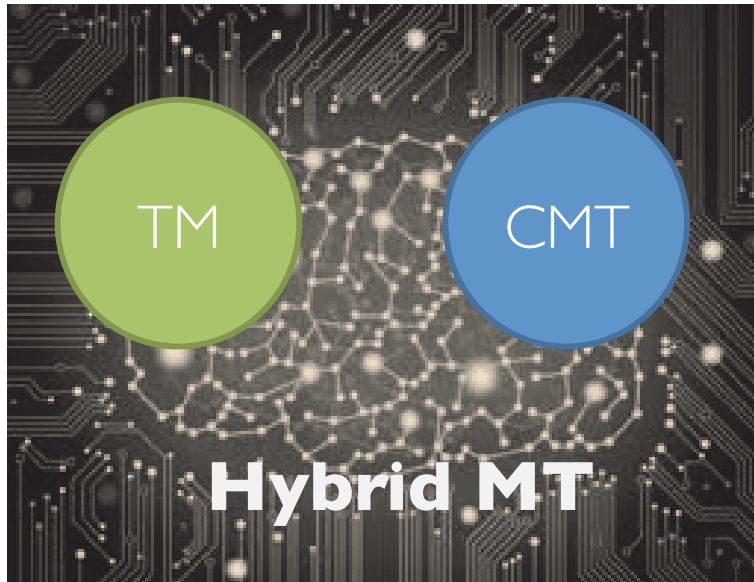
- First pass: use strict matching to translate known words and phrases



Customized Machine Translation

- Second pass: translate the remaining text using machine translation system

Aiming higher: Post Editing for Quality



Post Editors

- High quality
- High consistency
- Cost and time effective



Customized Machine Translation

- A statistical machine translation system
 - Train specific to the domain of the text that needs to be translated
- General practice
 - Use Moses
 - Train on the data of translation memory
 - Follow recipe of a competition grade system to ensure high quality



English to Arabic CMT

- Best competition grade pipeline involves
 - Arabic (de-) tokenization
 - Splitting morphologically rich words into smaller segments and vice-versa
 - +2.5 BLEU points improvement
 - Arabic (de-) normalization
 - Mapping different forms of a letter to one form and vice versa
 - +0.5 BLEU point improvement

This ensures high quality but **does not guarantee less frustration for post-editors**



Why?

Translation output requires:

- De-tokenization and de-normalization
- De-normalization introduces character-level errors
 - Frustrating for the post-editor to correct
 - Time inefficient



Recommended Practices for CMT of English-Arabic

- Don't normalize

But

- Always tokenize
 - Improve coverage of words
 - Better translations

Let's Talk about BL Case Numbers!

We compare:

- Translation Memory (TM) only
- Hybrid MT (TM + CMT)

Looking at:

- Effectiveness
- Quality
- Consistency

Also:

- Translator
- Hybrid MT + Post editing (PE)

Effectiveness of TM

Exact match

50%
segments

BUT
COVERS
ONLY

7%
words

Fuzzy match

84%
segments

BUT
COVERS
ONLY

13.5%
words

More than 85% of words still need to be translated !!!!

* Based on an assessment over X documents

Effectiveness of CMT

100% AND 99.9%
segments words

translated!

Effectiveness of Hybrid MT

- High precision
 - TM exact matches
- High recall
 - CMT to produce high quality translations

Assessing Quality

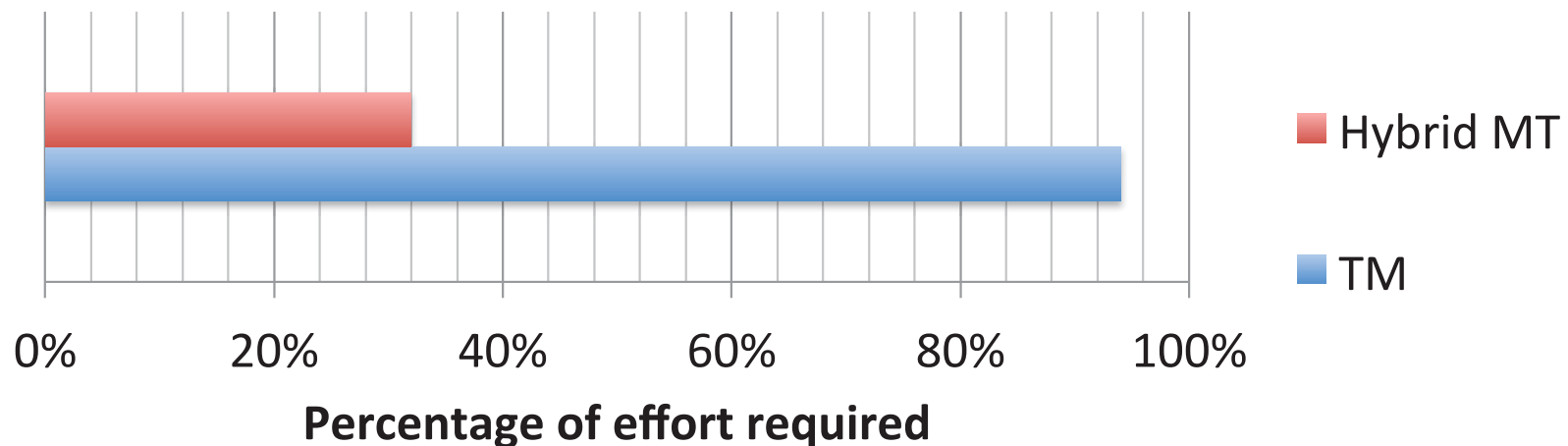
- BLEU
 - Compare output to ‘reference’ translation

	Strict	Partial
TM	7.07	21.01
TM + CMT	54.60	48.54

CMT alone BLEU scores are 53.90

Assessing Quality

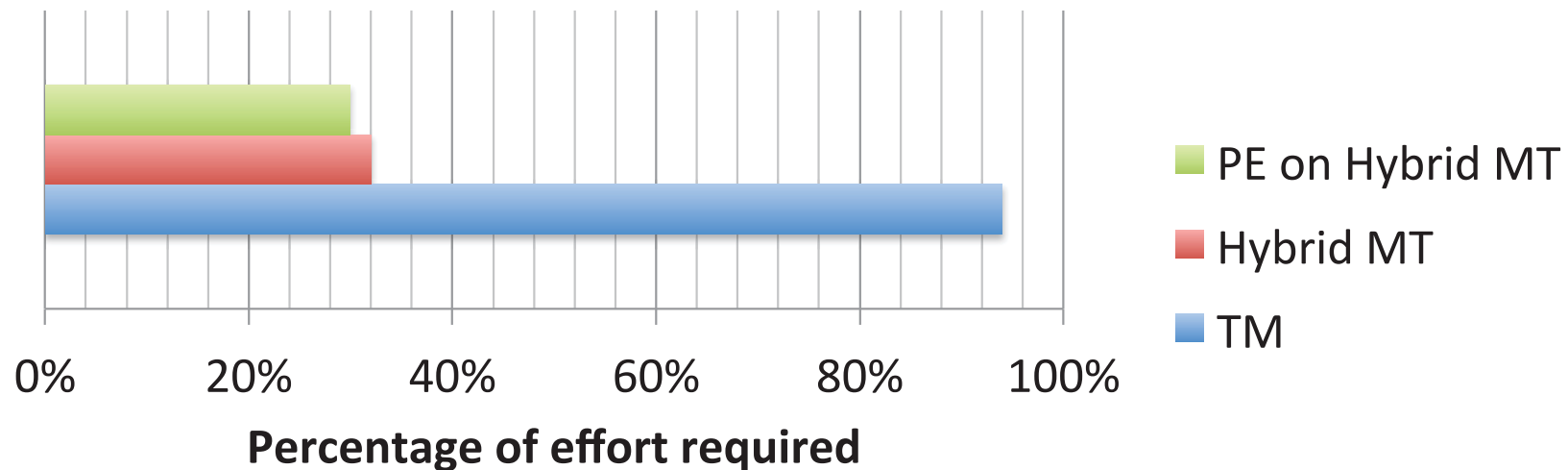
- TER: Translation Error Rate
 - How much effort is needed to get perfect translation
 - Compare Hybrid MT output to 'reference' translation



Hybrid MT can improve beyond that!!!

Assessing Quality

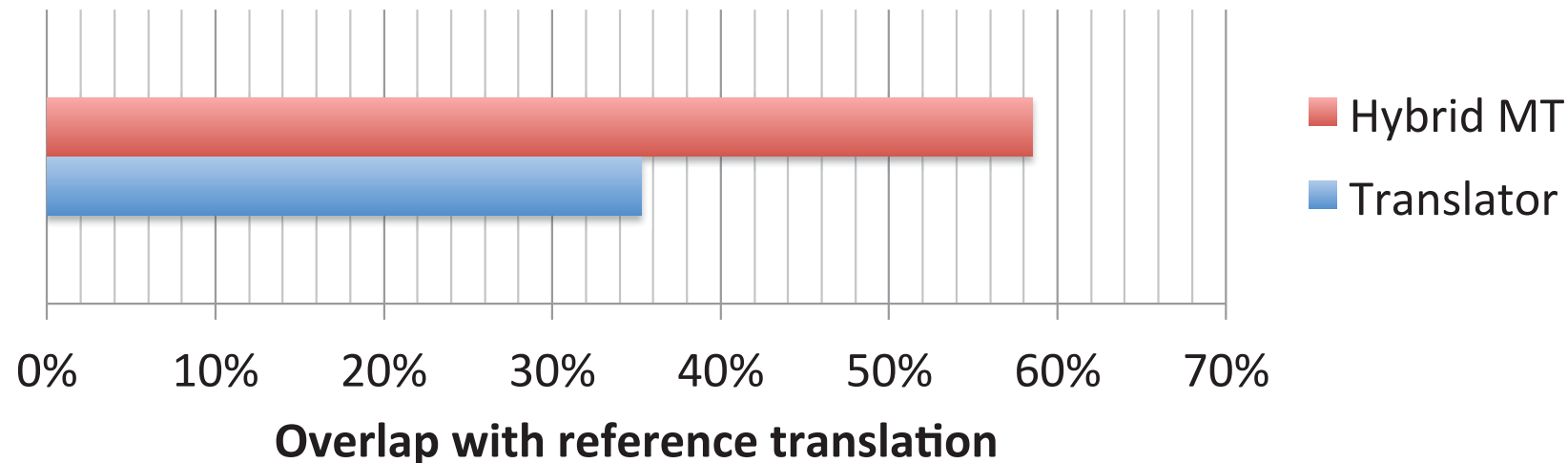
- TER vs. Post editing effort
 - Similar effort estimation using post-editing of Hybrid MT



* PE is based on an assessment over 4 documents, using a junior translator

Consistency of Hybrid MT

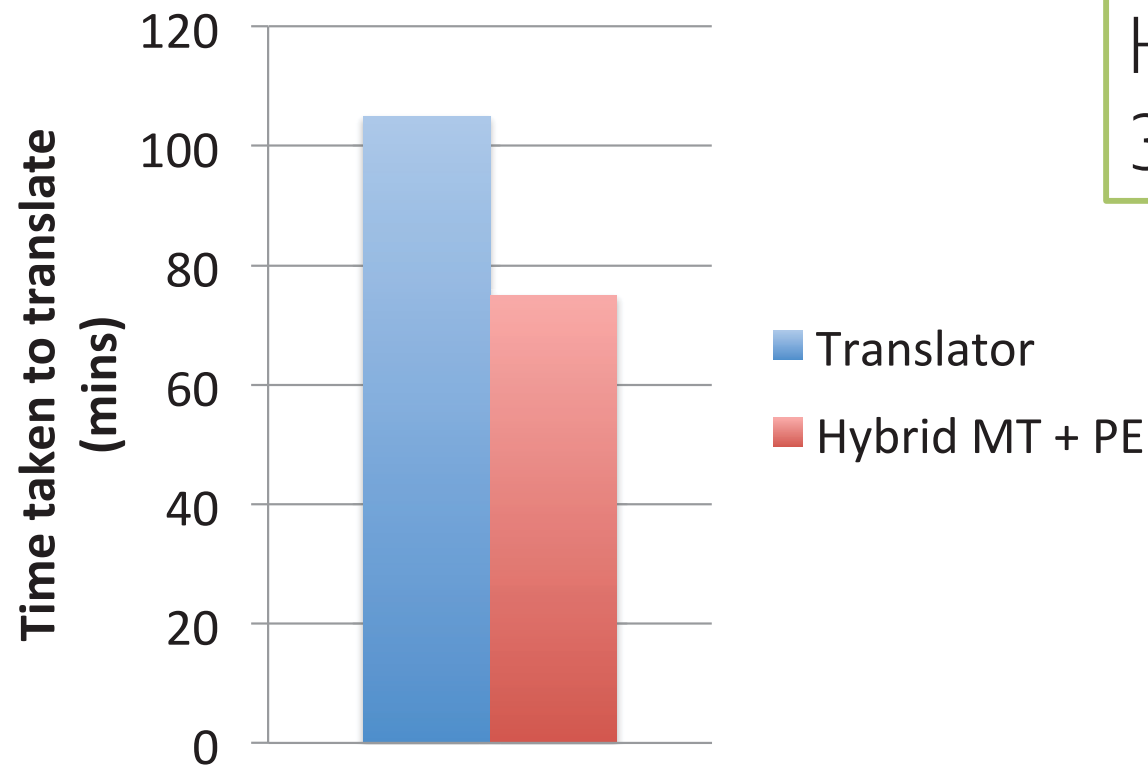
- We compared Hybrid MT versus a junior translator
- We measured consistency with reference translations



Hybrid MT is more consistent with reference translations

Speedup of Hybrid MT

- We compared Hybrid MT versus a junior translator



Hybrid MT+PE is 30% more efficient

Conclusion

- Hybrid MT
 - High precision and high recall
- Hybrid MT plus Post-editing
 - Efficient in terms of both time and cost
 - Improved consistency

References

- Ahmed Abdelali, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak. Farasa: A Fast and Furious Segmenter for Arabic. In NAACL-2016, San Diego, US.
- Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In ACL-2007, Prague, Czech Republic
- Hassan Sajjad, Francisco Guzman, Preslav Nakov, Ahmed Abdelali, Kenton Murray, Fahad Al Obaidli, and Stephan Vogel. QCRI at IWSLT 2013: Experiments in Arabic-English and English-Arabic Spoken Language Translation. In IWSLT-2013, Heidelberg, Germany

Divide and Conquer Strategy for Large Data MT

Dimitar Shterionov
KantanLabs, KantanMT, Dublin, Ireland

dimitars@kantanmt.com

Abstract

In recent years Statistical Machine Translation (SMT) has established a dominant position among the variety of machine translation paradigms. Industrial Machine Translation computer systems, such as KantanMT, deliver fast and of high performance SMT solutions to the end user. KantanMT is a cloud-based platform that allows its users to build custom SMT engines and use them for translation via a batch or an online mode. In order to employ the full potential of the cloud we have developed an efficient method for asynchronous online translation. This method implements a producer-consumer technique that uses multiple queues as intermediate data storage units. Furthermore, each queue is associated with a priority that defines how quickly the queue can be consumed. That gives our users the control on the flow of translation requests, especially when it comes to large amounts of data.

In this paper we describe the design and the implementation of the new method and compare it to others. We then assess the improvement in the quality of service of our platform by empirical evaluation.

1 Introduction

In recent years Statistical Machine Translation (SMT) (Koehn (2010); Brown et al. (1993)) has established a dominant position among the variety of available machine translation paradigms. In 2007 Moses (Koehn et al. (2007)) was released – an open-source toolkit for SMT. While research efforts have been mainly focused on improving the core SMT technology, i.e., Moses and related pre- and post- processing techniques, we focused on bringing this technology to the end user in a highly scalable manner. Our MT platform, KantanMT¹ is fully distributed on the cloud. In order to employ the full potential of the cloud and provide to our users high quality translations for large amount of data with low response time, we have developed an efficient request handling system.

The system aims to optimize resource allocation and to improve the robustness and resilience of our platform as well as the quality of service (QoS). We analyse two approaches for processing translation requests in a distributed MT environment that are already employed in our online translation pipeline and compare them to our new method.

The first approach processes each request at the moment it has been received using a centralized *https* endpoint. According to this method the segment, sent via the request, is translated by a predefined SMT model and returned back to the user. Underlying is a load balancing mechanism to distribute the segments on a fleet of servers that are dedicated for translation. This method is synchronous (we refer to it as *SYNC*) and thus can introduce high delays between sending a request and receiving a translation.

Our second method implements an efficient producer-consumer technique based on a single intermediate data storage unit (IDSU). While it allows users to translate large amounts of

¹<https://kantanmt.com/>

segments asynchronously (thus we refer to it as *ASYN*C), it is still bound by the limits of the used IDSU. According to the *ASYN*C method, old requests need to be processed before new ones. Thus, it may cause delays for new requests.

The new method we have developed improves upon these restrictions by distributing requests in as many IDSUs as the user requests. It allows requests that arrive at different time points to be processed in *parallel* and *independently* from each other. Furthermore, this method adds an extra control layer that allows our users to assign priority to their requests – the higher the priority, the faster the requests are processed. That is, we take under consideration that some requests may be more important to our users than others, and as such would need to be handled faster. We refer to this method as the *PP-ASYN*C.

In this paper we describe the design and the implementation of the *PP-ASYN*C method and compare it to the other two. We then assess the improvement in the QoS of our platform by empirical evaluation.

2 Online translation with KantanMT

KantanMT (<https://kantanmt.com>) is a cloud-based SMT platform that provides machine translation services to its clients for more than 760 language pairs. It is based on the state-of-the-art Moses toolkit to train SMT models; these models are then used for decoding. In the remaining of this paper we use the term *KantanMT* or *SMT “engine”* to refer to the collection of SMT models and configuration files. The use of the Moses toolkit together with the distributed architecture of the system allows KantanMT engines to be built at very high speed and with low computational cost.

KantanMT platform is equipped with two translation modes – *batch* mode and *API* or *online* mode. In the batch mode clients provide a set of documents². The system then translates each of these files (one after the other) and returns their translated versions to the client.

The focus of this work is on the methods used in online translation mode. The online translation mode allows clients to send translation requests via the KantanMT API using an HTTP GET or an HTTP POST method. Each request specifies user identification (under the form of a KantanAPI token), the engine to be used for translation and the segment to be translated. In response to a given translation request, the user will receive a translation of the input segment. In the rest of this section we describe the three online translation methods. We focus on the *PP-ASYN*C method which is the most recent and innovative online translation method of KantanMT.

2.1 Synchronous online translation

*SYN*C is the most basic online translation method of KantanMT. According to this method the user sends a request for translation of one (or more) segments via the API call `translate` and receives either the translation of the segment or a failure notification. In the request (both HTTP GET and HTTP POST requests are supported) the user provides its unique API token, the machine translation (MT) profile to be used for translation and, finally, the segment. Example 2.1 shows an HTTP GET request for the *SYN*C method.

Example 2.1 An HTTP GET request from a user with token `1234567890123456`³ for translation of the segment “Welcome to our blog ‘100% Machine Translation’.” with the MT profile `MT-en-bg`.

<https://kantanmt.com/api/translate/1234567890123456/MT-en-bg/Welcome to our blog '100%25 Machine Translation'>. △

²KantanMT supports 26 file formats.

³This is an example token.

After using the token to verify the user, KantanMT translates the segment(s) using the specified MT profile (if such exists and is running). The response of the `translate` call to each translation request is either a successful translation or an error notification (in case a failure due to failed authentication, incorrect encoding, etc. has occurred).

In Figure 1 we show the architecture of the system for handling SYNC requests with respect to message-passing, i.e., the flow of information from the user to the platform and back. Although we refer to this method as synchronous, it has an underlying distributed architecture that allows several requests to be processed in parallel (i.e., several segments to be translated in parallel). This we achieve by using two load balancers (LB) – one to receive and distribute the requests and a second one that distributes the segment for translation among different machines.

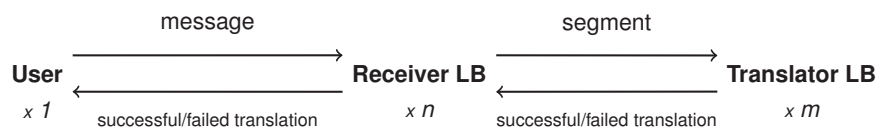


Figure 1: Message passing in SYNC.

2.2 Asynchronous online translation

Despite the capabilities of the LBs in the SYNC method to parallelise requests, thus letting multiple segments to be translated simultaneously, users still need to wait until one batch of segments is translated before they can submit new ones. This has three main drawbacks: (i) lower effectiveness of our platform – i.e., lower quality of service – enforcing users to manually operate the submission process; (ii) a large amount of requests submitted at the same time may overload the LBs and cause huge latencies and (iii) concurrent users may not be able to process their requests in parallel.

In order to tackle the aforementioned problems we introduce an additional layer of parallelism that allows users to submit huge batches of messages⁴. This method uses a queue as an intermediate data storage unit and processes translation requests asynchronously; we refer to this method as the *ASync* method.

The architecture of the *ASync* system implements a producer/consumer-based (Arpaci-Dusseau and Arpaci-Dusseau (2015)) approach where incoming requests are stored in a queue. A *consumer* then depopulates the queue sending each read segment for translation. To translate a segment, we use the SYNC method. If the translation succeeds it is forwarded to the user in the form of a notification sent to a user-defined endpoint. If the translation fails, the segment is redriven for a new translation attempt; if the number of failed attempts exceeds a predefined threshold a failure notification is sent to the user on a user-defined endpoint. We present the architecture of the *ASync* system in Figure 2.

3 Parallel priority-based online translation

The *ASync* method allows users to deal with large amount of requests asynchronously and receive response from our system when a translation is generated, or when a failure occurs. However, often a user may decide to, e.g., delay some translations in order to have other segments processed sooner, e.g., messages for a job that is close to a deadline need to be processed faster than others that are not urgent. When it comes to large volumes of data prioritization can be crucial for the on-time service delivery.

⁴In practice, we do not set any restriction on the number of requests that can be submitted at once.

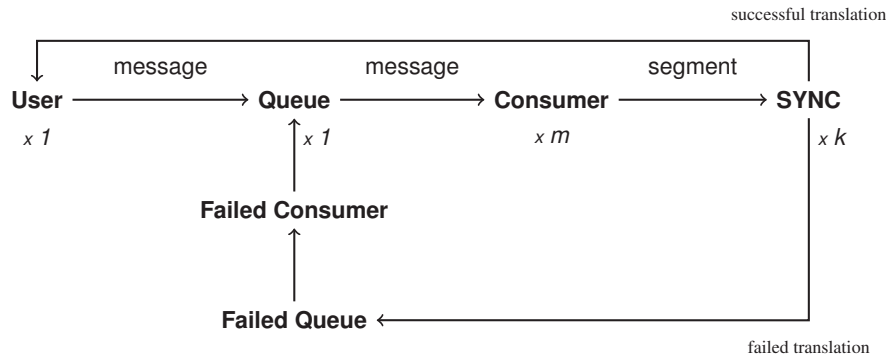


Figure 2: Message passing in ASYNC.

The ASYNC method is bound by the limits of the used data structure, i.e. the queue – each message is processed in the order that it is pushed into the queue: old messages already in the queue need to be processed before new ones. Such a scenario is given in Example 3.1.

Example 3.1 Consider that User A sends 2000000 requests at time T_0 . Next, consider that at time $T_1 = T_0 + 10$ minutes 10000 requests have been processed. That is, the consumption ratio is 1000 requests/minute. At time T_1 User B sends 10000 requests (stored at the end of the queue). While the 10000 requests of User B would require 10 minutes for processing – i.e., to translate the segments they carry – this user will need to wait for 1990 minutes = 33 hours and 10 minutes before they can be processed. \triangle

3.1 Method description

In order to provide to our users a mechanism to control the priority of their requests as well as to enforce parallelism we devised the *Parallel priority-based ASYNC*, (PP-ASYNC). The PP-ASYNC method uses multiple queues, each associated with a priority and a name. Each incoming translation request contains the priority with which it should be processed and the name of the queue in which it should be stored.

Each queue has a priority and is consumed at a predefined ratio associated with that priority. The consumption ratio defines how quickly a queue is read. The higher the queue priority, the greater the consumption ratio. Once a request is read from a queue the segment that it carries is sent for translation with the SYNC method.

Successful translations are then sent back to the user (via a success notification endpoint provided by the user). Failed translations are redriven into the correct priority queue for new translation attempt; if the number of translation attempts exceeds a predefined threshold then the segment is failed and the user is notified of the failure (via a failed notification endpoint).

The implementation of the PP-ASYNC method extends the implementation of the ASYNC method with: (i) a distributing mechanism that processes incoming requests and redirects them to the queue that matches the specified priority and queue name; if such a queue does not exist it is created before the request is pushed; and (ii) support for reading messages from multiple. Figure 3 shows the architecture of the system and the information flow.

3.2 Consumption ratio

The distributor (see Figure 3) receives a request from the user and distributes it to the correct queue according to the specified priority and queue name. The consumer (independently from the distributor) iterates over all queues, identifies the consumption ratio of each queue, based

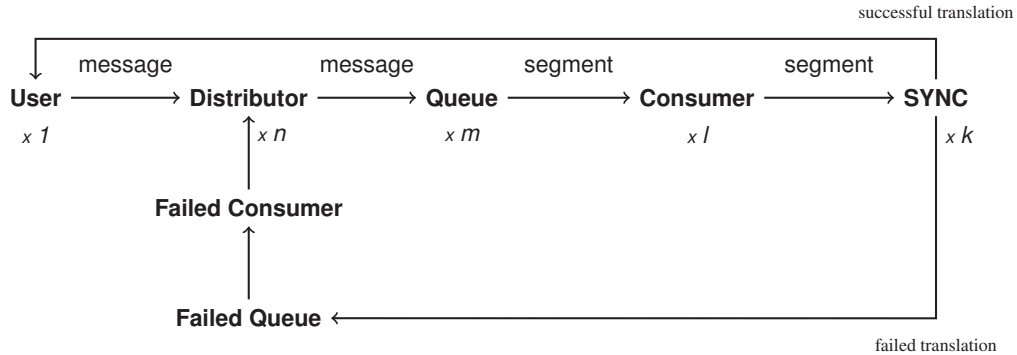


Figure 3: Message passing in PP-ASYNC.

on its priority, reads a number of requests from the queue, sends the segments they carry for translation and proceeds to the next queue. In our definition of the *PP-ASYNC* method, the consumption ratio of a queue defines how many requests are read from that queue at one iteration of the queue consumer. That is, for a queue $Q_i \in \{Q_1..Q_N\}$ with priority q_i there will be c_i requests read. For another queue $Q_j, j > i$ there will be $c_j > c_i^5$ requests that need to be read.

Our implementation allows multiple distributors and consumers to run simultaneously as neither of them has side effects. That is why the actual consumption ratio depends on the consumption ratio c_i of a queue Q_i and the number of consumers (M). For example, if there are two consumers that read simultaneously from a queue with consumption ratio 5 (messages per iteration), then ten requests will be read from that queue. A more detailed example is given in Example 3.2.

Example 3.2 Consider that User A sends 2000000 requests to a queue with priority 1 (Q_1) at time T_0 and that there are no other queues at that time⁶. Assume that the consumption ratio for priority 1 is 100 requests per iteration and the consumption ratio for priority 2 is 500. Also, consider that in 1 minute one consumer can read 100 requests and there are 10 consumers running simultaneously. Then, in 10 minutes, that is, at time $T_1 = T_0 + 10$ minutes 10000 requests will have been processed from Q_1 . At time T_1 User B has just send 10000 requests all of which with priority 2, i.e., they are stored in a queue with priority 2 – Q_2 . The consumers become aware of the new queue and proceed to processing it. According to its priority each consumer will process 500 requests (in one iteration). Given that each consumer reads from a queue at the speed of 100 requests per minutes, at time $T_2 = T_0 + 5$ minutes the consumers will stop reading from Q_2 and start a new iteration. They will consume 1000 requests from Q_1 and at time $T_3 = T_2 + 1$ minute the consumers will move to Q_2 . At time $T_4 = T_3 + 5$ minutes all requests from Q_2 will be read, that is, approximately 11 minutes after they have been submitted. \triangle

To determine the optimal consumption ratio we empirically evaluated different values and compared the system's performance. Our tests showed that the most efficient consumption ratio is the one linear to the priority of the queue. In our implementation we have selected the simplest linear dependency: the consumption ratio equals the priority of the queue. That is, for

⁵If the number of requests that needs to be read from a queue is larger than the number of requests available in the queue, then all of them are read, the segments they carry are sent for translation, and the consumer proceeds to the next queue.

⁶Or all other queues are empty.

a queue Q_i with priority q_i in one iteration of a handler there will be q_i requests read from that queue. We aimed at a balanced consumption ratio where, on the one hand, queues with higher priority are consumed faster than queues of lower priority but on the other hand, lower-priority queues are still processed within reasonable time.

In order to avoid concurrences between different users we allow two queues with the same priority to have different names. In this way, each user can benefit equally from the PP-ASYNC. In addition, we have implemented a queue with *infinity* priority – consumers process messages from this queue until it is empty. Using this queue leads to a exactly the same behaviour as the ASYNC method.

4 Empirical evaluation

We performed a series of tests that aimed to empirically evaluate the new online translation method, i.e., the PP-ASYNC method. There are two objectives that we targeted with our test: (i) compare the performance (i.e., translation speed) of the ASYNC and the PP-ASYNC methods and (ii) show that the new method eliminates/reduces delays for newly incoming messages.

We ran our experiments on a Windows Server 2012 machine with 8-core Intel CPU, 15GB or RAM and 160GB of SSD. For all our tests we used 2 engines – E_1 and E_2 – of different size⁷. Engine E_1 is considered *large* as it is trained on approximately 108000000 words; engine E_2 is trained on approximately 500000 words and we refer to it as small.

4.1 Delays for new messages in the PP-ASYNC method

This experiment aims to reveal whether the PP-ASYNC method reduces the delay for processing new requests. We performed 5 tests. For each of them we used 10000 requests randomly selected from an English text. We also used 100 consumers. Specific details about the tests are shown in Table 1.

Test	Queues	Details
PP-ASYNC sequential	– 5 priority queues $Q_1..Q_5$. Queue $Q_i \in \{Q_1, \dots, Q_5\}$ has consumption ratio i .	– 1000 translation requests for engine E_1 are first stored in each queue. – 1000 translation requests for engine E_2 are stored in each queue afterwards.
PP-ASYNC parallel	– 5 priority queues $Q_1..Q_5$. Queue $Q_i \in \{Q_1, \dots, Q_5\}$ has consumption ratio i .	– 1000 translation requests for engine E_1 and 1000 translation requests for engine E_2 are in a random order.
PP-ASYNC distributed	– 5 priority queues $Q_1^{E_1}..Q_5^{E_1}$. Queue $Q_i^{E_1} \in \{QE_{11}, \dots, QE_{15}\}$ has consumption ratio i . – 5 priority queues $Q_1^{E_2}..Q_5^{E_2}$. Queue $Q_i^{E_2} \in \{QE_{21}, \dots, QE_{25}\}$ has consumption ratio i . Total: 10 queues.	– 1000 translation requests for engine E_1 in each queue $Q_i^{E_1}$. – 1000 translation requests for engine E_2 in each queue $Q_i^{E_2}$.
ASYNC sequential	– 1 queue	– 5000 translation requests for engine E_1 are first stored in the queue and 5000 translation requests for engine E_2 are stored in the queue afterwards.
ASYNC parallel	– 1 queue	– 5000 translation requests for engine E_1 and 5000 translation requests for E_2 are stored in a random order.

Table 1: Tests for determining the effect of the PP-ASYNC method on request delays.

We ran each experiment and measured the time when a request is consumed from a queue. We then measure the time difference between the first processed request (at overall) and the

⁷We use the term “engine size” to refer to the number of words which were used to train the engine.

Test name	Delay time (in minute)				Total
	E_1		E_2		
	First	Last	First	Last	
PP-ASYNC sequential	0.00	30.29	7.61	38.17	38.17
ASYNC sequential	0.00	22.65	20.98	44.90	44.90
PP-ASYNC parallel	0.00	34.14	0.02	34.13	34.14
ASYNC parallel	0.00	28.93	0.02	25.58	28.93
PP-ASYNC distributed	0.00	33.34	0.46	31.85	33.34

Table 2: Time until the first message for a given engine is read from a queue.

first processed request for a specific engine. That is, we compute the delay before the requests to a specific engine are accessed for the first time. For example, a given queue has requests for engine E_1 and engine E_2 . The first request to be read will have no delay (0.00 minutes). Let us say that it is from engine E_1 . 10 minutes after that a consumer reads the first request for engine E_2 resulting in a delay of 10 minutes for engine E_2 .

Our results for each of the five tests are summarized in Table 2.

From Table 2 we notice, first that for the parallel tests there is no difference between the *ASYNC* and *PP-ASYNC* methods. That is because the requests are randomly distributed and, in practice, they are not prioritized.

The results for the *PP-ASYNC* sequential, *ASYNC* sequential and *PP-ASYNC* distributed are of greater interest as they show the benefits of using the *PP-ASYNC* method. Namely, from the comparison of the *PP-ASYNC* sequential and the *ASYNC* sequential we notice that using multiple queues decreases the delay significantly. As for the *PP-ASYNC distributed* test the delay is reduced even more, shown that using separate queues to store the requests for different engines is a preferred option when it comes to quick system response. In Figure 4 we compare the delays represented as a percentage of the total processing time.

Figure 4 again shows that the *PP-ASYNC* method with multiple queues designated to one engine has the best performance, i.e., almost no delay⁸. On large scale, i.e., when users send huge volumes of translation requests, such delay reductions may be crucial for their operation.

4.2 Comparison between *ASYNC* and *PP-ASYNC*

As shown in Section 4.1 the worst case scenario for the *PP-ASYNC* method is when messages are randomly distributed among different queues (see Table 2). Then there is no practical difference between the *PP-ASYNC* and the *ASYNC* method. We therefore compare the two methods in such a scenario in order to test whether there is a degradation in performance due to the additional distribution mechanism of the *PP-ASYNC* method. For this experiment we used 50000 requests for 2 engines and invoked 100 consumers. The two tests we executed are described in Table 3.

We ran three iterations for each test and we measured the time consumed from the moment the first request is read from a queue until it is emptied. We show the results from these tests in Table 4 in minutes.

The results summarised in Table 4 confirm that in the worst case scenario for the *PP-ASYNC* method the performance of the online translation of KantanMT does not degrade.

⁸In practice the delay is approximately one second.

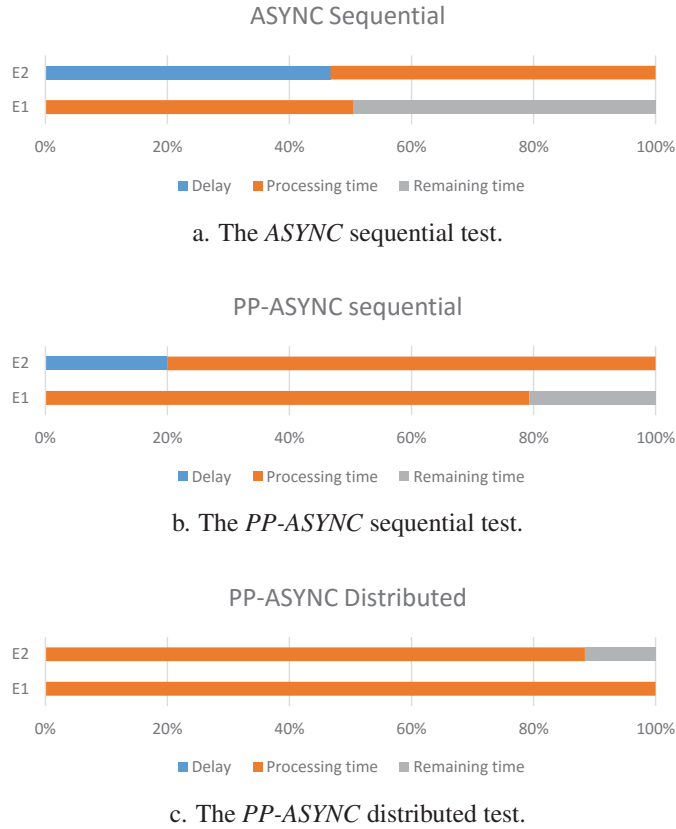


Figure 4: Comparison of delays and processing times, relative to the total time.

Test	Queues	Details
<i>PP-ASYNC</i> comparison	– 6 priority queues $Q_{-1}, Q_1..Q_5$. Queue $Q_i \in \{Q_1, \dots, Q_5\}$ has consumption ratio i . Q_{-1} has infinite consumption ratio.	– 25000 translation requests for engine E_1 and 25000 translation requests for engine E_2 are in a random order.
<i>ASYNC</i> comparison	– 1 queue	– 25000 translation requests for engine E_1 and 25000 translation requests for E_2 are stored in a random order.

Table 3: Tests for comparing the *PP-ASYNC* and *ASYNC* in the worst case scenario.

5 Conclusions

In this paper we presented the online translation methods of KantanMT. We focused on our most innovative method – the *PP-ASYNC*. It uses multiple intermediate data storage units to store users requests; efficient producer-consumer technique is used to distribute the requests efficiently for translation. Each intermediate data storage unit is implemented as a queue and is assigned a priority. The priority defines at what rate the queue is consumed.

The new method extends the *ASYNC* method by introducing multiple queues and priorities for each queue. By using priorities this method allows our users to move forward the translation of important segments. Such a mechanism is crucial to our users, especially when it comes to

Test name	Consumption ratio	Iteration 1	Iteration 2	Iteration 3	Average
PP-ASYNC comparison	infinite	17.09	27.91	67.58	37.53
	1	57.84	55.38	85.75	66.32
	2	49.31	46.62	75.68	57.20
	3	44.91	42.71	73.22	53.61
	4	39.56	37.20	65.49	47.42
	5	36.07	30.09	60.88	42.35
Total		72.93	75.66	114.73	87.78
ASYNC comparison	-	83.24	72.80	110.29	88.78

Table 4: Consumption time (in minutes).

processing large volumes of data.

Our empirical evaluation showed that while in the worse case scenario the *PP-ASYNC* is as efficient as the *ASYNC* method, in general it reduces delays drastically. As such, our method is suitable for processing concurrently/in parallel requests from different users. Furthermore, the distributed architecture of our platform allows the *PP-ASYNC* method to handle large amounts of data efficiently and respond on time.

In the future we aim at optimizing this method by building an intelligent system to enforce extra control on the sysetm with less human interventions.

References

- Arpaci-Dusseau, R. H. and Arpaci-Dusseau, A. C. (2015). *Operating systems: Three easy pieces*. Arpaci-Dusseau.
- Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.
- Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1st edition.
- Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi Itc-Irst, N., Cowan, B., Shen, W., Moran Mit, C., Zens, R., Aachen, R., Dyer, C., Constantin, A., Colledge, W., and Cornell, E. H. (2007). Moses: Open source toolkit for statistical machine translation. pages 177–180.



The Reasonable Effectiveness of Data

Achim Ruopp, Director of Data Cloud

Agenda

- ▶ Introduction
- ▶ Data
 - ▶ Collect
 - ▶ Combine
 - ▶ Select
- ▶ Grey Box Testing
- ▶ Lessons Learned

ModernMT Project

Horizon 2020 Innovation Action
3M € funding
3 years: 2015-2017



Goal:

deliver a large-scale commercial online **machine translation** service based on a new open-source distributed architecture.



Horizon 2020
European Union funding
for Research & Innovation

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 645487.

ModernMT Consortium

Business



Research



Modern MT in a nutshell

- Zero training time
- Manages context
- Learns from users
- Scales with data and users



For the first three points see Marcello Frederico's talk:
Machine Translation Adaptation from Translation Memories in ModernMT

Introduction

The Unreasonable Effectiveness of Data

- ▶ 2009 article by Alon Halevy, Peter Norvig, and Fernando Pereira
- ▶ Large sets of unlabeled data
 - ▶ “invariably, simple models and a lot of data trump more elaborate models based on less data.”
 - ▶ Translation “a natural task done every day”
 - ▶ “a threshold of sufficient data”
- ▶ Representational model: “For many tasks, words and word combinations provide all the representational machinery we need to learn from text.
- ▶ Solving the “semantic interpretation problem” based on text

Introduction

The Next Challenge: Data Efficiency

- ▶ Kamran and Sima'an, 2015

- ▶ “blind concatenation of all available training data may shift translation probabilities away from the domain that the user is interested in”

- ▶ Eetemadi et al, 2015

- ▶ “We now find ourselves, however, the victims of our own success, in that it has become increasingly difficult to train on such large sets of data, due to limitations in memory, processing power, and ultimately, speed”
- ▶ “training data has a wide quality spectrum. A variety of methods for data cleaning and data selection have been developed to address these issues.”

Amir Kamran and Khalil Sima'an, **Technical report, DatAptor STW project number 12271**, University of Amsterdam, 2015

▶ Proceedings of AMTA 2016, Oct. 2-4, Hilton Park, **Saulh Eetemadi, William Lewis, Kristina Toutanova, and Hayder Radha. 2015. Survey of data-selection methods in statistical machine translation. *Machine Translation* 29, 3-4**

Austin, Oct 28 - Nov 1, 2016 | p. 129



Collect Translated's MyMemory



- ▶ The world's largest Translation Memory
- ▶ Seeded originally with translation data from public organizations and the web
- ▶ Search
 - ▶ Edit search results
 - ▶ CAT Tool plug-ins
- ▶ Users can upload TMs
 - ▶ Private or publicly shared
- ▶ TM matching
 - ▶ No mass downloading
- ▶ September 2016: 1.491.231.338 human contributions
- ▶ 100s of millions of words per ModernMT language pair

Collect TAUS Data Cloud



- ▶ Largest industry-shared repository of translation data
- ▶ A neutral and secure repository platform for
 - ▶ Sharing/pooling translation data based on a reciprocity model
 - ▶ Searching domain-specific or general data
 - ▶ Leveraging Translation Data
- ▶ Solid legal framework established by 45 founding members
- ▶ Addresses the shortage of available in-domain parallel data from the industry
- ▶ September 2016: 72,476,886,904 words in the repository
- ▶ 10s to 100s of millions of words per ModernMT language pair

Collect

The Web – Crawling it is hard

- ▶ The Web is large - even the so-called Surface or Indexable Web
- ▶ The Web is messy
- ▶ The Web is constantly in flux
- ▶ Not many organizations crawl the entire indexable web
 - ▶ Google - about 49 billion web pages in index (Source: <http://www.worldwidewebsite.com/>)
 - ▶ Microsoft - about 20 billion web pages in index (Source: <http://www.worldwidewebsite.com/>)
- ▶ Other crawls are focused crawls on a subset with certain criteria/goals
 - ▶ Still hard for the same reasons

Collect

CommonCrawl to the Rescue

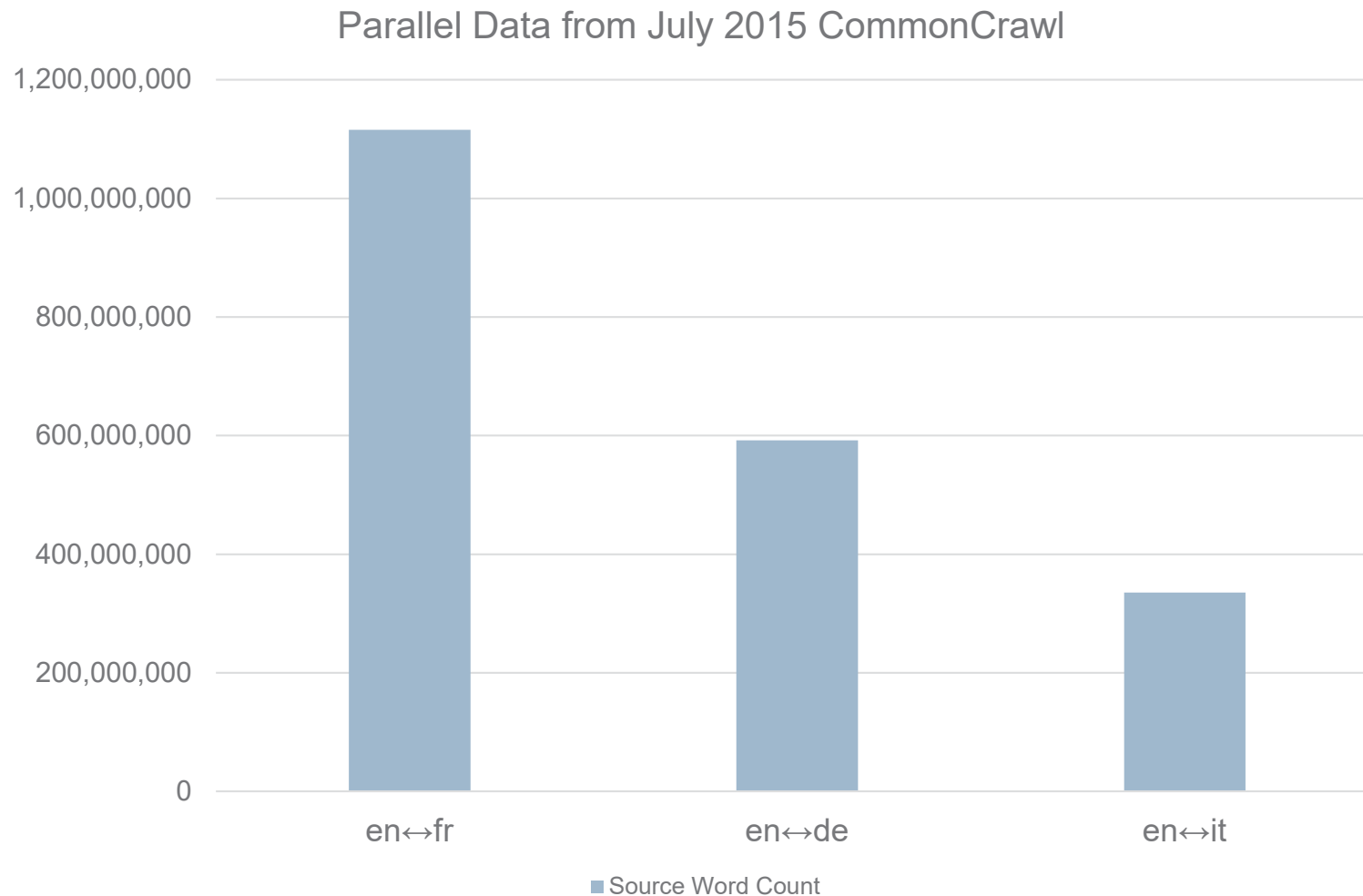
- ▶ commoncrawl.org
 - ▶ “CommonCrawl is a 501(c)(3) non-profit organization dedicated to providing a copy of the internet to internet researchers, companies and individuals at no cost for the purpose of research and analysis.”
- ▶ On average 1.5 billion unique URLs per crawl
- ▶ A very good resource for sourcing bilingual and monolingual data for machine translation purposes
 - ▶ Prototype developed by academic developers in 2012/2013 showed potential to mine parallel corpora with millions of source words

Collect

CommonCrawl to the Rescue

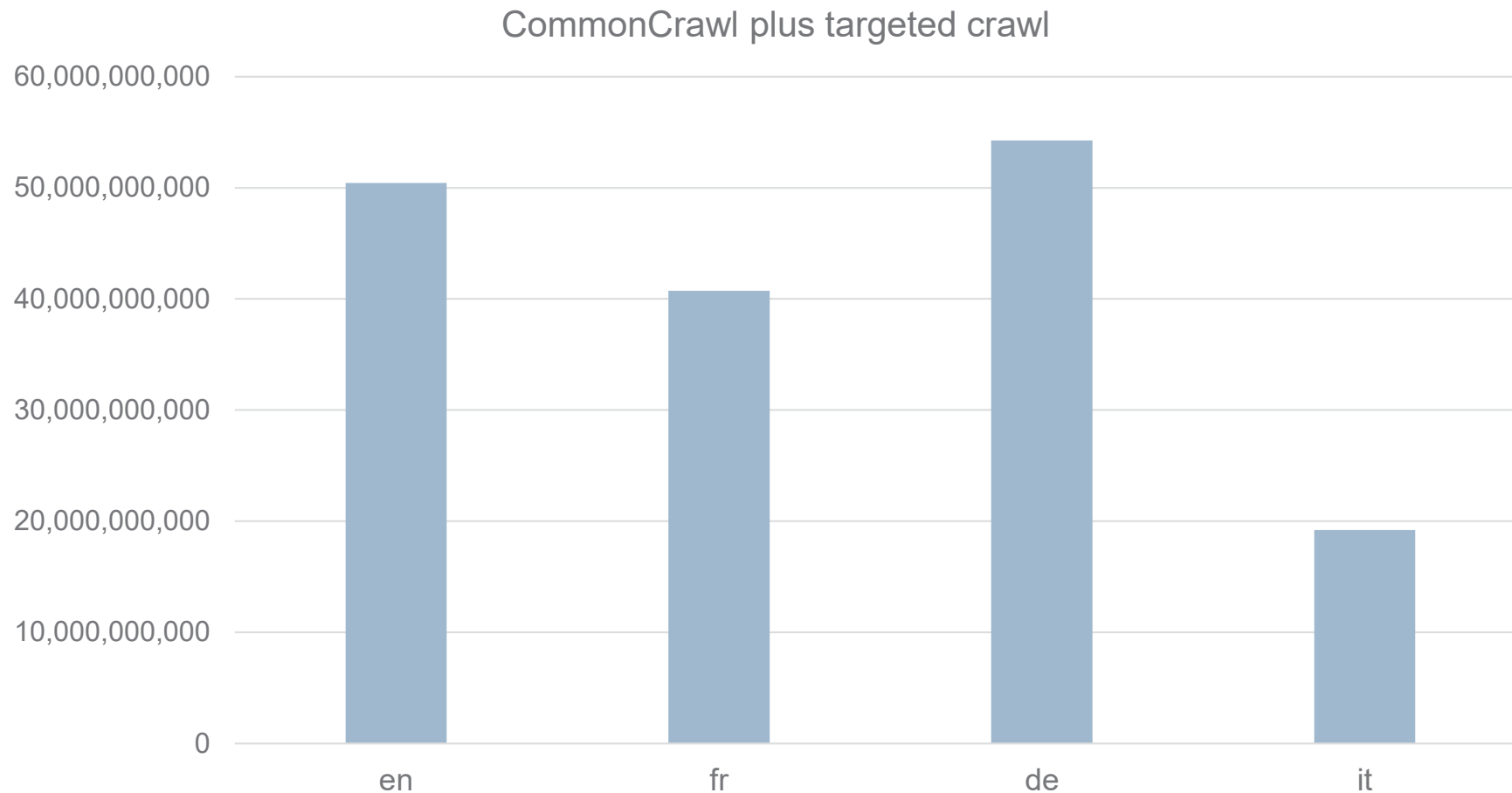
- ▶ Implemented data collection pipeline based on prototype techniques
- ▶ Collecting monolingual and bilingual data
- ▶ Open sourced at <https://github.com/ModernMT/DataCollection>
- ▶ We are making the indices of parallel pages we discover available
 - ▶ Saves running half of the data collection pipeline
 - ▶ Each user still has to download their own data
 - ▶ Avoids potential copyright issues

Collect CommonCrawl Parallel Data



- Original source language not detectable – data can be used in both directions
- Not deduplicated – separated by registered web domains

Collect CommonCrawl Monolingual Data



- Deduplicated
- More raw English, French and Italian data available

Combine

- ▶ As plain text corpora
 - ▶ TMX files are converted to plain text
- ▶ “Document” concept
 - ▶ One TMX, one document
 - ▶ One site (web domain), one document
- ▶ Uniform pre-processing and post-processing for all data
- ▶ Repository meta-data data not unified/combined
 - ▶ Ontologies for domain/content type too different
 - ▶ Not available for web crawled content
 - ▶ Is it even useful?

Select

Segment Level Data Cleaning

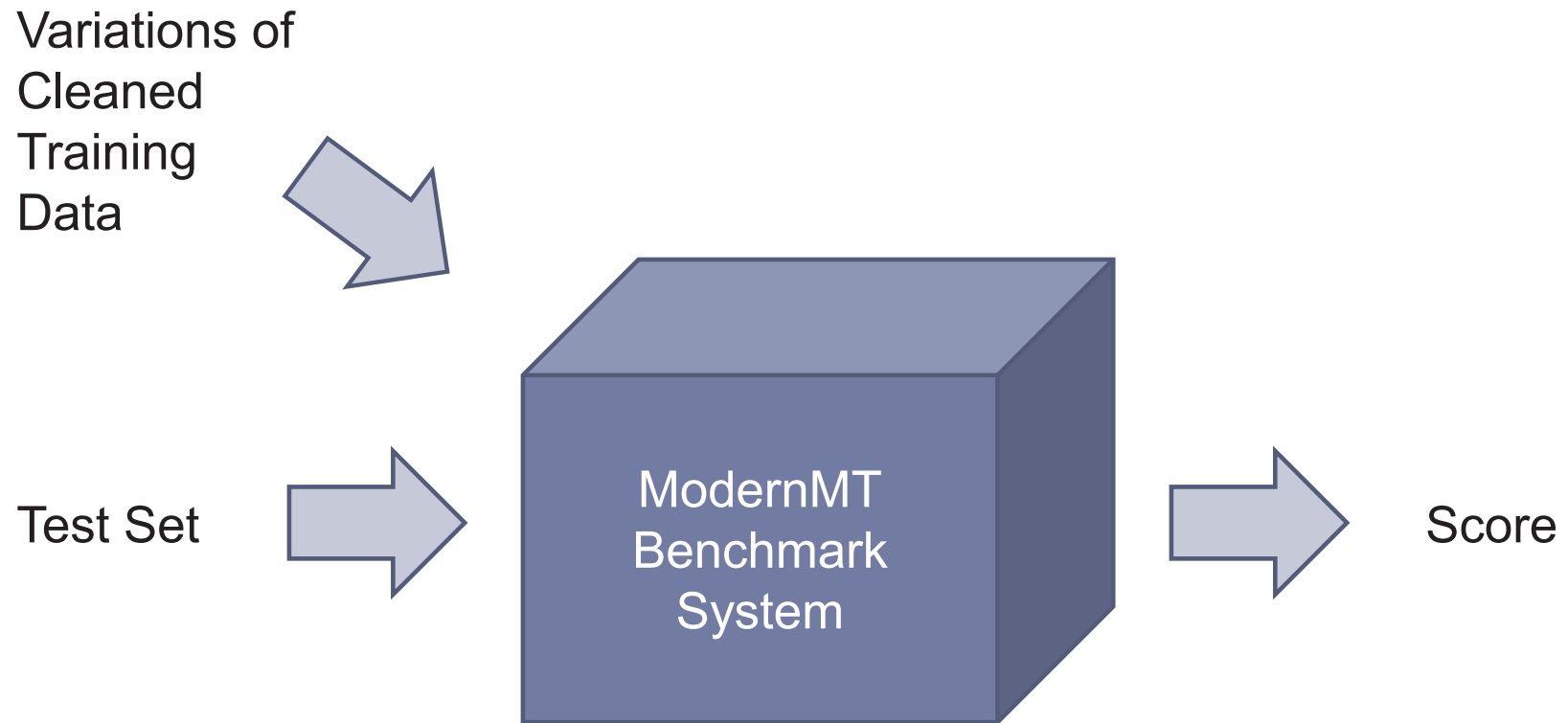
- ▶ Clean data from all sources to a uniform level
- ▶ Some quality indicators
 - ▶ Empty or identical source/target
 - ▶ Mismatches in number of sentences
 - ▶ Very lengthy segments
 - ▶ No alphabetical characters in segments
 - ▶ Inconsistency in tags or dates, etc.
 - ▶ Big difference in segment length between source and target
 - ▶ Language identification

Select Document Level

▶ Context Analyzer

- ▶ Training data “Document” selection via cosine similarity between user-provided translation input and training data
- ▶ Proven to be competitive with more complex training data selection methods for domain adaptation

Grey Box Testing



Lessons learned so far

- ▶ CommonCrawl is a great source for MT training data – please donate!
 - ▶ Some common sense cleaning of web data is necessary
 - ▶ More complex cleaning has diminishing returns
 - ▶ Web data helps with translating text with wide ranging topics/tones
 - ▶ You might not need it if you have focused domains/topics with data available
 - ▶ Web data does not help to translate items out-of-vocabulary named entities/abbreviations specific to the text you are translating
- ⇒ Test set design/compilation is crucial



MT for Uralic Languages: Yandex Approach

Irina Galinskaya
Alexey Baytin
Yandex, LLC

galinskaya@yandex-team.ru
baytin@yandex-team.ru

1. Abstract

The Uralic language group is certainly interesting – it is spotted in the very distant regions of Europe and Asia. Spoken by about 25 million people, Uralic languages have spread to both sides of the Ural Mountains, reaching Balkans, Baltic region, Scandinavia, Karelia, Volga region, Western Siberia and the seaside of the Arctic Ocean.

With the exception of the three major Uralic languages (Hungarian, Finnish, Estonian), relatively small Sami and a few quite small Baltic languages (Veps, Ingrian, Livonian), all other Uralic languages are spoken on the territory of the Russian Federation. Most of them have official status in the corresponding federal regions and national autonomies. They are taught at schools, studied in universities and their usage is legally obliged in official documents. There are many enthusiasts who are trying to preserve and develop these languages. Nevertheless, it is clear that Russian is prevailing, so, in fact, most of small Uralic languages are used mostly in colloquial speech.

In terms of complexity, MT for Uralic languages is a very difficult and challenging task. First, these languages have very distinct lexicon, morphology and syntax. Second, there are many dialects, which are quite distinct as well. Third, Uralic languages have highly productive morphology which leads to a strong data sparsity in SMT. Fourth (and the worst), there are very few electronic documents available for most of these languages.

Our general approach to Uralic group was as follows. We divided its languages into three subgroups:

- 1) with more than 1M native speakers (Hungarian, Finnish, Estonian);
- 2) with 100K to 1M native speakers (Udmurt, Meadow Mari);
- 3) with less than 100K native speakers (Hill Mari, Karelian, Nenets).

For the major subgroup we were able to collect a sufficient amount of parallel documents from the web, and thus to build quite good baseline translation systems. Automorphology and compound splitting were used to further improve translation quality.

For Udmurt and Meadow Mari, languages from the second subgroup, we managed to crawl only a modest parallel corpora and therefore were forced to rely on a hand-crafted lexicon and morphology. The pipeline included the following steps:

- 1) used Bible, Wikipedia and human-made dictionaries as a main lexical base;
- 2) developed morphological analyzers, built lemmatized models and implemented “lemma-to-lemma” decoding;
- 3) post-processed lemmatized translations by synthesizing proper Russian word forms.

Lack of available documents in electronic form (and sometimes even in paper form) for languages of the third subgroup poses the question of how to build translation and language models without data. In attempts to find an answer to this question, we made a little shift to a quite specific group of small languages – magic ones. Some of them (like Elvish

dialects) have linguistic artifacts in the form of lexicon and well elaborated phonology and morphology. They are also known to be connected with Uralic languages. We considered Sindarin as an archetype of under-resourced languages and decided first to experiment with it. Some results were very promising, so we became more optimistic in upcoming efforts with MT for low-resourced Uralic languages and have successfully developed a translation for Hill Mari.

The STAR logo consists of the word "STAR" in a white, sans-serif font, with a stylized triangle above the letter 'A'. The logo is set against a dark red rectangular background.

STAR

A photograph of a man in a light-colored shirt holding a woman's hand, walking away from the camera on a city street at sunset. The street is lined with buildings and a red double-decker bus is visible in the distance. The scene is bathed in a warm, golden light.

Connecting your visions, technologies and customers

STAR Group

AMTA 2016 Commercial MT Users and Translators Track

Seamlessly integrating machine translation into existing translation processes (STAR MT and Transit NXT)

Nadira Hofmann, STAR Language Technology & Solutions

Outline

▲ What customers want to know

- Can we benefit from MT?
- Is it possible to evaluate the MT system?
- MT black box and “MT psychology”:

How do we involve our translators and project managers?

Typical TM customers

▲ Existing tools

- Translation Memory / Terminology Management systems
- Workflow systems
- Third-party systems

▲ Text types

- Technical documents, software localization, legal texts, subtitling, etc.
- Structured documents
- Languages – every single one you can think of

▲ In-house translators, freelancers, LSPs

Typical MT requirements

- ▲ No extra tool for project managers, translators and “non-experts”
- ▲ Specific, customized engines
- ▲ Analysis of MT quality
- ▲ Alternative to online translation services
- ▲ No cloud solution
- ▲ Integrated use with a Translation Memory system
 - Retaining benefits of TM (pretranslation / fuzzy matches)

How do we support our (TM) MT customers?

▲ Proof of Concept for evaluating STAR MT

- **Stage 1:** Engine training and initial analysis
- **Stage 2:** Pilot phase in productive environment
- **Stage 3:** Productive analysis of pilot phase results

Stage 1 – Engine training

▲ Creation of MT training packages based on:

- Customer-specific Translation Memory
- Customer-specific terminology

▲ Deployment of pilot engine(s)

- During pilot phase: HTTPS access
- Later: MT server on customer's premises

Stage 1 – Sentence Bleu lists

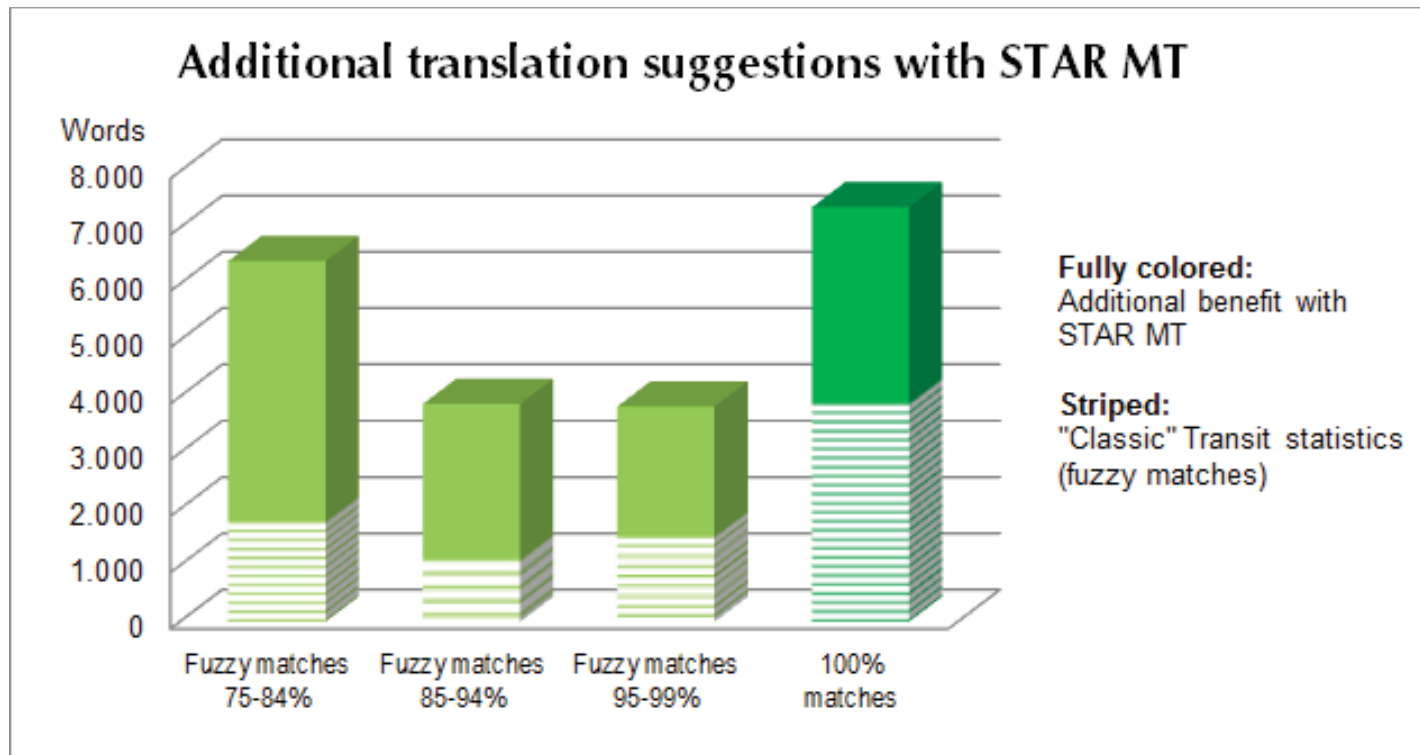
▲ First impressions of MT quality

Engine Info: Deutsche Bahn - Pilot Engine			
BLEU	Source Text	MT Output	Human Translation
0,825080	Abschließend wird der Bewerber- / Bieterkreis vorgeschlagen.	Finally, the pool of candidates / tenderers is proposed.	To conclude , the pool of candidates/tenderers is proposed.
0,815355	Fragen zur TSI PRM: Messungen der Sprachverständlichkeit 3 - RFU 073	Questions regarding TSI PRM: Measuring the intelligibility of spoken information 3 - RFU 073	Questions regarding TSI PRM: Measuring the intelligibility of spoken information 2 - RFU 073
0,783129	Die Strecke verläuft überwiegend auf Brücken oder in Tunneln - 63 von 107 Kilometer.	The line runs mainly on bridges or in tunnels - 63 of 107 kilometres.	The line is routed mainly on bridges or in tunnels - 63 of 107 kilometres.
0,753922	Wir schätzen sie als zuverlässige, pflichtbewusste und ehrliche Mitarbeiterin.	We appreciate it as a reliable, conscientious and honest employee.	We value her as a reliable, conscientious and honest employee.
0,668740	Ungültiges Datum - TT.MM.JJ	Invalid date - DD.MM.YY	Invalid date - TT.MM.JJ

Stage 1 – Initial analysis

- ▲ Analysis of productive jobs of the last 3 to 4 months that have been translated without MT support
 - Jobs are translated again with MT (pretranslated segments excluded)
 - MT results are compared with human translations

Stage 1 – Initial analysis



Example: Initial analysis for one of our customers ("Technology" division)

Stage 1 – Initial analysis

Division	100% matches	Fuzzy matches	No matches	Pretranslated
IT	+244%	+286%	-17%	remains as is
Company	+112%	+86%	-10%	
Technology	+89%	+215%	-24%	
HR	+84%	+115%	-6%	
Legal	+84%	+61%	-4%	
Traffic	+81%	+132%	-13%	
Finance	+27%	+109%	-25%	
CEF	+13%	+113%	-19%	

Examples: Initial analyses for one of our customers (all divisions)

Stage 2 – Involving everyone in the process

▲ On-site workshop for all stakeholders

▲ Involving all translators, who:

- are informed at an early stage about the planned MT system
- receive a feedback sheet with queries regarding:
 - project topic and text type
 - “perceived” benefit and “perceived” quality
 - linguistic and terminological quality

Stage 2 – Integration into existing processes

▲ One-off adjustment of:

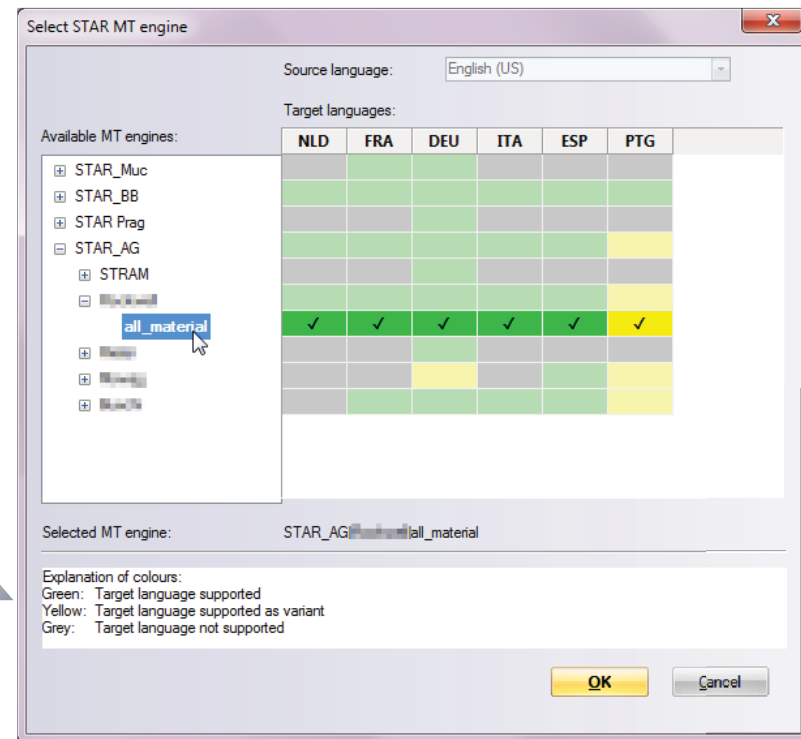
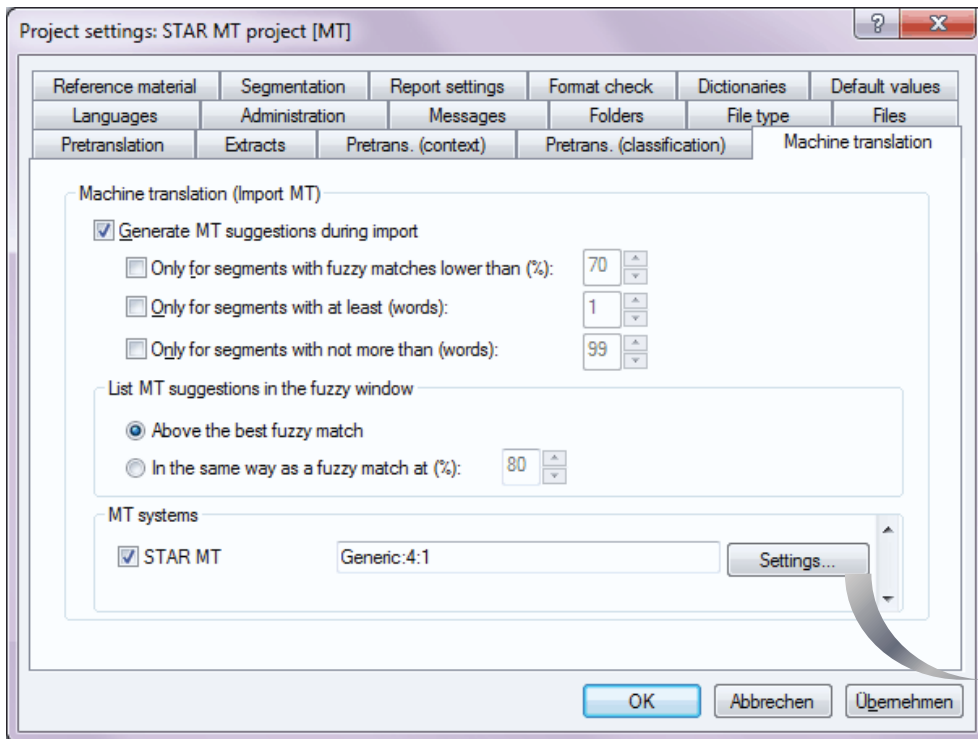
- project templates (manual translation processes)
- workflow settings (automatic translation processes)

▲ Project management steps remain the same

▲ In Transit, MT suggestions are automatically:

- generated during project import
(for all segments that have not been pretranslated)
- packed into the project package during project exchange

MT project settings in Transit



Stage 2 – “Look and feel” for translators

- ▲ No additional tool and no access to MT engine required
- ▲ Work in the Transit editor as usual
- ▲ MT suggestions:
 - are provided with the project package
 - are displayed and used like fuzzy matches
- ▲ MT quality assurance
 - formal checks, terminology, markups, translation variants, etc.

“Look and feel” for translators

- ▲ Instead of “No fuzzy match found”:

		c1005_3_dhm_appguide
		carbon steel.
430		In materials with an uneven structure when chip breaking problems
432		More advantageous for long-series production. <1-inb>
434		Uniform and extremely long workpieces.

		c1005_3_dhm
430	*	In materials with an uneven structure when chip breaking problems
432		More advantageous for long-series production. <1>
434		Uniform and extremely long workpieces.
436	~	Nécessite une machine spéciale de forage de trous profonds.

Source Fuzzy		
		Machine Translation (Import MT)
New		More advantageous for long-series production. <1>
MT		Plus avantageux pour long-series production. <1>

- ▲ Markups are automatically inserted
- ▲ Untranslated words are indicated

“Look and feel” for translators

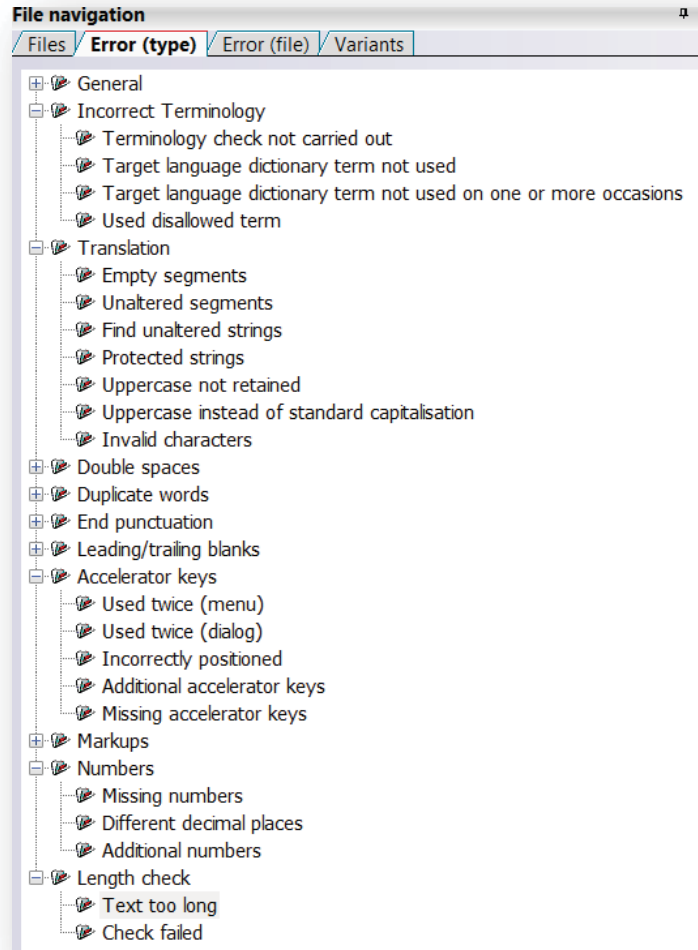
- ▲ In addition to fuzzy matches:

Source Fuzzy	
	Machine Translation (Import MT)
New	Ergebnisse zu den zuginduzierten Drucklasten neben dem Gleis
MT	Results for the train-induced pressure loads in addition to the track
87%	D:\Transit NXT\projects\MT\3184_16_TSI Zulassungen SET Hrach_deem\15-21427-T.TV
Ref	Ergebnisse zu den zuginduzierten Strömungsgeschwindigkeiten neben dem Gleis
New	Ergebnisse zu den zuginduzierten Drucklasten neben dem Gleis
Ref	Results for the train-induced flow velocities next to the track

- ▲ MT suggestion, validated by fuzzy match (TM):

Source Fuzzy	
	TM-validated MT
New	Highest material removal rate
	Débit copeaux maximum

Quality assurance of MT-translated segments



- Formal errors are displayed in the “File navigation” window
- Spelling checks based on MS Office speller, reference material and/or dictionaries
- Source / translation variants check
- Segment filter for MT-translated segments

Stage 2 – Web application (STAR MT Translate)

▲ Allows specific evaluation of MT quality by language experts

- Direct access to pilot engines
- Translation of individual sentences or paragraphs

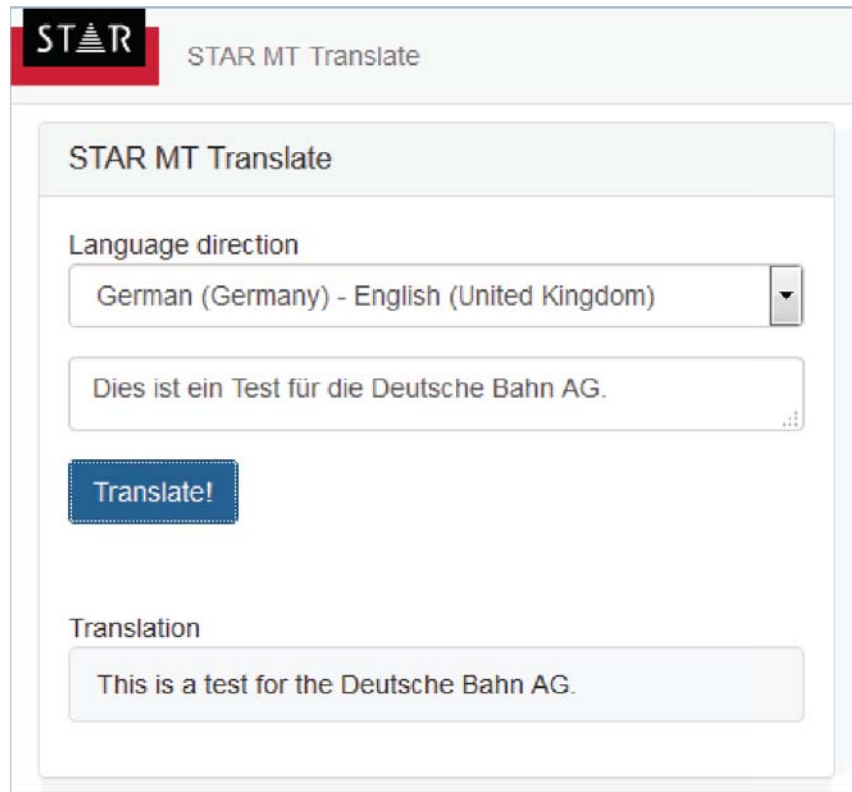
▲ In general:

Alternative to online services (also for “non-experts”)

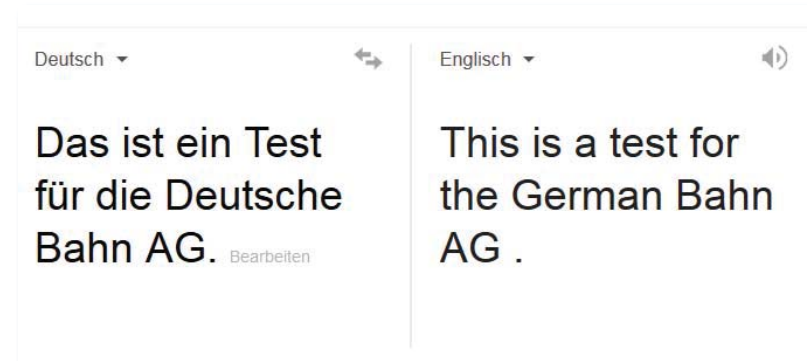
- Confidential corporate data stays “in-house”
- Translations use style and terminology of corporate language
- Translation of entire documents (Office, PDF)

An example

STAR MT Translate:



Google Translate:



Stage 3 – Productive analysis

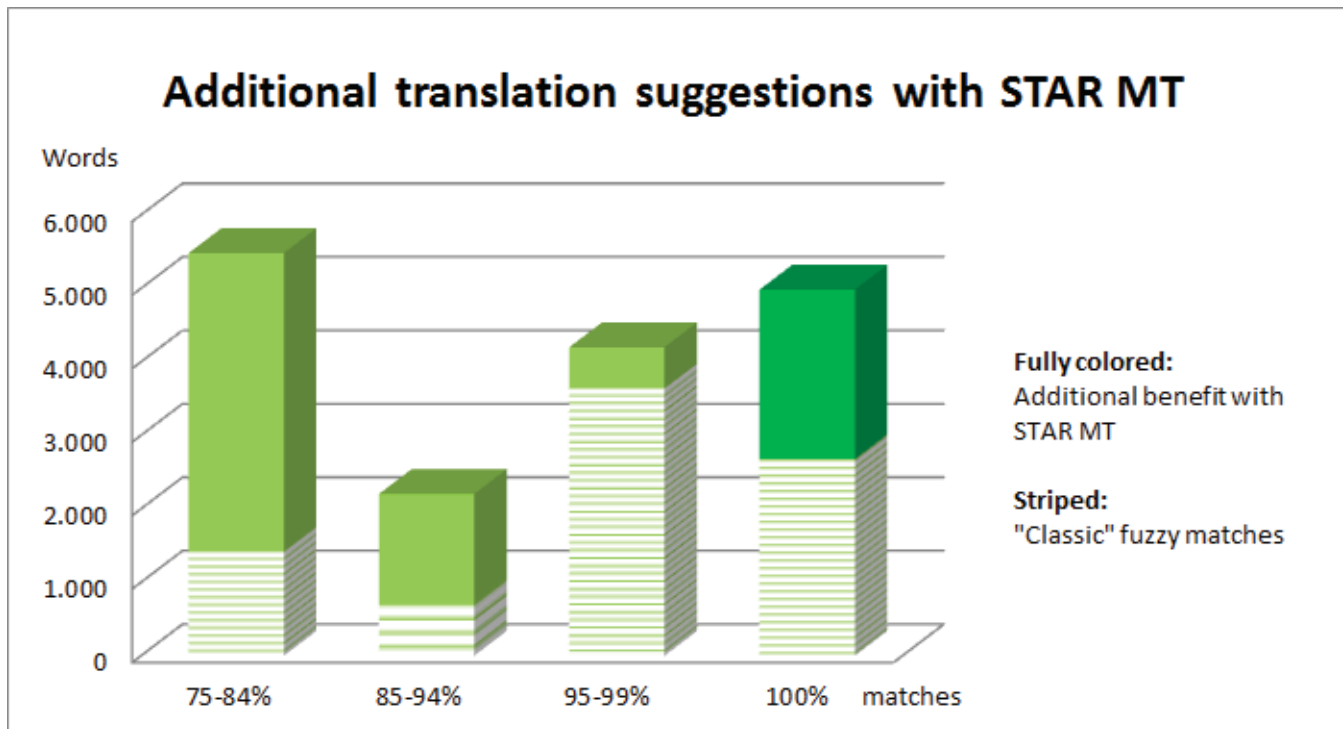
▲ Evaluation of productive analysis

- Objective benefit
- Comparison of initial analysis and productive analysis

▲ Evaluation of feedback sheets

- “Perceived” benefit
- Comparison of objective and “perceived” benefit

Stage 3 – Productive analysis



100% matches:	+86%
Fuzzy matches:	+104%
No matches:	-20%

Example: Productive analysis for one of our customers ("Technology" division)

Lessions learned

- ▲ A successful pilot phase requires
 - duration of several months
 - representative amount of productive jobs from “real life”
 - customer-specific MT scenario (IT infrastructure, MT case study)
 - all stakeholders to be involved
 - a smooth integration into daily work
 - a good cooperation between system provider and customer

Lessons learned

▲ Translators

- have high expectations of MT
- search for MT mistranslations
- have to be prepared for using MT well
- are diverse: “Facebook generation” vs. Traditional translators

▲ Proof of Concept offers a good indication of

- the expected benefit
- an accounting model with win-win situation for customers and translators

**Questions and comments
welcome!**

nadira.hofmann@star-group.net

Building a Translation Memory to Improve Machine Translation Coverage and Quality

Duncan Gillespie
Benjamin Russell
Etsy, Inc.

dgillespie@etsy.com
brussell@etsy.com

Abstract

In this paper we discuss the motivation and process for planning, building, and monitoring a translation memory (TM) that serves both human- and machine-translated text segments in a production e-commerce environment. We consider the quality improvements associated with serving human translations for commonly used and mis-translated strings, and the cost benefits of avoiding multiple re-translations of the same source text segments. We cover the technical considerations and architecture for each stage of the TM pipeline, and review the results of using and monitoring the TM in a production setting.

1 Machine Translation at Etsy

Etsy is an online marketplace for handmade and vintage items, specializing in unique goods. It is important our global member base can communicate with one another, even when they speak different languages. Machine translation is a valuable tool for facilitating multilingual interactions on our site and in our apps. An example of the translation interface shown to users can be seen in Figure 1.

Listing descriptions account for the bulk of text we machine translate. We have about 40 million active listings at an average length of around 1,000 characters, with hundreds of thousands of listings created or edited each day. These listings are machine translated into six languages. We also provide machine translation for listing reviews, forum posts, and conversations (messaging between members). For these translations, we send text to a third party machine translation service,¹ with whom we have a fixed monthly budget, imposing a limit on the number of characters we translate per month.

While a user can request a listing translation if we don't already have one (we call this on-demand translation), translating a listing beforehand and showing a visitor the translation automatically (we call this pre-translation) provides a more fluid browsing experience. Pre-translation also allows listings to surface in search results in multiple languages, both for searches on Etsy and on external search engines like Google.

2 The Benefits of a Translation Memory

Many of the strings we machine translate from one language to another are text segments we've seen before. Our most common segments are used in millions of listings, with a relatively small subset of distinct segments accounting for a very large proportion of the content. For example, the sentence "Thanks for looking!" appears in around 500,000 active listings on Etsy, and

¹Microsoft Translator

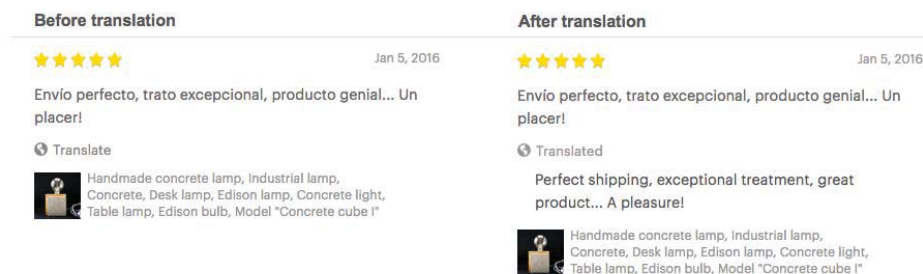


Figure 1: An example review translation on Etsy's website.

has appeared in over 3 million now inactive listings. More broadly, the distribution of unique paragraphs in Etsy listings has a classical Zipfian shape (see Figure 2), with the top segments appearing on the order of 10^7 times and approximately 10^9 distinct segments.

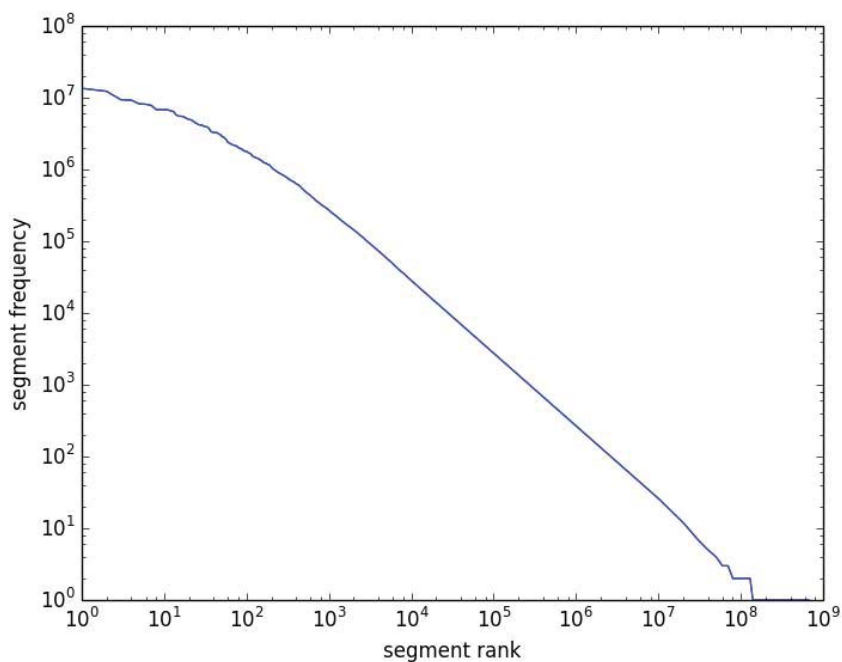


Figure 2: Frequency and rank of text segments (titles, tags, and description paragraphs) appearing in listings on Etsy. The distribution of segments roughly conforms to a Zipfian shape, where a string's rank is inversely proportional to its frequency.

Prior to undertaking this project, a single text segment that appeared in thousands of listings on Etsy was re-translated once for every listing. It would also be re-translated any time a seller edited a listing. This meant our translation budget was being spent on millions of repeat translations that would be better used to translate unique content into more languages.

To solve this problem, we built a translation memory. At its simplest, a translation memory stores a text segment in one language and a corresponding translation of that segment in another language. Translation memories are traditionally used by human translators to avoid re-translating the same text segments multiple times, and to ensure consistency of translations, thereby improving quality (Christensen and Schjoldager, 2010; Reinke, 2013). Combining machine translation with human-translated strings in a translation memory has also been a topic of research, and has been shown to have a positive effect on overall translation quality (Marcu, 2001; Koehn and Senellart, 2010).

For our purposes, storing strings in a translation memory allows us to serve translations for these strings from our own databases, rather than making repeated requests to the translation service. Storing these translations for later reuse has important implications on both quality and coverage. In terms of quality, the translation memory allows us to translate individual strings instead of translating one block of text as a whole. This means we can see which text segments are most commonly used on Etsy and have these segments human translated. Serving human translations instead of machine translations for these common segments improves the overall quality of our translations.

Secondly, storing common translations in the translation memory and serving them ourselves also allows us to drastically reduce the number of duplicate segments we send to the translation service. This process lets us translate seven times more content for the same cost, increasing our overall language coverage.

3 Initial Considerations

We had two main concerns when planning the translation memory architecture. First, we had to plan for adequate **capacity**. The more text segments we store in the translation memory, the greater our coverage. However, storing every paragraph from each of our more than 35 million active listings, and a translation of that paragraph for each of our supported languages, would mean an exceptionally large database table. We wanted to keep the table limit under a billion rows to make sure it was maintainable under our existing MySQL infrastructure.

Second, we needed to provide a mechanism for periodic **deletions**. The translation service's quality is continually improving, and to take full advantage of these improvements we need to periodically refresh entries in the translation memory by deleting older translations. We wanted to be able to delete several hundred million rows on a monthly basis without straining system resources.

4 The Translation Memory Architecture

The translation memory consists of several separate services, each handling different tasks. The services act sequentially upon a given text segment, only sending segments to the third party service after exhausting all other possible translation sources. A full diagram of the pipeline is shown in Figure 3. A brief overview of each step:

4.1 Splitting into segments

The first step of the translation pipeline is splitting blocks of text into individual segments. The two main choices here were splitting by sentence or splitting by paragraph. We chose the latter for a few reasons. Splitting by sentence gave us more granularity, but our estimated translation memory hit rate was only 5% higher with sentences versus paragraphs. The increased hit rate wasn't high enough to warrant the extra logic needed to split by sentence, nor the increase in table rows needed to store every sentence, instead of just every paragraph. Moreover, although automatic sentence boundary detection systems can be quite good, Read et al. (2012) evaluated the most popular systems on user-generated content and found that accuracy peaked at around

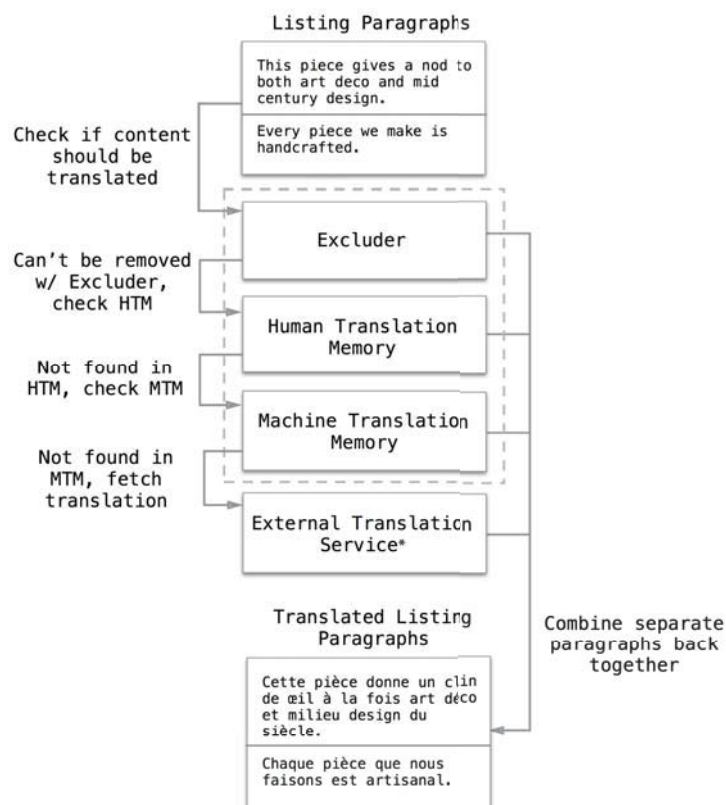


Figure 3: An overview of the translation memory pipeline. The external translation service is Microsoft Translator.

95%. In contrast, using newline characters to split paragraphs is a straightforward and error-free way to segment text.

4.2 Excluder

The Excluder is the first service we use to process translations. It removes any content we don't want to translate, specifically lines containing only links, numbers, or non-alphanumeric characters.

4.3 Human Translation Memory (HTM)

After excluding non-translatable strings, and before looking for a machine translation, we check first for an existing human translation. Human translations are provided by Etsy's professional translators (the same people who translate Etsy's static site content). These strings are stored in a separate table from the Machine Translation Memory and are updated using an internal tool we built, pictured in Figure 4.

4.4 Machine Translation Memory (MTM)

We use sharded MySQL tables to store our machine translation entries. Sharded tables are a well-established pattern at Etsy, and the system works especially well for handling the large row

Human Translation Memory Recent Entries Upload TSV New Entry

Source Language	Source Content	Translated Language	Translated Content	Update Date		
en	gold	ja	ゴールド	Jan 12, 2016, 1:49 pm	edit	delete
en	elastic	ja	伸縮性のある	Jan 12, 2016, 1:49 pm	edit	delete
en	Necklace	ja	ネックレス	Jan 12, 2016, 1:49 pm	edit	delete
en	sale	ja	セール	Jan 12, 2016, 1:49 pm	edit	delete
en	copper	ja	銅	Jan 12, 2016, 1:49 pm	edit	delete
en	Metal	ja	金属	Jan 12, 2016, 1:49 pm	edit	delete
en	lace	ja	レース	Jan 12, 2016, 1:49 pm	edit	delete
en	winter	ja	冬	Jan 12, 2016, 1:49 pm	edit	delete

Figure 4: The interface for managing human-translated segments.

count needed to accommodate the text segments. As mentioned earlier, we periodically want to delete older entries in the MTM to clear out unused translations, and make way for improved translations from the translation service. We partition the MTM table by date to accommodate these bulk deletions. Partitioning allows us to quickly drop all the translations from a certain month without worrying about straining system resources by deleting millions of individual entries.

4.5 External Translation Service

If there is new translatable content that doesn't exist in either our HTM or MTM, we send it to the translation service. Once translated, we store the segment in the MTM so it can be used again later.

4.6 Re-stitching segments

Once each of the segments has been processed by one of our four services, we stitch them all back together in the proper order.

5 The Results

We implemented the Excluder, HTM, and MTM in that order. Implementing the Excluder first allowed us to refine the text splitting, restitching, and monitoring aspects of the pipeline before worrying about data access. Next we built the HTM and populated it with several hundred translations of the most common terms on Etsy. Finally, at the end of November 2015, we began storing and serving translations from the MTM.

5.1 Coverage

As you can see from the graphs in Figure 5, we now only send out 14% of our translations to the translation service, and the rest we can handle internally. Practically, this means we can pre-translate over seven times more text on the same budget. Prior to implementing the translation memory, we pre-translated all non-English listings into English, and a majority of the rest of our listings into French and German. With the translation memory in place, we are pre-translating all eligible listings into English, French, German, Italian, Spanish, and Dutch, with plans to scale to additional languages.

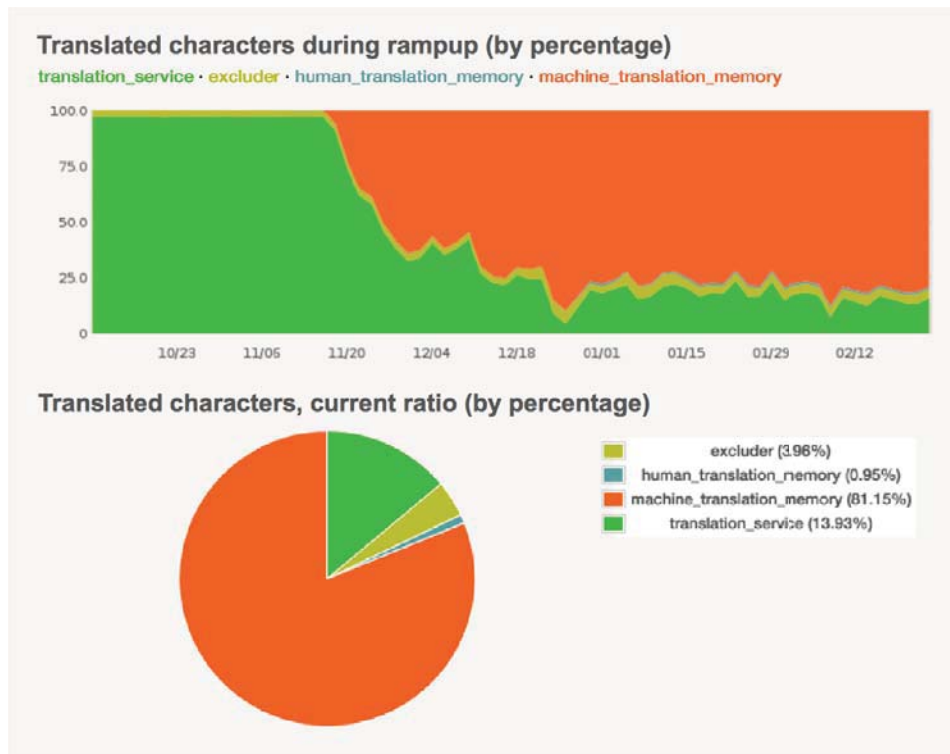


Figure 5: With the translation memory in place, we only need to send out a fraction of the listing segments to the third-party service for re-translation.

5.2 Quality

1% of our translations (by character count), are now served by the human translation memory. These HTM segments are mostly listing tags. These tags are important for search results and are easily mis-translated by an MT system because they lack the context a human translator can infer more easily. Additionally, human translators are better at conveying the colloquial tone often used by sellers in their listing descriptions. With the HTM in place, the most common paragraph on Etsy, “Thanks for looking!” is human translated into the friendlier, “Merci pour la visite !” rather than the awkward, “Merci pour la recherche !” The English equivalent of this difference would be, “Thanks for visiting!” versus “Thanks for researching!”

5.3 Monitoring

Since a majority of our translation requests are now routed to the MTM rather than the third-party translation service, we monitor our translations to make sure they are sufficiently similar to those served by the translation service. To do this, we sample 0.1% of the translations served from the MTM and send an asynchronous call to the translation service to provide a reference translation of the string. Then we log the similarity (the percentage of characters in common) and Levenshtein distance (also known as edit distance) between the two translations. As shown in Figure 6, we track these metrics to ensure the stored MTM translations don’t drift too far from the original third party translations.

For comparison, as you can see in Figure 7, the similarity for HTM translations is not as

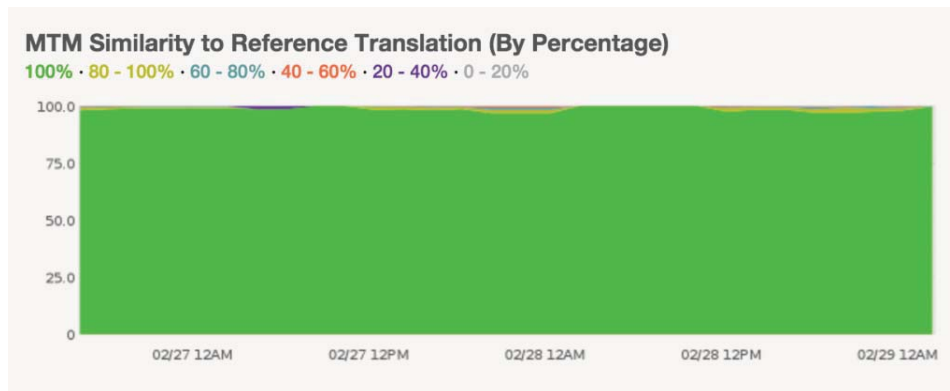


Figure 6: Tracking translation drift allows us to understand the difference between the translations we serve internally and the translations we would get from using the third party translation service. Too much drift means we are not refreshing the translations often enough.

high, reflecting the fact that these translations were not originally drawn from the third party translation service.

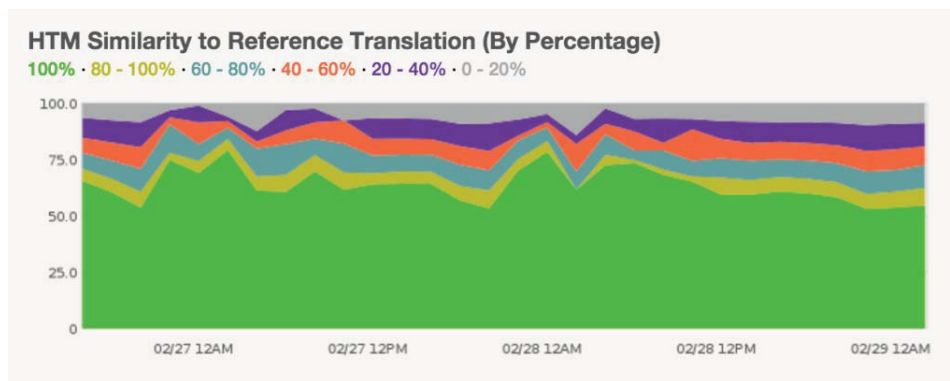


Figure 7: The translations from our human translation are expected to be significantly different than the machine translations we would be serving otherwise.

6 Additional Benefits

6.1 Correcting mis-translations

Statistical machine translation engines are trained on large amounts of data, and sometimes this data contains mistakes. The translation memory gives us more granular control over the translated content we serve, allowing us to override incorrect translations while the translation service we use works on a fix. In Figure 8 is an example where “Realistic bird” is mis-translated into German as “Islamicrovolutionservice.”

With the translation memory, we can easily correct problematic translations like this by adding an entry to the human translation memory with the original listing title and the correct German translation.



Figure 8: Correcting mis-translations quickly is important to maintain the trust of our buyers.

6.2 Respecting sellers paragraph choices

Handling paragraph splitting ourselves has the additional benefit of improving the quality of translation for many of our listings. Etsy sellers frequently include lists of attributes and other information without punctuation in their listings. For example, a listing description might contain the following three lines:

Dimensioni 24 × 18 cm
Spedizione in una scatola protettiva in legno
Verr fornito il codice di monitoraggio (tracking code)

The translation service often combines these lists into a single sentence, producing a translation like this:

Size 24 × 18 cm in a Shipping box wooden protective supplies the tracking code (tracking code)

By splitting on paragraphs, our sellers' choice of where to put line breaks is now always retained in the translated output, generating a more accurate (and visually appealing) translation like this:

Size 24 × 18 cm
Shipping in a protective wooden box
You will be given the tracking code (tracking code)

Splitting on paragraphs prior to sending strings out for translation is an improvement we could have made independent of the translation memory, but it came automatically with the infrastructure needed to build the pipeline.

7 Conclusion and Future Work

Greater accuracy for listing translations means buyers can find the items they are looking for more easily, and sellers' listings are more faithfully represented when translated. The translation memory allows us to bring this internationalized Etsy experience to more users in more languages, making it easier to connect buyers and sellers from around the world. It also helps our return on investment in machine translation, since we are able to translate more content with the same budget.

To further the translation memory quality improvements, we recently started using an Etsy-customized engine built on top of the third-party translation service's generic engine. We saw

a significant increase in purchase rate from users who interacted with the customized engine instead of the generic engine (Russell and Gillespie, forthcoming 2016).

Future efforts will be focused on improving translation quality in our search experience. Serving more human translations instead of machine translations will be especially important in search, where queries are often short strings lacking context. Quality improvements in search ensure all users have access to listings in our global marketplace, regardless of mismatches between query language and listing language.

8 Acknowledgements

We'd like to thank product manager Ray Flournoy for his guidance and input during the planning and construction of the translation memory.

References

- Christensen, T. P. and Schjoldager, A. (2010). Translation-memory (TM) research: What do we know and how do we know it. *Hermes*, 44:89–101.
- Koehn, P. and Senellart, J. (2010). Convergence of translation memory and statistical machine translation. In *Proceedings of AMTA Workshop on MT Research and the Translation Industry*, pages 21–31.
- Macklovitch, E. and Russell, G. (2000). What's been forgotten in translation memory. In *Conference of the Association for Machine Translation in the Americas*, pages 137–146. Springer.
- Marcu, D. (2001). Towards a unified approach to memory- and statistical-based machine translation. In *Proceedings of the 39th annual meeting on association for computational linguistics*, pages 386–393. Association for Computational Linguistics.
- Read, J., Dridan, R., Oepen, S., and Solberg, L. J. (2012). Sentence boundary detection: A long solved problem? In *COLING (Posters)*, pages 985–994.
- Reinke, U. (2013). State of the art in translation memory technology. *Translation: Computation, Corpora, Cognition*, 3(1):27–48.
- Russell, B. and Gillespie, D. (forthcoming 2016). Measuring the behavioral impact of machine translation quality improvements with A/B testing. In *Conference on Empirical Methods in Natural Language Processing*.

Enhancing a Production TM-MT Environment Using a Quotation TM

Hitokazu Matsushita
Steve Richardson

The Church of Jesus Christ of Latter-day Saints
50 East North Temple Street, Salt Lake City, UT, USA

hitokazu.matsushita@ldschurch.org
stephen.richardson@ldschurch.org

Abstract

In a typical TM-MT environment, translations for segments are provided from TMs based on matching criteria while the remaining segments are translated completely by MT. However, this binary approach does not always produce desirable translations. For example, even though a contiguous portion of a sentence to be translated may exactly match a TM entry or a frequently occurring sub-segment in many TM entries, if the match for the entire sentence does not exceed some arbitrary threshold, the smaller matches will not be used, and the entire sentence will be machine translated, resulting in a less than perfect translation, even for those portions that could have matched perfectly. In this report, we describe our approach to flexibly combine the capability of MT and TMs, applying exact TM matches to sub-segments of sentences and allowing MT to handle the remaining portions of the sentences. We specifically focus on the scenario where the matched phrases, clauses, and/or sentences are quotations in the text to be translated.

1 Introduction

In recent years, individual translators, language service providers, and large enterprises have more actively utilized hybrid systems of translation memories (TMs) and machine translation (MT) to increase translation productivity (Reinke, 2013). In a typical TM-MT environment employed by those translation professionals, translations for segments are provided from TMs based on matching criteria (e.g., high fuzzy-match scores), while the remaining segments are translated completely by MT. However, this binary approach does not always produce desirable translations. For example, even though a contiguous portion of a sentence to be translated may exactly match a TM entry or a frequently occurring sub-segment in many TM entries, if the match for the entire sentence does not exceed some arbitrary threshold (typically around 70%), the smaller matches will not be used, and the entire sentence will be machine translated, resulting in a less than perfect translation, even for those portions that could have matched perfectly. This is especially unfortunate if potentially matching sub-segments consist of frequently quoted or frequently occurring text for which only the exact human translations are acceptable in a production environment.

In this report, we describe our approach to flexibly combine the capability of MT and TMs, applying exact TM matches to sub-segments of sentences and allowing MT to handle the remaining portions of the sentences. We specifically focus on the scenario where the matched phrases, clauses, and/or sentences are quotations in the text to be translated.

2 Background

The Church of Jesus Christ of Latter-day Saints (henceforth, the Church) translates a wide variety of English materials into more than 100 languages to support communication among more than 15 million members around the world (Richardson, 2012). To facilitate its translation processes, the Church provides a TM-MT hybrid system using SDL WorldServer¹ and the Microsoft Translator Hub² for their human translators in the various locations of the world. The hybrid system functions based on the binary approach mentioned in the Introduction above, using a 75% fuzzy match threshold to determine whether the TM matches or MT outputs are used as translation candidates.

As a religious organization, the Church's scriptural canon consists of four volumes: the Holy Bible, the Book of Mormon, the Doctrine and Covenants, the Pearl of Great Price.³ These scriptures are translated in many languages under the strict supervision of Church authorities.⁴ An important aspect of the Church's translation effort is that many of the documents to be translated contain verses or phrases quoted from these scriptures. When scripture text is quoted in Church publications, the corresponding phrases or clauses in the current editions of the same scriptures must be strictly used when the publications are translated. In general, scripture quotes appear in Church publications in the following three forms:

1. All, or almost all, of a verse as a single segment
e.g., Peace be unto thy soul; thine adversity and thine afflictions shall be but a small moment;⁵
2. All, or almost all, of a verse as part of a segment
e.g., Remember the yearning hope of a father as expressed by John: "I have no greater joy than to hear that my children walk in truth."⁶
3. A smaller part of a verse as part of a segment
e.g., God has said that His purpose is "to bring to pass the immortality and eternal life of man" (Moses 1:39).

In item 1 above, the corresponding TM entry will be matched with a 100% or high fuzzy match score because the entire verse is stored in the TM. On the other hand, the verse surrounded by the quotation marks in item 2 above cannot be matched by the TM due to the text preceding the quote, which lowers the fuzzy match score to 68%.⁷ Furthermore, the quote in item 3 is even more problematic since it is only a part of the entire verse of scripture and it is also only part of the segment. Although the quote in item 3 is part of the most frequently quoted verse in all of the Church's scriptures, the desired TM match cannot be applied due to the low fuzzy match score, and it will be machine-translated along with the rest of the segment.

Our focus in this study is to apply correct quote translations to segments containing quotes like those in items 2 and 3 above. In this study, we investigate a method to collect scripture quotes to form a quotation TM and apply translations for quotes embedded in segments in

¹<http://www.sdl.com/cxc/language/translation-management/worldserver/>

²<https://hub.microsofttranslator.com>

³<http://www.scriptures.lds.org>

⁴The Bible is an exceptional case; typically, specific editions translated by authoritative organizations in the various countries or areas are approved for Church use.

⁵Doctrine and Covenants 121:7.

⁶John 1:4.

⁷Calculated by the character-based Levenshtein distance.

a manner similar to TM matches, while the remaining parts of those segments are machine-translated. This approach can be effective in reducing the amount of post-editing by human translators if the quotes recurring in Church documents are correctly identified and the proper translations are applied. While this approach is particularly relevant to our context, we feel that it could be applied in any context where correct human translations of quotes from canonical sources must be included in the publications of organizations attempting to use MT. In the following section, we discuss the previous work related to the focus of this study.

3 Related Work

Many methods have been proposed for combining TM and MT technology. A popular approach is to integrate TM matches directly into the MT decoder. Biçici and Dymetman (2008) extract phrases commonly found in a sentence to be translated along with their fuzzy-matched TM entries and put weights on those phrases in the MT phrase table to favor them during decoding. Dandapat et al. (2012) investigate enhancement of MT outputs using a hybrid example-based and statistical MT system in a approach similar to that of Biçici and Dymetman (2008). Wang et al. (2013) propose a phrase-based translation model which includes TM information as parameters of the model in order to dynamically choose the best phrase matches during decoding. Li et al. (2016) extend this idea and apply it to syntax-based MT systems to address translations of non-contiguous phrases.

Koehn and Senellart (2010a,b) discuss an approach to extract matching portions in a sentence to be translated and a TM entry, and then constrain the MT decoder with an XML frame to translate only the unmatched portions using hierarchical translation models combined with suffix arrays. A similar approach is also reported by Zhechev and Van Genabith (2010). Furthermore, Ma et al. (2011) and He et al. (2011) extend the approach by Koehn and Senellart (2010a) and investigate a method to identify the most promising translation among all the fuzzy-matched TM entries using support vector machines (SVMs) trained on various linguistic features extracted from TM data.

Other studies focus on an MT-system-agnostic approach, where fuzzy-matched TM entries are identified and applied to sentences to be translated before they are sent to MT systems. Espla-Gomis et al. (2011) and Ortega et al. (2014) investigate a method to patch sentences to be translated with elements in fuzzy-matched TM entries to improve the output from a rule-based MT system in a computer-aided translation (CAT) environment. He et al. (2010) discuss automatic quality estimation of statistical MT outputs, which determines whether the MT outputs are suitable for post-editing based on an SVM approach with features similar to those described in He et al. (2011).

In this study, we employ an MT-system-agnostic approach using XML frames. We apply TM entries to input segments before submitting them to an MT system, marking the entries that match quotes in the segments with XML frames to constrain the MT system to processes only the text outside of the framed portions of the segments, similar to the approach of Koehn and Senellart (2010a). Unlike the previous studies summarized above, however, which consider cases where various elements of TM entries are used to repair parts of sentences to be translated wherever they may be applied, we confine our focus solely to the application of TM entries to quoted text, as we described in Section 2 above. This is because of the nature of scripture quotes, where the approved translations must absolutely be used. In this sense, the limitation we impose is essential in our production environment, where it must be highly likely that we impact the quality of MT output in only a positive way. In the following sections, we describe our method to prepare scripture TM data in the form of a quotation TM to be used in the patching process.

4 Quotation TM Creation

As shown in the examples in Section 2, the quotes used in sentences are often small portions of the original scripture verses, such as syntactic constituent phrases (e.g., NP and VP) and dependent clauses. In such cases, we cannot apply the scripture translation units (TUs) in the TM directly because most of them are aligned at the verse or sentence level. To properly apply sub-sentential quotes to sentences to be translated, the TUs must be re-aligned at a much finer level of granularity. Several previous studies used aligned phrases in the phrase tables generated during the MT training process (e.g., Biçici and Dymetman 2008; Dandapat et al. 2012). However, these phrases are non-syntactic sequences of words and can be very noisy due to errors made in the word alignment process.

To obtain more finely aligned and linguistically well-formed scripture quotes, we process the original scripture TUs with a bilingual segment alignment algorithm proposed by Deng et al. (2007). The reasons we chose this algorithm are:

1. The algorithm performs alignment processes using small syntactic entities generated by segmentation based on non-terminal punctuation marks.
2. The algorithm considers non-monotonic alignment cases, which frequently occur in the alignment process of linguistically divergent language pairs.

With this algorithm, we obtain scripture TUs aligned at a sub-sentential level. We create a quotation TM with these scripture TUs and use that TM to process sentences in the subsequent quote application process. In the following subsections, we describe the alignment algorithm in detail.

4.1 Two-Step Segment Alignment

The traditional method for bilingual segment alignment is based on dynamic programming (DP) with the assumption that the segments can be aligned monotonically (Gale and Church 1993; Moore 2002, *inter alia*). This assumption is reasonably effective if one aligns sequences of full sentences in a bitext, but it is not so applicable if the alignment process is at the sub-sentence level, especially for language pairs with a significant linguistic distance from one another such as English and Japanese. To overcome this issue, we use the two-step alignment approach proposed by Deng et al. (2007), which aligns segments with DP and divisive clustering (DC) algorithms in a sequential manner. With this approach, the sub-sentential segments are aligned both monotonically and non-monotonically, and desirable quote TUs with finer granularity are collected.

4.1.1 Monotonic DP Alignment

The segment alignment process using DP is typically based on probabilistic models that employ features such as segment lengths and word alignment probabilities (Braune and Fraser, 2010; Mújdricza-Maydt et al., 2013). Deng et al. (2007) use length and word alignment features based on the Bayesian hierarchical model in Figure 1. In this figure s represents source-language (SL)

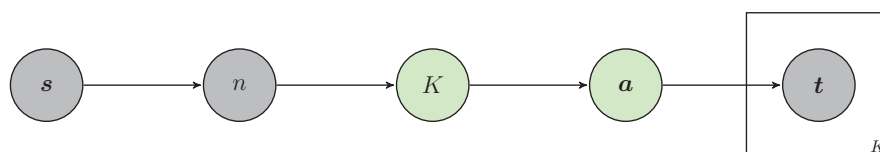


Figure 1: Graphical Model of DP Alignment (Deng et al., 2007)

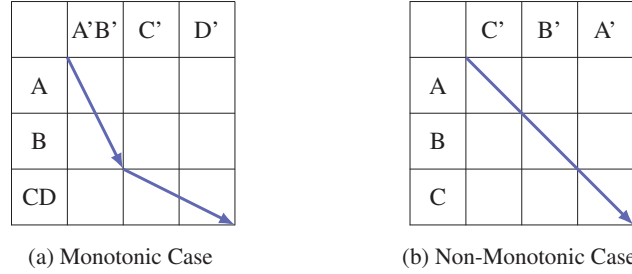


Figure 2: Examples of DP Alignment with Monotonic and Non-Monotonic Pairs

segments of a certain length, measured by the segment count; n represents the number of target-language (TL) segments; K represents the number of aligned segment pairs; \mathbf{a} represents the sequence of K aligned segment pairs (i.e., “beads” in Brown et al. 1991); and \mathbf{t} represents the target segments, which can be split into K chunks to form aligned TUs with \mathbf{s} . The problem is to estimate K and \mathbf{a} using the observed values \mathbf{s} , n , and \mathbf{t} . Based on this assumption, the segment alignment model is formulated as follows:

$$P(\mathbf{t}|\mathbf{s}, n, K, \mathbf{a}) = \prod_{k=1}^K P(\mathbf{t}_{a_k}|\mathbf{s}_{a_k}), \quad (1)$$

where

$$P(\mathbf{t}_{a_k}|\mathbf{s}_{a_k}) = \frac{P(u|v)}{(v+1)^u} \prod_{j=1}^u \sum_{i=0}^v t(f_j|e_i). \quad (2)$$

$P(\mathbf{t}_{a_k}|\mathbf{s}_{a_k})$ is an extension of the IBM1 model for an individually aligned TUs (See Brown et al. 1993). $P(u|v)$ in Equation 2 is the segment length model where u is the word count of SL segments in the bead a_k , v is the word count of the TL segments in a_k , and $t(f_j|e_i)$ is the word alignment model where f and e indicate the SL and TL words in the bead, respectively. The DP algorithm searches for the optimal values of K and \mathbf{a} which determine the best aligned TU sequence by maximizing $P(\mathbf{t}|\mathbf{S}, n, K, \mathbf{a})$ in Equation 1. This algorithm works effectively if the alignment process is monotonic, i.e., where there are no TL segments in a transposed order. Figure 2a shows a monotonic alignment example in DP. In this example, A, B, C, and D represent SL segments, and A', B', C', and D' are the corresponding translations. C and D are combined into one source segment, and A' and B' are combined into one target segment to artificially create differences in these hypothetical segment sequences. The optimal aligned TUs are correctly discovered with the alignment model in Equation 1 by pursuing the best scores yielded by the local model in Equation 2, as indicated by the arrows shown in Figure 2a. However, the non-monotonic case depicted in Figure 2b is problematic because the segment order in the TL is the complete opposite of that in the SL. In this case, the only valid alignment is A-B-C and C'-B'-A', which is the same as the entire original TU, unless the alignment algorithm allows for deletions and insertions (i.e., 1-0 and 0-1 mappings) in the bead types. Such transposed cases are highly likely to occur when the segments to be aligned are at the sub-sentence level. To address such problematic cases, we use the DC alignment method. In the following section, we describe this method in detail.

4.1.2 Divisive Clustering

The divisive clustering (DC) method described by Deng et al. (2007) is an effective approach to overcome the non-monotonic alignment problem. Figure 3 shows how the alignment is

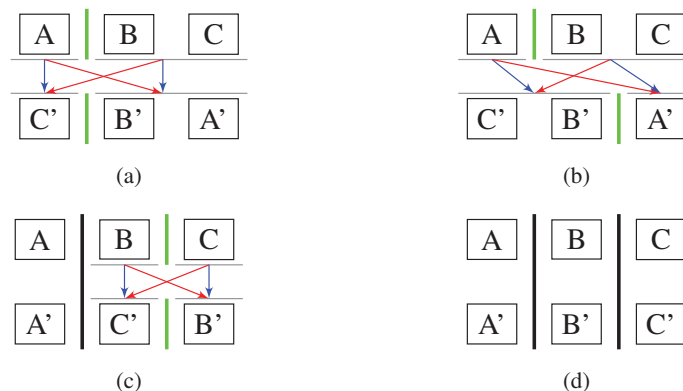


Figure 3: Example of Divisive Clustering Alignment

accomplished with the example case depicted in Figure 2b. First, the algorithm divides SL and TL segments, as shown by the green vertical lines in Figure 3a. Then it compares the divided segments in both monotonic and non-monotonic orders as indicated by the blue and red arrows in Figure 3a, and it records the respective alignment probabilities computed by the local model in Equation 2 as the costs. Next, it moves the split point of the TL segments, as shown in Figure 3b, and performs the same probability computation. The algorithm determines that the non-monotonic case in Figure 3b is the best among all four cost values, and reorders the TL segments so that A and A' and BC and C'B' are aligned, as shown in Figure 3c. Since A and A' consist of single segments, no further alignment process is applied. For BC and C'B', the same alignment process is followed, and the non-monotonic alignment case in Figure 3c is chosen. Since B-B' and C-C' consist of single segment pairs, the TL segments are simply reordered and the entire alignment process is completed.

In the overall alignment process, we use both DP and DC methods in a sequential manner, as described in Deng et al. (2007). We apply DP alignment to the original scripture TUs segmented only by terminal punctuation marks and then apply DC alignment to each of the resulting DP-aligned TUs, which are re-segmented by both terminal and non-terminal punctuation marks. We use this two-step alignment process in order to first obtain “large” aligned TUs with the DP alignment, and then process those TUs with the DC alignment to generate “small” aligned TUs. This approach is based on the assumption that the use of terminal punctuation marks between SL and TL is generally consistent; thus allowing the use of the time-efficient DP algorithm. Also, the resulting DP-aligned TUs narrow down the search space explored by the DC alignment, since the latter process is confined within each DP-aligned TU rather than the much larger original TU.

4.1.3 Iterative Segment Alignment

In typical scenarios, alignment algorithms collect TUs with high alignment probabilities computed by a simple model, such as a segment-length model, to obtain alignment features or train a word alignment model. With the features or model, the algorithm re-aligns the original TUs to identify aligned TUs with better accuracies as the final output (e.g., Moore 2002; Braune and Fraser 2010). This approach is effective only if the aligned TUs with high alignment probabilities are identified (i.e., precision-conscious alignment). However, this approach is not desirable in our scenario because quotes in sentences to be translated cannot be matched if only aligned TUs with high precision exist in the quotation TM. We need to collect aligned TUs with alignment scores that are as high as possible without sacrificing recall so that the quotation TMs can

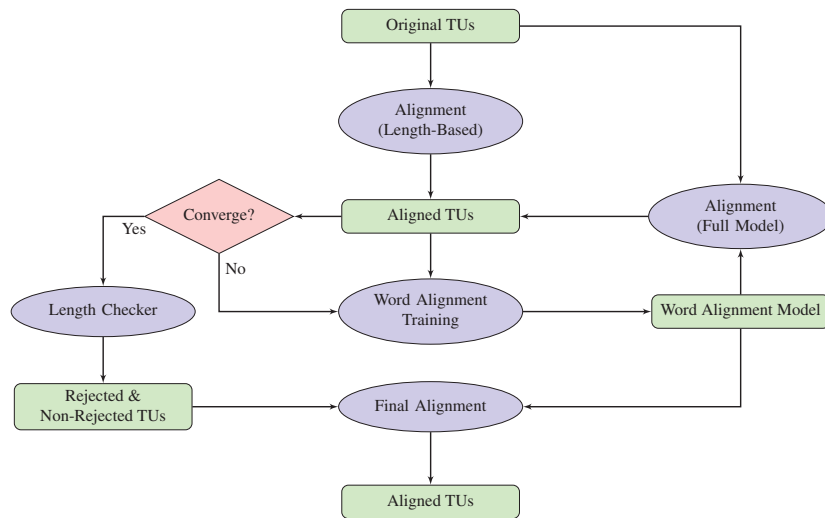


Figure 4: Iterative Segment Alignment Process

be used to match the largest number of potential quotes in the sentences.

To create a quotation TM, we use the alignment process illustrated in Figure 4, which begins with the original verse or sentence aligned TUs and iteratively identifies high probability sub-sentential TUs to be used in the creation of the quote TUs the TM will contain. First, the DP+DC algorithm aligns original TUs with a pre-computed length model to create the first collection of aligned sub-sentential TUs. This set of aligned TUs is then used for word alignment training. After the training process, the DP+DC process re-aligns the original TUs using the full model in Equation 2. If the alignment result is better than the previous alignment result, the new set of aligned TUs is used to train a new word alignment model and conduct the DP+DC alignment again. Once the alignment results converge and do not exhibit any improvement, the original TUs are aligned with the final version of the word alignment model in the final alignment process. Before this process, we evaluate the aligned TUs using a length checker based on the Poisson model used by Moore (2002) to avoid abnormal length discrepancies between the SL and TL segments of the aligned TUs. In this final alignment process, we aggressively re-align the TUs rejected by the length checker, segmenting them with all punctuation marks in both DP and DC processes. If the newly aligned TUs are accepted by the length checker, then we add them to the aligned TU set and export it as the output.

Figure 5 shows the increase in the aligned TU count for our English-Japanese (EN-JA) scripture TM data, which results from the iterative alignment process. The first alignment process using only the length model increases the aligned TU count slightly (44456 → 48899). Then there is a very substantial increase in the next iteration, where both length and word alignment models are used in the alignment process. In the next iterations, the aligned TU counts end up fluctuating somewhere between 117800 and 121550. We arbitrarily stopped the alignment process at iteration 15, assuming that the convergence point has been reached at or before this iteration. With the word alignment model created at this iteration, we aligned TUs using the aforementioned final alignment method, and obtained 143100 aligned TUs as the final output.

4.2 Quote Generation

Once the aligned TUs are generated, they are used to create quote TUs to be applied to sentences to be translated. To accommodate various types of quotes, we generate collections of segment n -

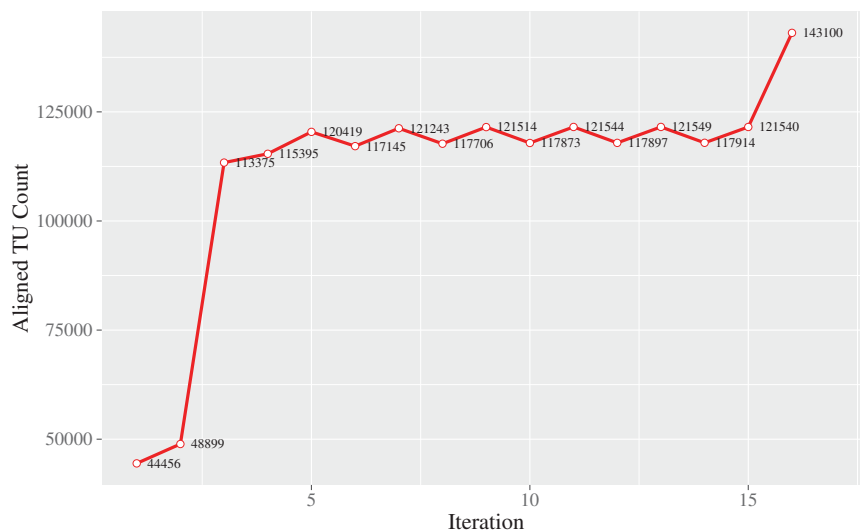


Figure 5: Aligned TU Increase over Alignment Iterations with EN-JA Scripture Dataset

grams with the aligned TUs. Figure 6 shows a simple example. Based on the identified aligned

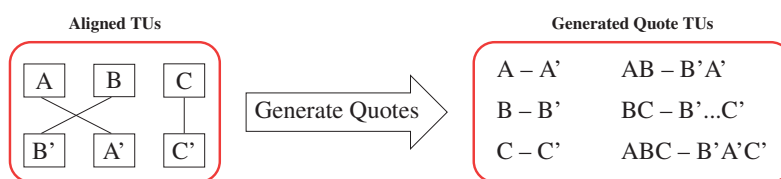


Figure 6: Example of Quote TU Generation using Aligned TUs

TUs, potential quote TUs are generated. In this case, six different quote TUs are generated based on the possible combinations of SL segments in the aligned TUs. Because the order of SL segments is different from that of the TL segments, appropriate treatments, such as swapping (e.g., A'B' → B'A' for AB) and ellipsis (e.g., B'...C' for BC) in Figure 6, need to be applied to the corresponding TL segments when segment n -grams are formed based on the SL segment order. To accomplish this, we keep track of the order of TL segments as they are aligned with SL segments using a generic tree data structure (i.e., a parent node with an arbitrary number of child nodes). The tree is then referred to in a depth-first order during the quote generation process as multiple SL segments are used to form a quote. If swapping or ellipsis cases are identified while traversing the tree, the corresponding treatments are applied. After quote TUs are generated and TL ordering is modified as needed, the TUs are stored in a quotation TM and applied to sentences to be translated as described in the following section.

5 Quote-Applied MT Inputs

Using the quotation TM, quote TUs are applied to sentences before they are machine-translated. Figure 7 shows an example of the quote application process. Once a sentence with an embedded quote is identified, the double- or single-quoted portion of the sentence is matched against the contents of the quotation TM. In our experiment, we used 98% as the fuzzy match threshold to

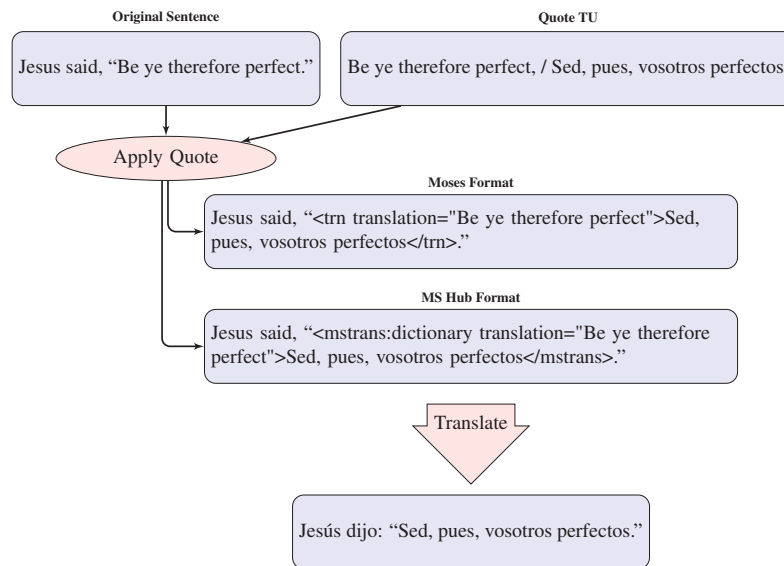


Figure 7: Example of Applying a Quote TU to a Sentence to be Machine-Translated

accommodate potential minor differences in punctuation. If a match is found, the identified quote TU is applied to the sentence using a system-defined XML frame. Figure 7 shows example XML-framed sentences for the Moses toolkit⁸(Koehn et al., 2007) and the Microsoft Translator Hub.⁹ The modified sentence is then sent to the MT system and translated.

6 Experiments

With the quote generation and application approach described above, we created scripture quotation TMs in eight language pairs and applied the quotes to test sets selected for experimentation. In the following section, we describe the experimental configuration.

6.1 Experiment Configuration

For the experiments, we used our in-house TM data for eight language pairs (with English as SL) to build trained MT systems and conduct the evaluation. The dataset sizes (based on the number of English segments) are shown in Table 1. The MT engine used for the experiments was the Microsoft Translator Hub.¹⁰ For system tuning, we provided separate tuning sets (2500 TUs) extracted from the same data source as the training sets. The test sets consisted of TUs collected from the speeches given at the two most recent semi-annual general conferences of the Church held in October 2015 and April 2016,¹¹ which are not contained in the training sets. We cleaned the training and test datasets with Okapi¹² and segmented them using in-house

⁸Any arbitrary tag names other than “trn” can be used. See <http://www.statmt.org/moses/?n=Advanced.Hybrid> for more detail.

⁹Usage is described at <https://social.msdn.microsoft.com/Forums/en-US/de55e04f-7bc8-4a03-8a6d-13d43b2f739b/wordaround-about-donottranslate-list?forum=translatorhub>

¹⁰<https://hub.microsofttranslator.com>

¹¹The texts of these speeches are available in over 90 languages at <https://www.lds.org/general-conference/2015/10> and <https://www.lds.org/general-conference/2016/04>, respectively.

¹²<http://okapiframework.org/>

Language	Training Set Size		Test Set Size	
	#TUs	#Tokens	#Sentences	#Tokens
Spanish (ES)	2,103,203	28,500,765	7,228	170,220
Portuguese (PT)	1,605,441	23,315,397	8,096	161,120
German (DE)	1,178,506	17,447,421	8,305	161,604
French (FR)	1,544,140	21,762,962	8,358	171,837
Italian (IT)	1,538,076	22,717,683	8,851	183,928
Russian (RU)	1,247,934	17,265,925	9,093	184,289
Japanese (JA)	1,026,201	15,801,673	6,385	150,647
Chinese (ZH)	1,233,531	18,809,448	6,208	136,248

Table 1: MT Training and Test Datasets

SRX¹³ rules. The data were also sentence-aligned with an aligner similar to that of Moore (2002). Unaligned segments were excluded to reduce noise. The scripture quotation TMs for the eight languages were prepared using the quote generation process described in Section 4. They included all the quote TUs derived from verse-aligned TMs containing the four volumes of scripture described in Section 2.

7 Results

We examined the effectiveness of the quote application method with an automatic evaluation using case-sensitive BLEU (Papineni et al., 2002) and a human evaluation using Dynamic Quality Framework¹⁴ (DQF, TAUS 2016). The details are described in the following sections.

7.1 Automatic Evaluation

Language	#Sentences	No Quotes Applied	Quotes Applied	Difference
ES	7,228	38.18	39.58	1.40
PT	8,096	40.15	41.63	1.48
DE	8,305	27.11	29.41	2.30
FR	8,358	39.82	41.02	1.20
IT	8,851	37.36	38.73	1.37
RU	9,093	28.15	29.59	1.44
JA	6,385	19.58	22.41	2.83
ZH	6,208	18.68	21.72	3.04

Table 2: Translation Quality Results (BLEU) of All Test Sentences

Table 2 shows the evaluation results for the test set of each language. The column “No Quotes Applied” indicates the BLEU scores for the entire test sets without applying any quotes (i.e., regular MT output). The column “Quotes Applied” shows the scores for the test sets with scripture quotes applied to the source sentences before being machine-translated. The column “Difference” indicates the difference in BLEU score between No Quotes Applied and Quotes Applied. As expected, applying quotes produces better results across all eight languages by

¹³<http://www.ttt.org/oscarStandards/srx/>

¹⁴<http://dqf.taus.net>

roughly 1.2 to 3 BLEU points, since we are substituting accurate reference translations in portions of the MT output. Of course, the improvement could also be offset somewhat by agreement problems and other grammatical anomalies caused by inserting the quoted text. But this modest improvement in BLEU score across the entire test set does not really reflect the true impact of quote application, since it is also dependent on the number of sentences in the test sets that actually contain quotes. To focus particularly on those test sentences containing quotes, we computed their BLEU scores separately. These scores and their relative impact are shown in Table 3. In this case, the score differences are much more significant, extending roughly from

Language	#Sentences	No Quotes Applied	Quotes Applied	Difference
ES	426	47.18	59.52	12.34
PT	408	50.55	61.70	11.15
DE	401	41.78	66.42	24.64
FR	415	52.80	64.47	11.67
IT	428	45.47	59.19	13.72
RU	280	41.93	61.31	19.38
JA	369	22.44	44.28	21.84
ZH	357	25.88	48.72	22.84

Table 3: Translation Quality Results (BLEU) of Quote-Applied Sentences

11 to 24 BLEU points, and convincingly demonstrating the positive effect of quote application. In comparing the differences across the eight languages, we observe that quote application has a somewhat lesser positive effect on those languages that are closer to English in grammar and syntax (e.g., Spanish, Portuguese, French, and Italian, with differences between 11 and 13 BLEU points), and a greater positive effect on the languages that are more divergent from English (e.g., German, Russian, Japanese, and Chinese, with differences between 19 and 24 BLEU points). This only makes sense, given that the BLEU scores of these latter systems are generally lower than the former ones.

7.2 Human Evaluation

Table 4 shows the DQF MT system comparison results of three languages (Portuguese, German, and Japanese). For this evaluation, we randomly chose 100 translations of the quote-applied sentences and the corresponding regular MT translations. The evaluation of each language was conducted by two bilingual raters: Rater 1 was a non-translator, and Rater 2 was a professional translator. In the evaluation process, the raters were asked to rank the translations of each of the 100 English sentence (ties were allowed). The percentages in the third and fourth columns of Table 4 are weighted rank scores reported by the DQF system, and the numbers in the parentheses indicate the number of times that the raters chose the translations generated by a particular method (Quotes Applied or No Quotes Applied) as being better than the other. For these numbers, we excluded the ties. The fifth column shows the number of rankings in agreement between the two raters.

Overall, the Quotes Applied translations were chosen by the raters more often than the No Quotes Applied translations across all three languages. This correlates strongly with the BLEU score results in Table 3. In particular, the Quotes Applied translations were most often preferred by German raters. For the other two languages, although the professional translators were more conservative about choosing Quotes Applied translations than non-translators, they still preferred Quote Applied translations significantly more often than No Quote Applied ones. We speculate that accurate machine translations of the quoted text, together with poten-

Language	Rater	Weighted Ranking Score		#Agreement
		No Quotes Applied	Quotes Applied	
PT	1	38.80% (6)	61.20% (77)	82
	2	42.12% (20)	57.88% (69)	
DE	1	36.93% (7)	63.07% (87)	96
	2	37.54% (7)	62.46% (84)	
JA	1	37.30% (5)	62.70% (84)	87
	2	44.05% (12)	55.95% (52)	

Table 4: Human Evaluation Results of Three Languages

tial agreement and other grammatical issues arising from the insertion of the quotes, may have influenced some of the choices made by the raters. Nevertheless, the results of these three language evaluations confirm the potential benefit of the quote application method in the machine translation of documents containing frequent scripture quotes.

8 Conclusions

In this paper we discussed our approach to enhance MT output using quotation TMs. We demonstrated the positive effect of the quote application process with real-world data for eight major languages used in a production translation environment, as measured by both automatic and human evaluations. A key factor in the success of creating and using a quotation TM, which is also a unique contribution of this work, is the generation of quote TUs, which are linguistically well-formed sub-sentential quote segments, aligned with their corresponding translations with a high degree of accuracy.

Although our focus was particularly on the translation of quotes in documents from the religion domain, our approach may be utilized in other translation scenarios and domains as well. For example, a document to be translated may contain quotes from canonical or standard texts in a heterogeneous domain. The texts containing those quotes can be processed to create a quotation TM, and quote TUs from this TM can be applied to sentences to be translated if the quoted portions in the sentences are correctly identified. Such simple domain adaptation approaches can be effective in translating mixed-domain documents more accurately in a production environment.

For future work, we will investigate approaches to dynamically generate quote TUs based on finding quotable text in a sentence without relying on segments generated by punctuation marks. A challenge in this case will be to capture the corresponding translation for the quotable text in order to create the sub-sentential quote TU. To address this problem, we will explore the idea of forced alignment in speech recognition, or a similar concept, as a good starting point (e.g., Alkhoul et al. 2016). If accurate forced alignment of words in each quote TU is possible, and if it can be obtained automatically with existing or new techniques, then it will be possible to extract the corresponding TL segment and form a quote TU dynamically.

Acknowledgments

We thank William Byrne at SDL and Cambridge University for making the original sentence alignment code available. We also thank Ryan Lee at the LDS Church for providing support for the experiments and for his insightful comments on this study.

References

- Alkhouli, T., Bretschner, G., Peter, J.-T., Hethnawi, M., Guta, A., and Ney, H. (2016). Alignment-based neural machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 54–65, Berlin, Germany. Association for Computational Linguistics.
- Biçici, E. and Dymetman, M. (2008). Dynamic translation memory: using statistical machine translation to improve translation memory fuzzy matches. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 454–465. Springer.
- Braune, F. and Fraser, A. (2010). Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 81–89. Association for Computational Linguistics.
- Brown, P. F., Lai, J. C., and Mercer, R. L. (1991). Aligning sentences in parallel corpora. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, pages 169–176. Association for Computational Linguistics.
- Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.
- Dandapat, S., Morrissey, S., Way, A., and Van Genabith, J. (2012). Combining EBMT, SMT, TM and IR technologies for quality and scale. In *Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)*, pages 48–58. Association for Computational Linguistics.
- Deng, Y., Kumar, S., and Byrne, W. (2007). Segmentation and alignment of parallel text for statistical machine translation. *Natural Language Engineering*, 13(3):235–260.
- Espla-Gomis, M., Sánchez-Martínez, F., Forcada, M. L., et al. (2011). Using machine translation in computer-aided translation to suggest the target-side words to change. Machine Translation Summit.
- Gale, W. A. and Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational linguistics*, 19(1):75–102.
- He, Y., Ma, Y., van Genabith, J., and Way, A. (2010). Bridging SMT and TM with translation recommendation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 622–630. Association for Computational Linguistics.
- He, Y., Ma, Y., Way, A., and van Genabith, J. (2011). Rich linguistic features for translation memory-inspired consistent translation. In *Proceedings of the Thirteenth Machine Translation Summit*, pages 456–463.
- Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

- Koehn, P. and Senellart, J. (2010a). Convergence of translation memory and statistical machine translation. In *Proceedings of AMTA Workshop on MT Research and the Translation Industry*, pages 21–31.
- Koehn, P. and Senellart, J. (2010b). Fast approximate string matching with suffix arrays and A* parsing. In *Meeting of the Association for Machine Translation of the Americas (AMTA)*.
- Li, L., Escartín, C. P., and Liu, Q. (2016). Combining translation memories and syntax-based smt. *Baltic Journal of Modern Computing*, 4(2):165–177.
- Ma, Y., He, Y., Way, A., and van Genabith, J. (2011). Consistent translation using discriminative learning: a translation memory-inspired approach. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1239–1248. Association for Computational Linguistics.
- Moore, R. C. (2002). Fast and accurate sentence alignment of bilingual corpora. In *Conference of the Association for Machine Translation in the Americas*, pages 135–144. Springer.
- Mújdricza-Maydt, É., Körkel-Qu, H., Riezler, S., and Padó, S. (2013). High-precision sentence alignment by bootstrapping from wood standard annotations. *The Prague Bulletin of Mathematical Linguistics*, 99:5–16.
- Ortega, J. E., Sánchez-Martínez, F., and Forcada, M. L. (2014). Using any machine translation source for fuzzy-match repair in a computer-aided translation setting. In *Proceedings of the 11th Biennial Conference of the Association for Machine Translation in the Americas*, volume 1, pages 42–53.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.
- Reinke, U. (2013). State of the art in translation memory technology. *Translation: Computation, Corpora, Cognition*, 3(1):27–48.
- Richardson, S. (2012). Using the Microsoft Translator Hub at the Church of Jesus Christ of Latter-day Saints. In *AMTA 2012, Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas*.
- TAUS (2016). TAUS Quality Dashboard: From quality evaluation to business intelligence. <https://www.taus.net/component/rsfiles/download-file/files?path=Reports%252FFree%2BReports%252FQualityDashboardDocument-%2BMarch2016.pdf>. [Online; accessed September 1, 2016].
- Wang, K., Zong, C., Su, K.-Y., et al. (2013). Integrating translation memory into phrase-based machine translation during decoding. In *ACL (1)*, pages 11–21.
- Zhechev, V. and Van Genabith, J. (2010). Seeding statistical machine translation with translation memory output through tree-based structural alignment. Association for Computational Linguistics.

Improving Machine Translation for Post-Editing via Real Time Adaptation

Dragos Munteanu

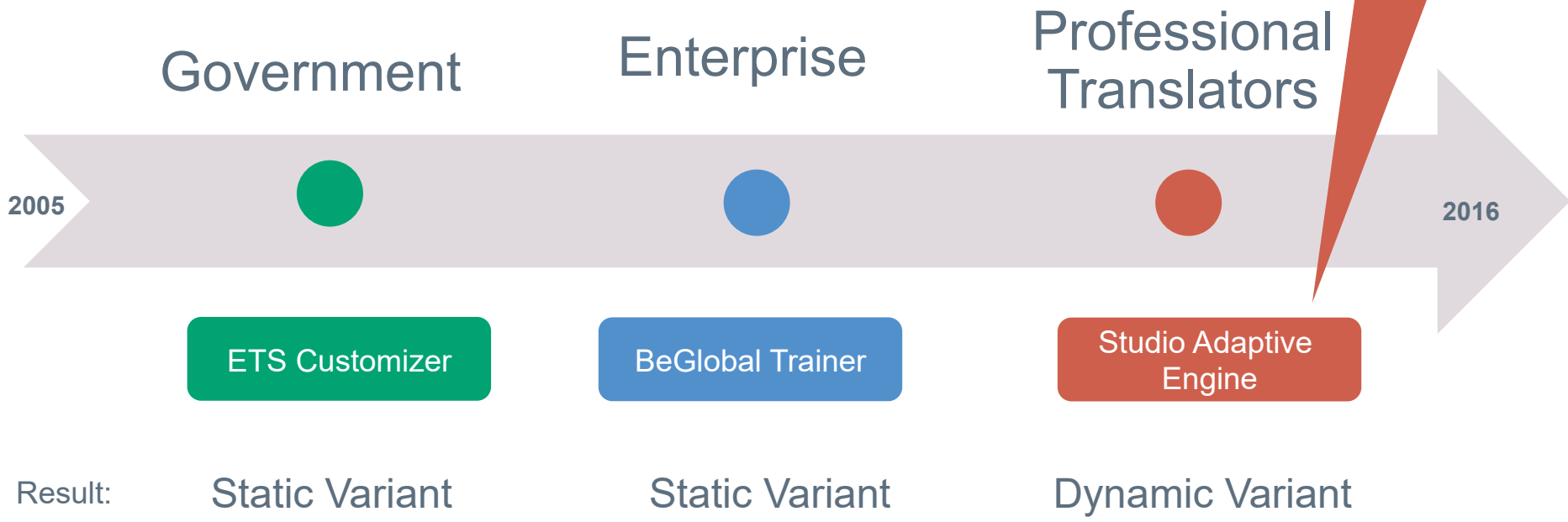
Director of Research and Development, Machine Translation



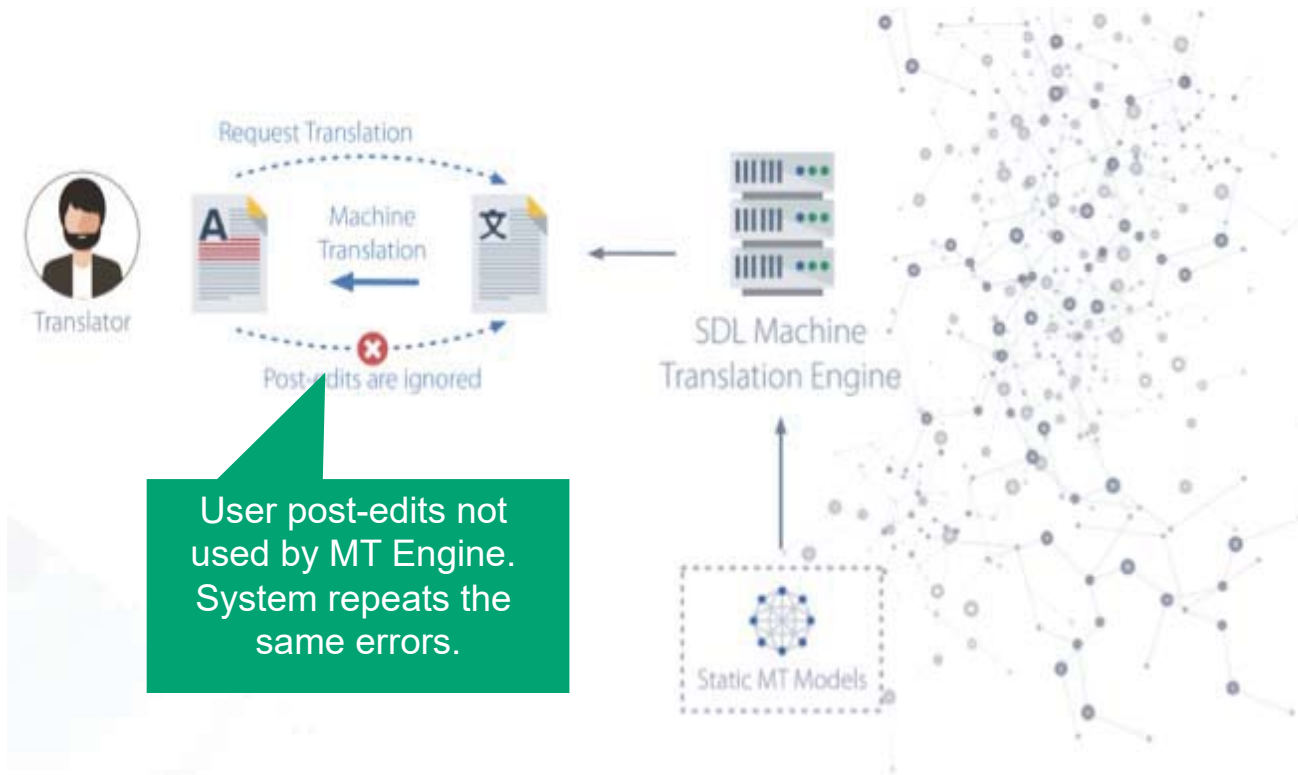
SDL Proprietary and Confidential

Adaptive Machine Translation at SDL

Focus of this presentation

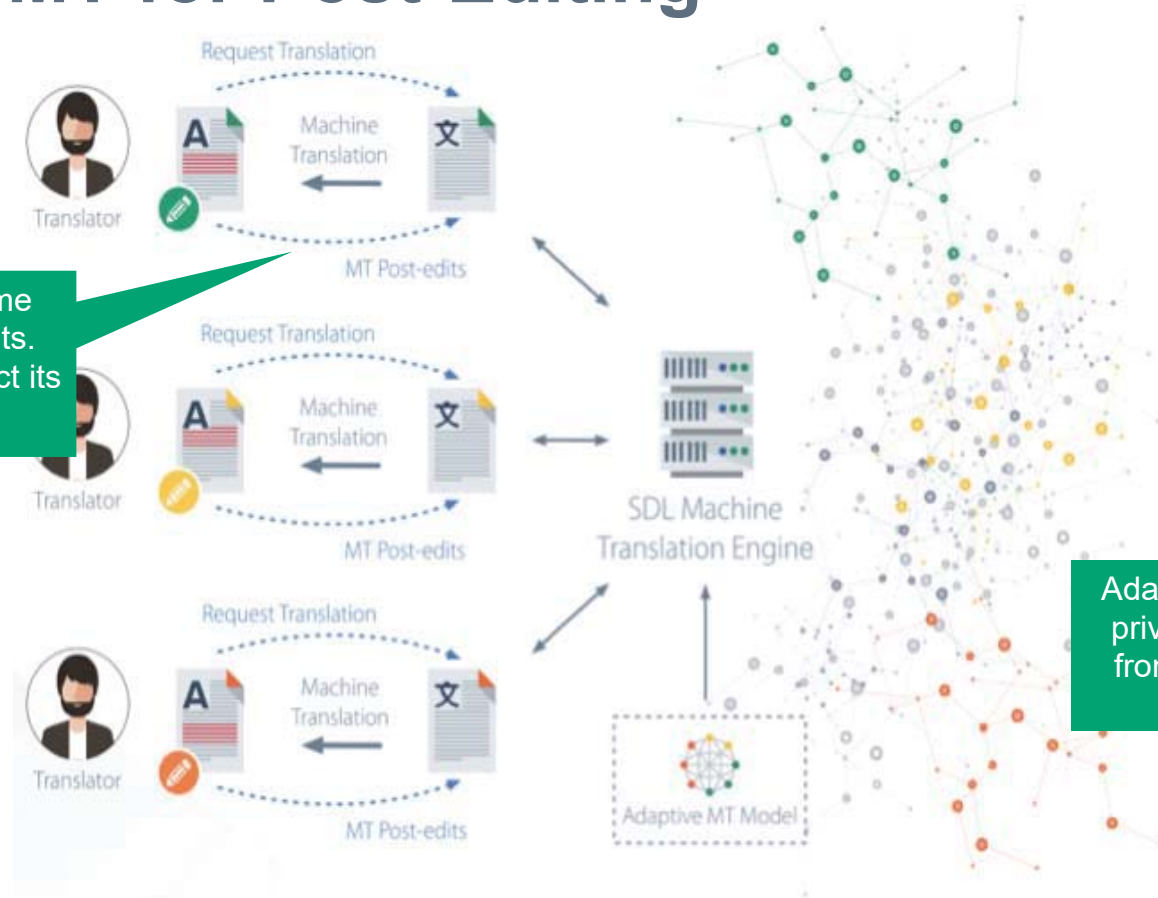


Standard MT for Post-Editing



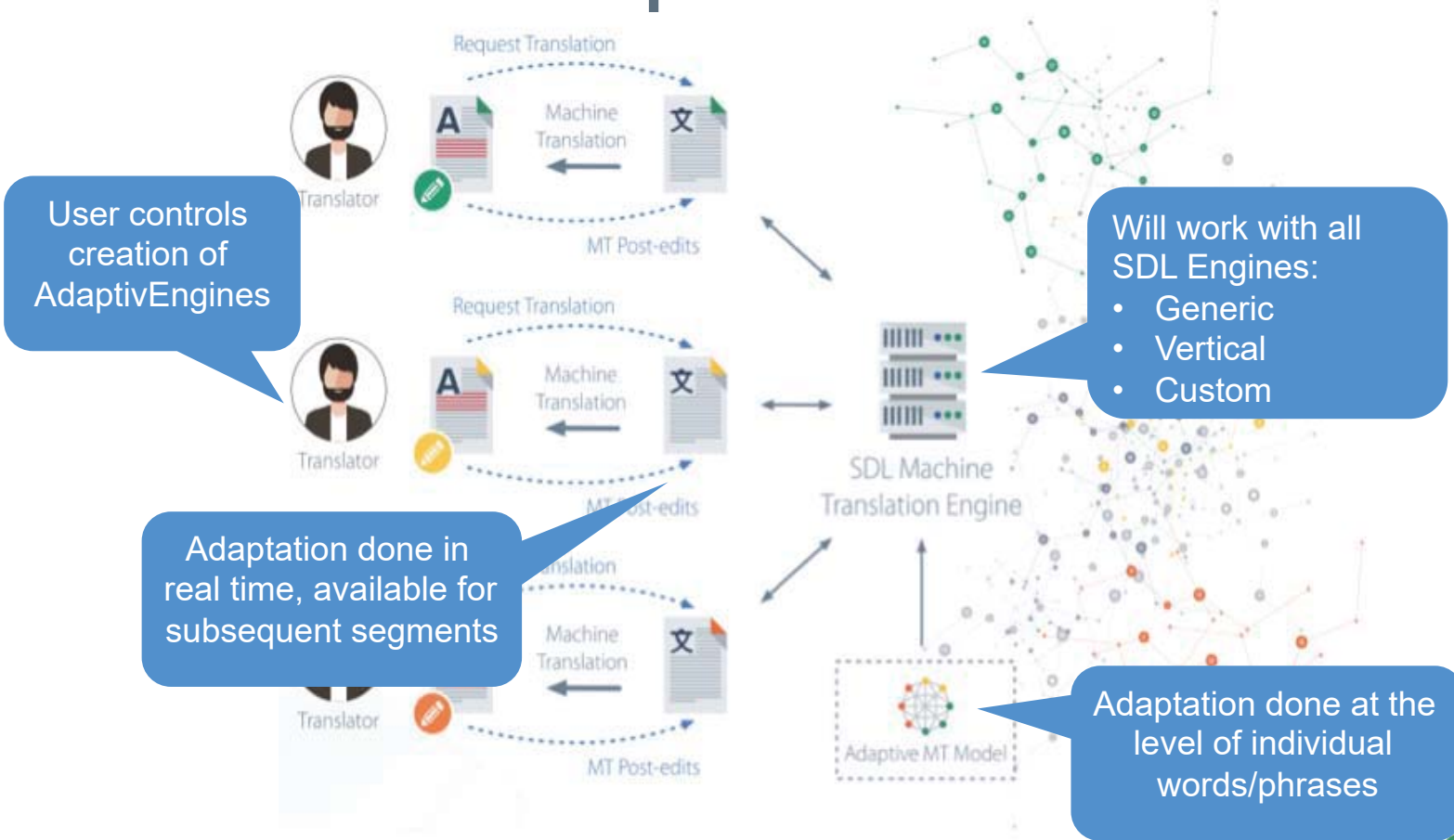
Adaptive MT for Post-Editing

Seamless and real-time learning from post-edits. System learns to correct its errors

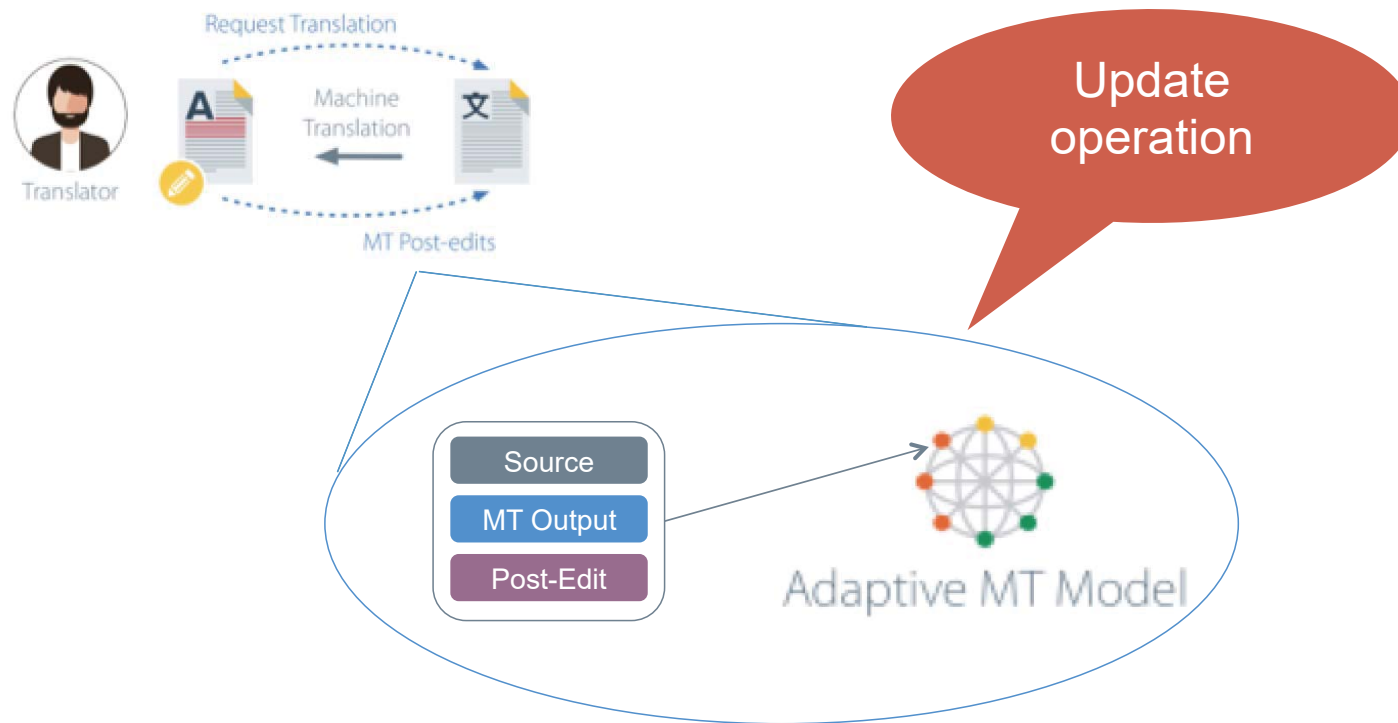


Adaptation provides data privacy: what is learned from a user is available only to that user

Features of SDL Adaptive MT



Learning from Post-Edits



Update Example

- An update of one of the statistical MT models, the translation model

Source	No further requirements are needed
MT	Pas d'autres exigences sont requises <i>No other requirements are needed</i>
PE	Aucune exigence supplémentaire n'est nécessaire <i>Any requirement additional is not necessary</i>

Update Example: Translation Model Adaptation

No further requirements are needed

Pas d'autres exigences sont requises

No other requirements are needed

Aucune exigence supplémentaire n'est nécessaire

Any requirement additional is not necessary

Original translations

needed -> requises

further -> autres

no further -> pas d'autres

requirements -> exigences

Update Example: Translation Model Adaptation

No further requirements are needed

Pas d'autres exigences sont requises

No other requirements are needed

Aucune exigence supplémentaire n'est nécessaire

Any requirement additional is not necessary

Good New Translations

needed -> ~~requis~~-nécessaire

further -> ~~autres~~-supplémentaire

no further requirements are -> aucune exigence supplémentaire n'est

Bad New Translations

are -> n'est

requirements -> exigence

no -> aucune

- Statistical features help choose good rules

Impact of Adaptive MT on Post-Edit productivity

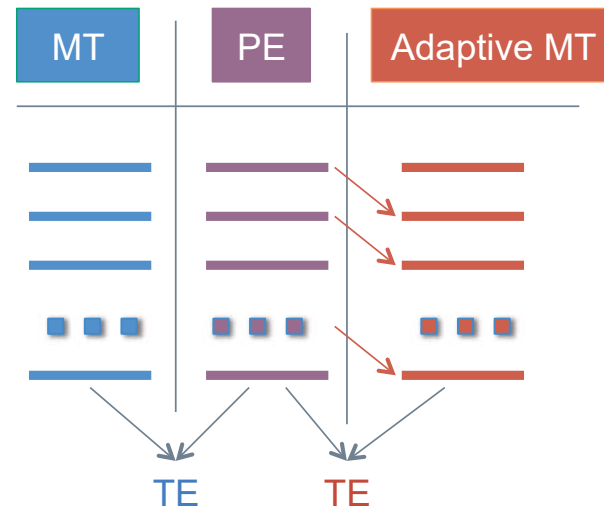
Test Data

- Two example projects, chosen to illustrate different AdaptiveMT impact
- Project A benefits little from AdaptiveMT while Project B benefits a lot

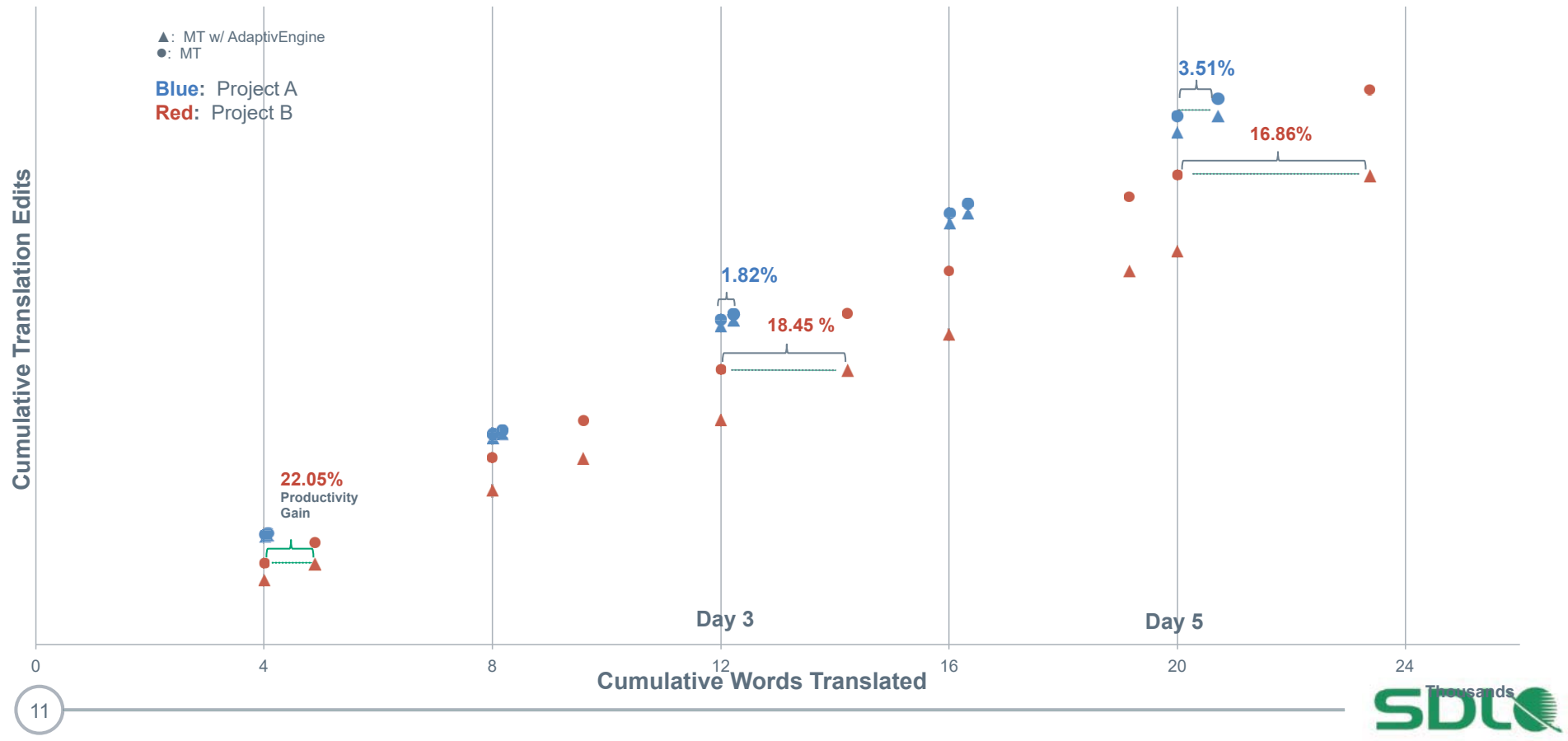
	Project A	Project B
Content Domain	Technical	Industrial
Total segments	2500	2500
Average words/segment	13	9
Longest segment length (words)	65	78

Experimental Framework

- Translate with both regular and Adaptive engine; measure Translation Edits required to create the same Post-Edit

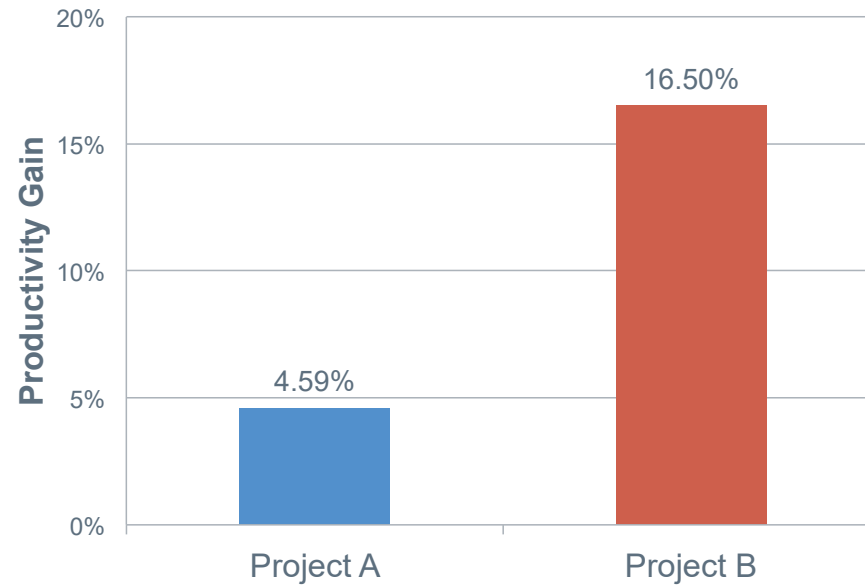


Impact of Adaptive MT: Productivity Gain



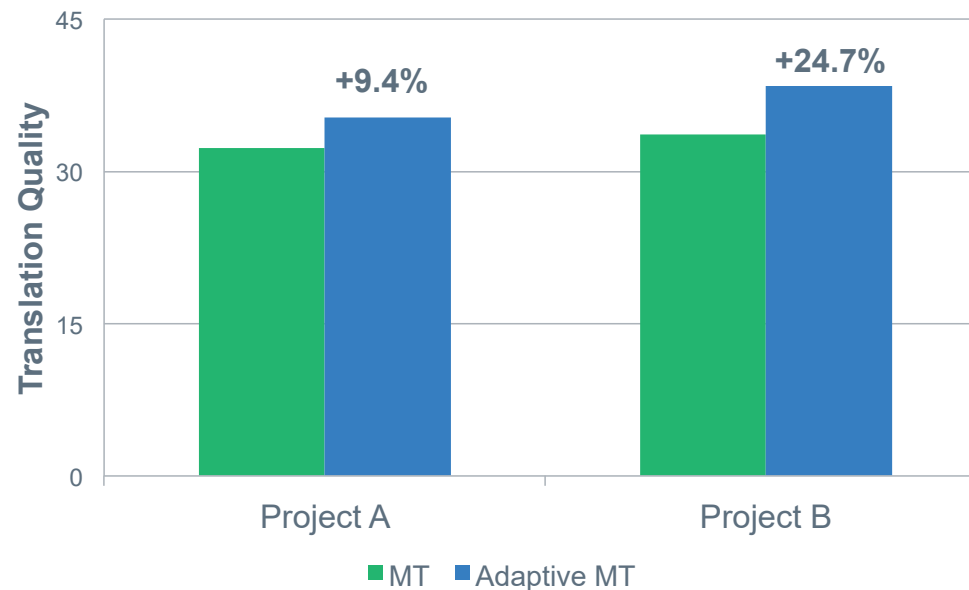
Impact of Adaptive MT: Productivity

- Overall productivity gain provided by the adaptive engine:



Impact of Adaptive MT: Translation Quality

- Dynamic statistical model adaptation to content lead to real-time and impactful quality improvement





Using Adaptive MT in SDL Trados Studio



SDL Trados Studio - Adaptive Engine Demo

File Home View Add-Ins Help

Project Settings Open For Translation Open For Review Open For Sign-off Batch Tasks Explore Containing Folder Add Files Delete Files Check Out Check In Cancel Check Out Create Project Package File details layout Go to last segment Bookmarks

Configuration Open File Actions GroupShare Packages Layout

Files

French (France)

Project folders

- Adaptive Engine Demo
- My Tasks
- Sent Tasks

Include subfolders

Welcome

Projects

Files

Reports

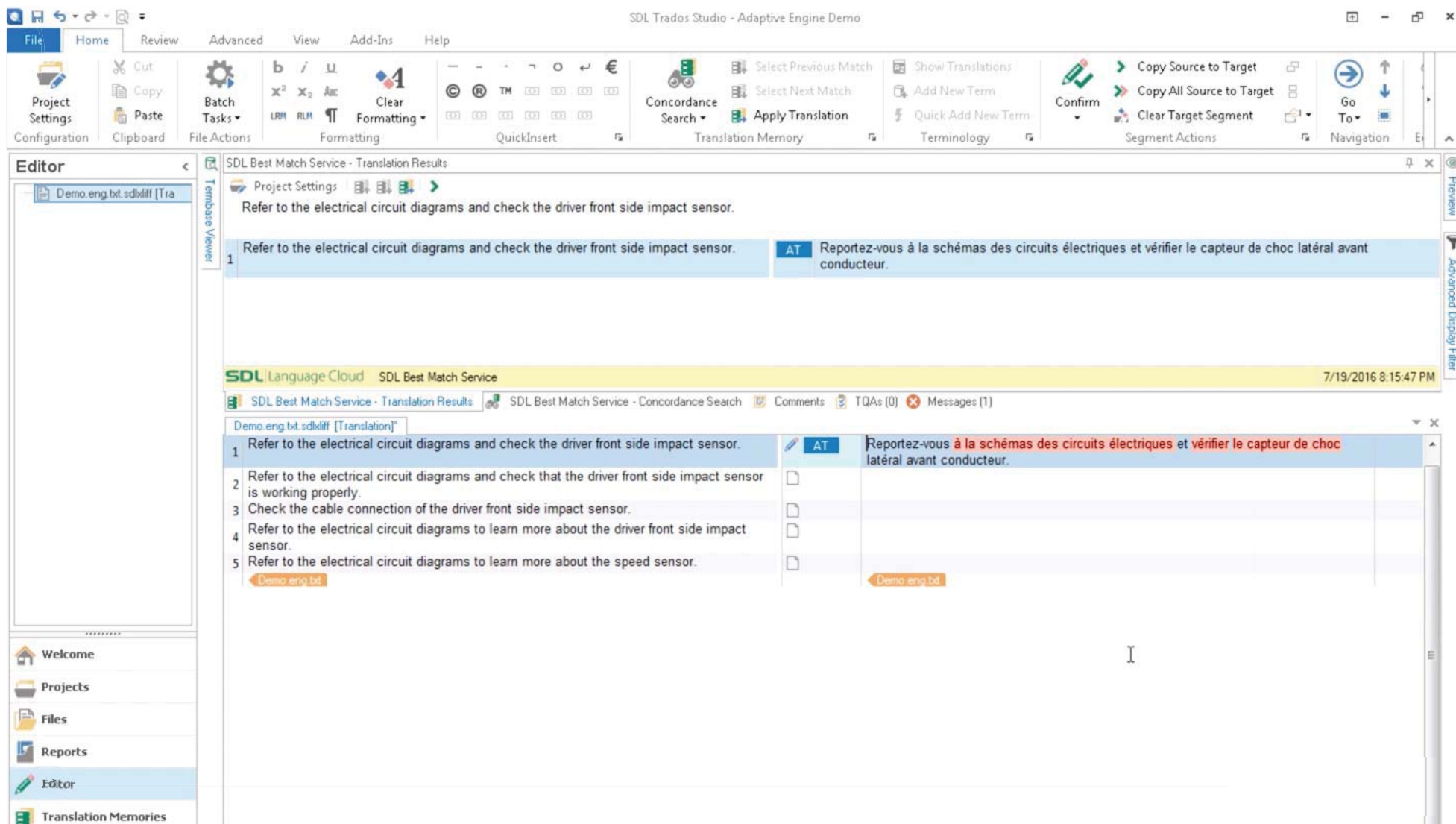
Editor

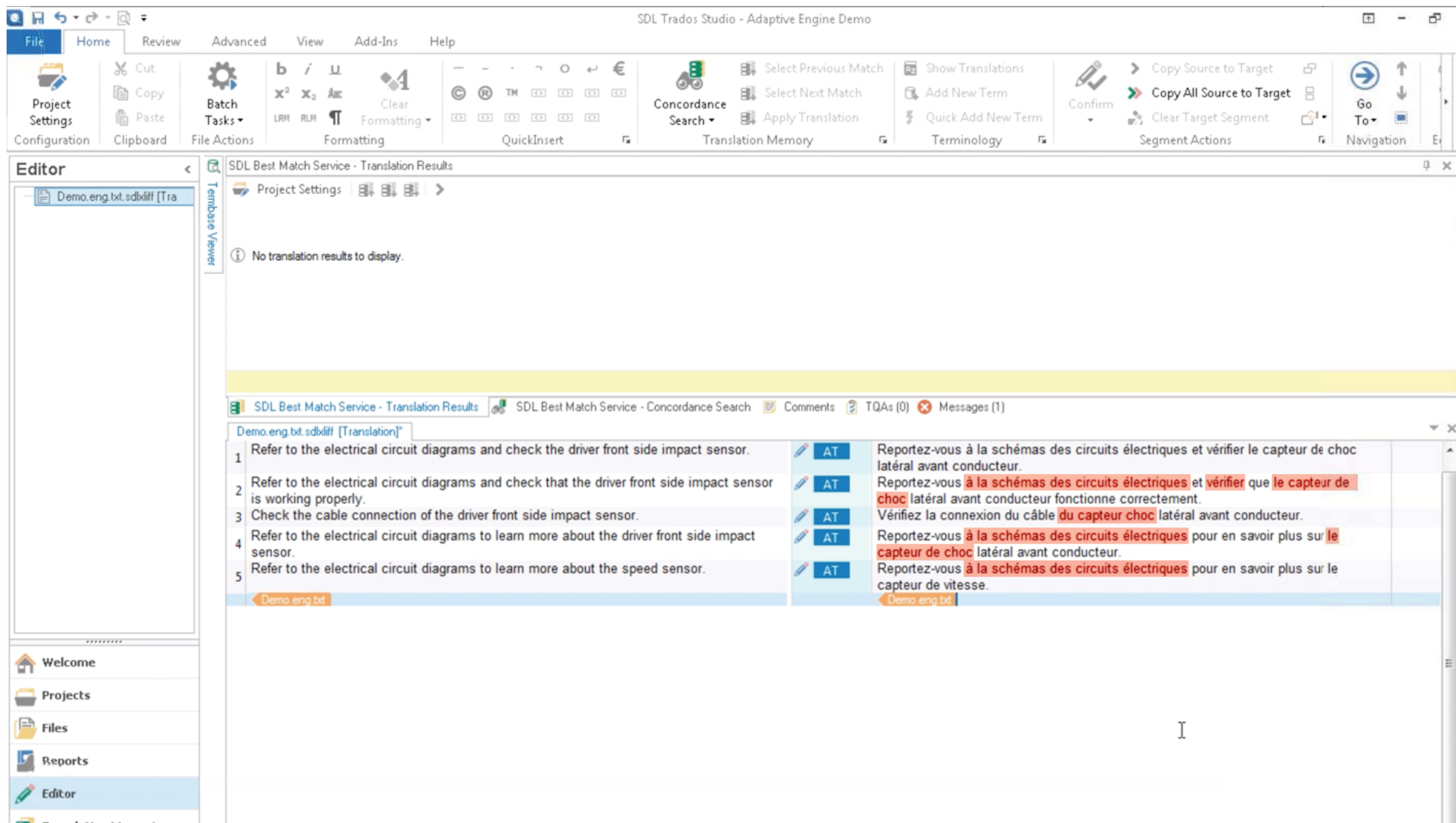
Translation Memories

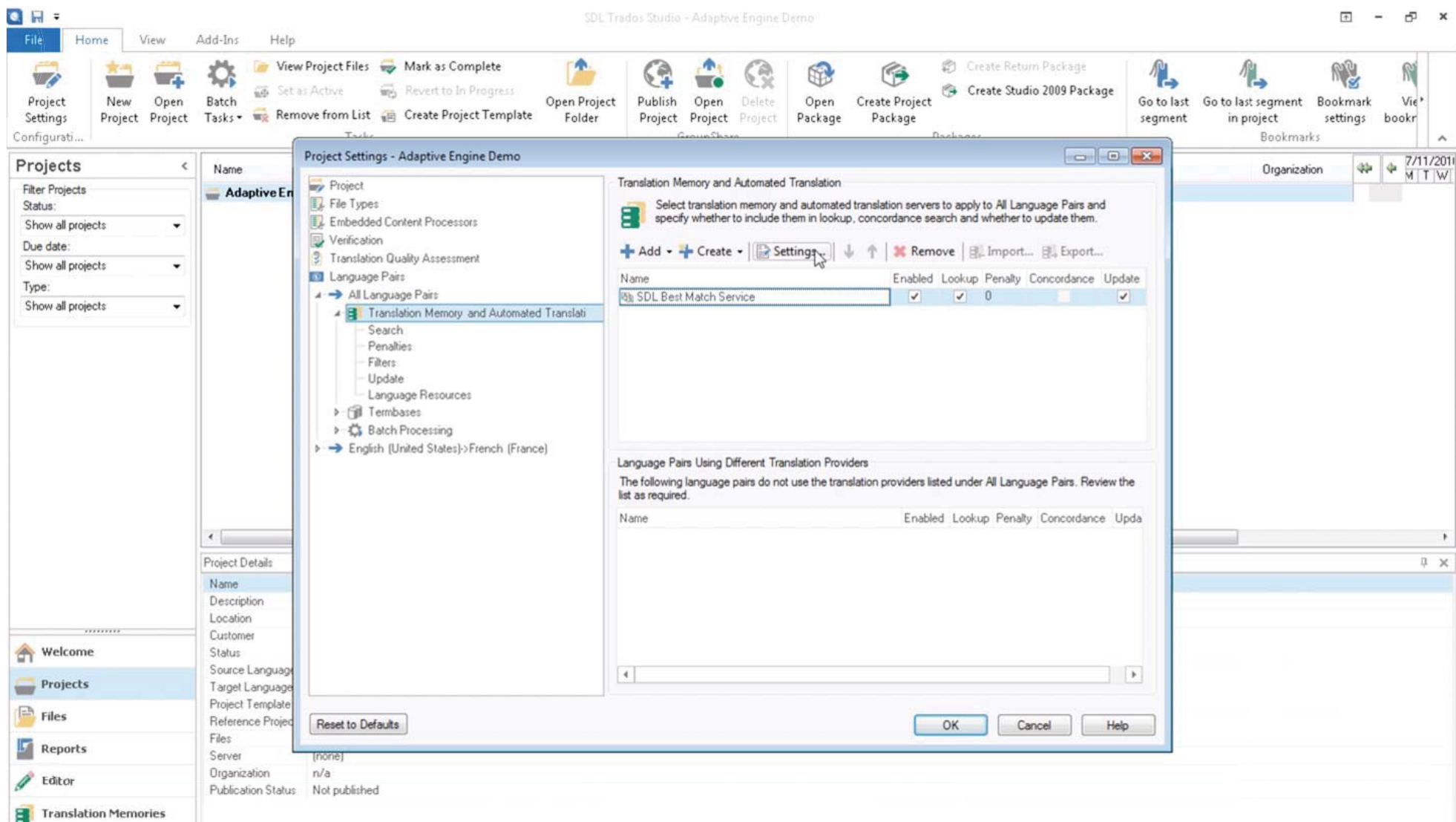
Name	Words	Status	Progress	Size	Usage	File Type Iden...	Path	Local State	Checked Out To
Demo.eng.txt.sdxliff	72	Unspecified		5 KB	Translatable	Plain Text v 1.0...	\	n/a	(none)

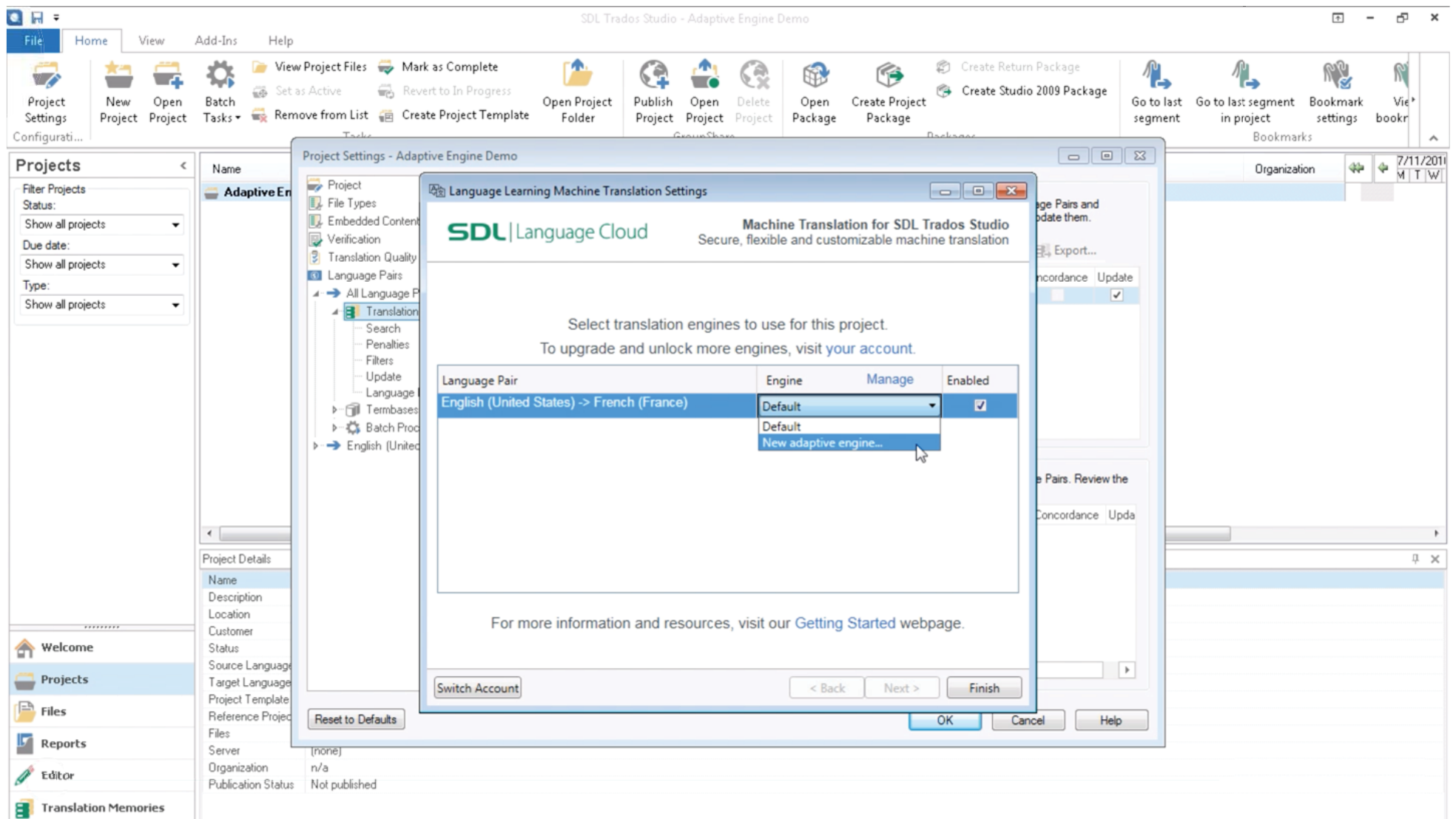
File Details

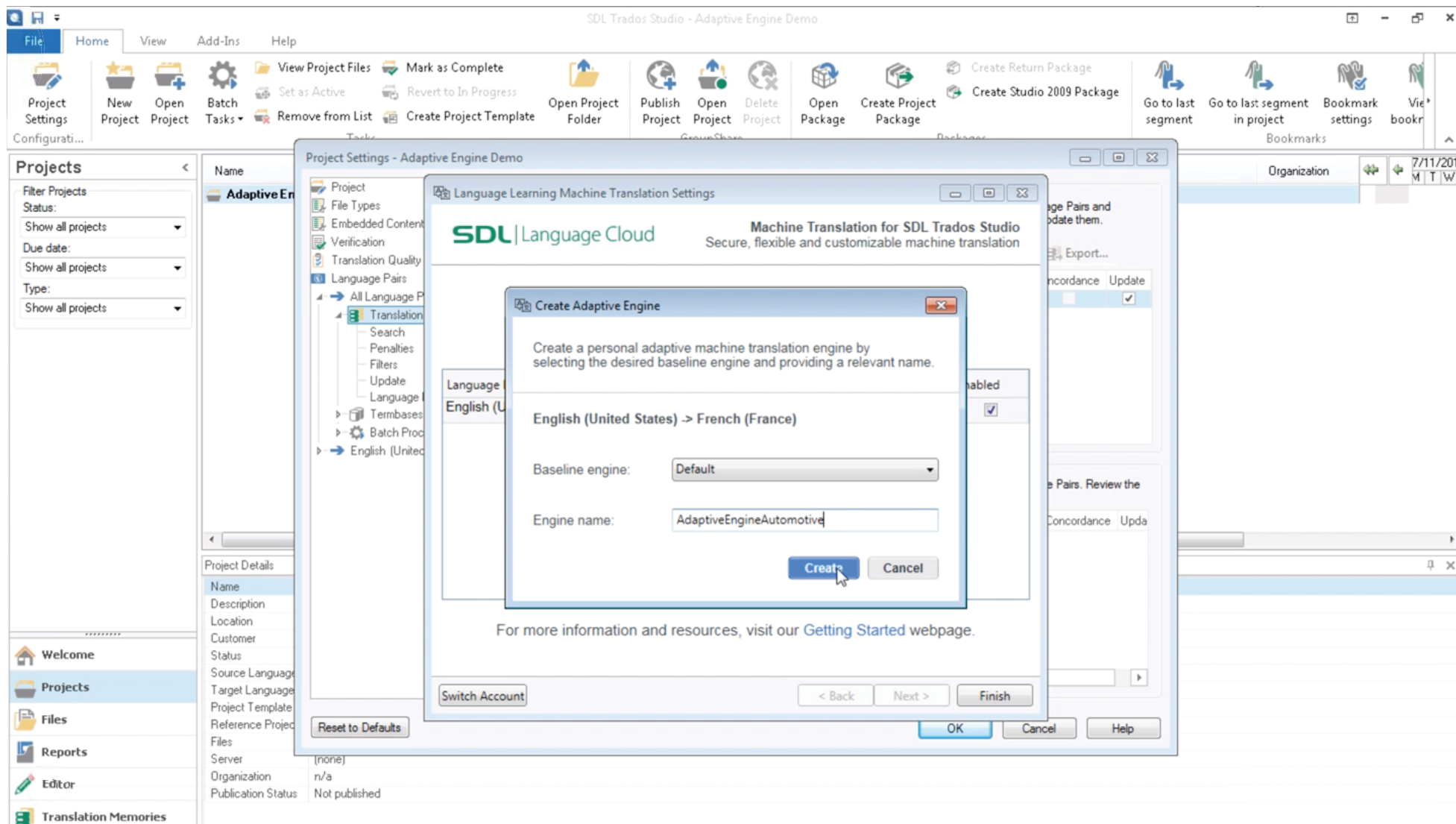
Name	Demo.eng.txt.sdxliff
Path	C:\Users\demo\Documents\Studio_2015\Projects\Adaptive Engine Demo\fr-FR\Demo.eng.txt.sdxliff
Usage	Translatable
Last Modified	7/19/2016 7:42:13 PM
Size	5 KB
Original Name	Demo.eng.txt
File Type Identifier	Plain Text v 1.0.0.0
Words	72

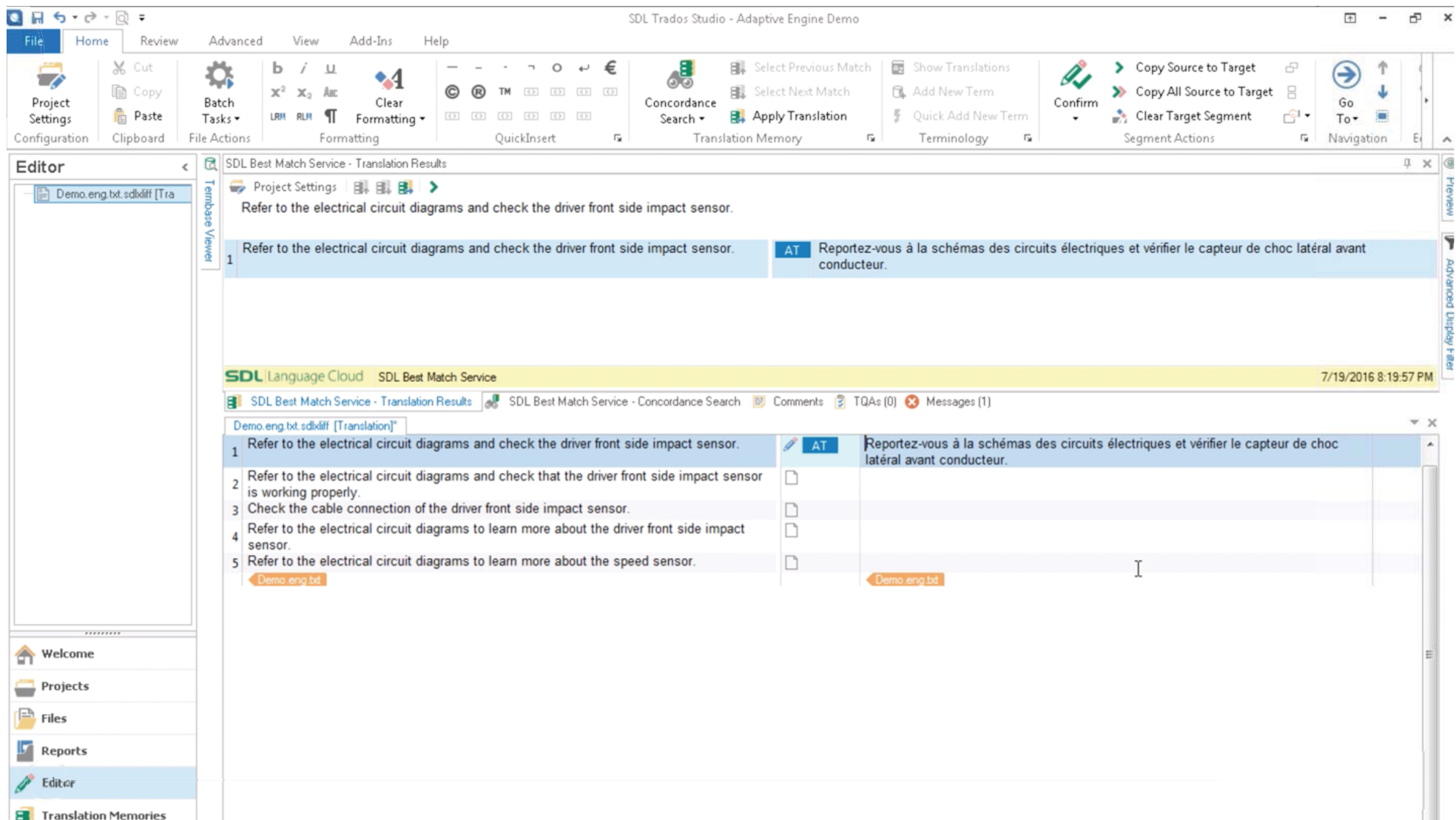


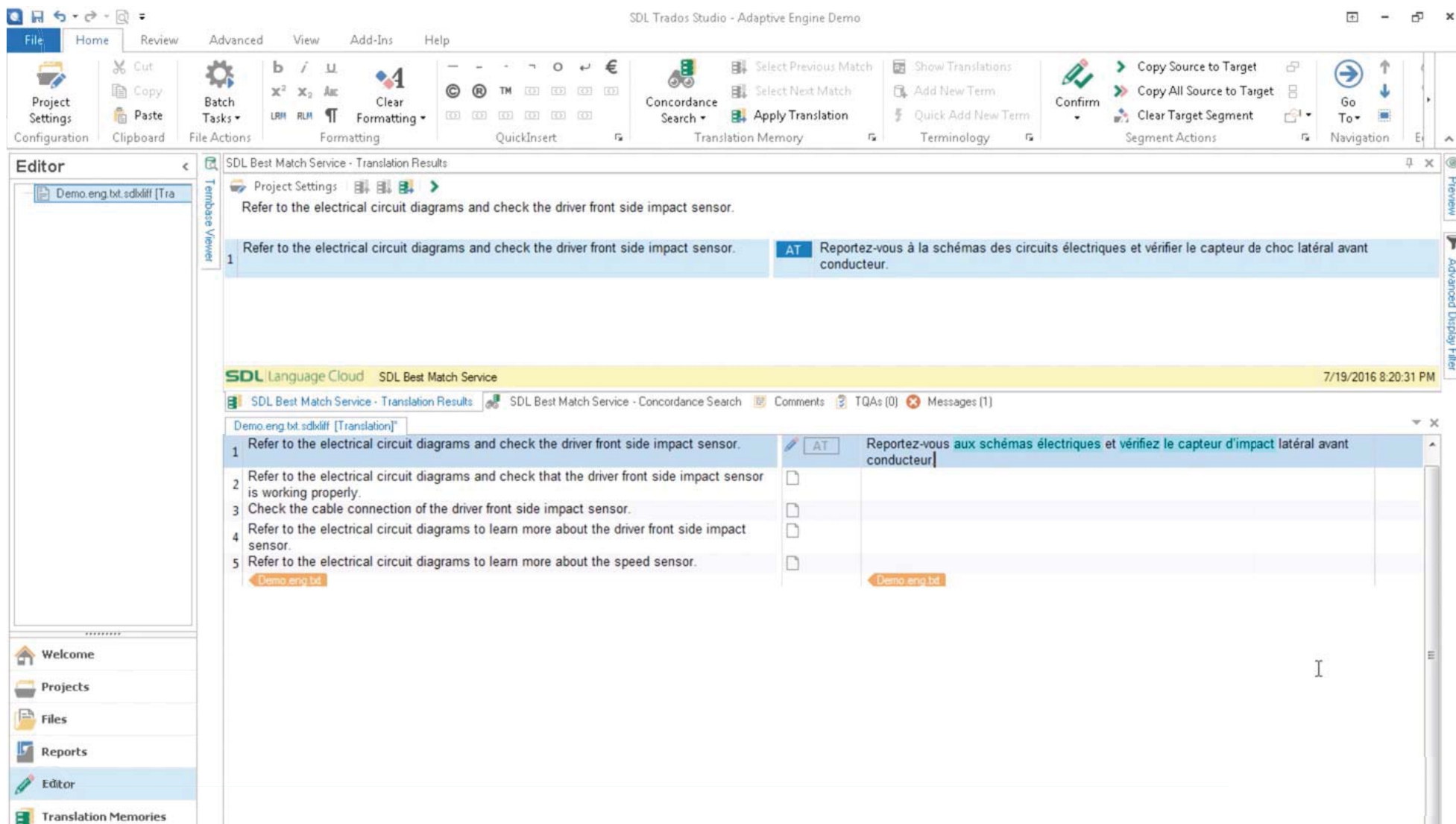


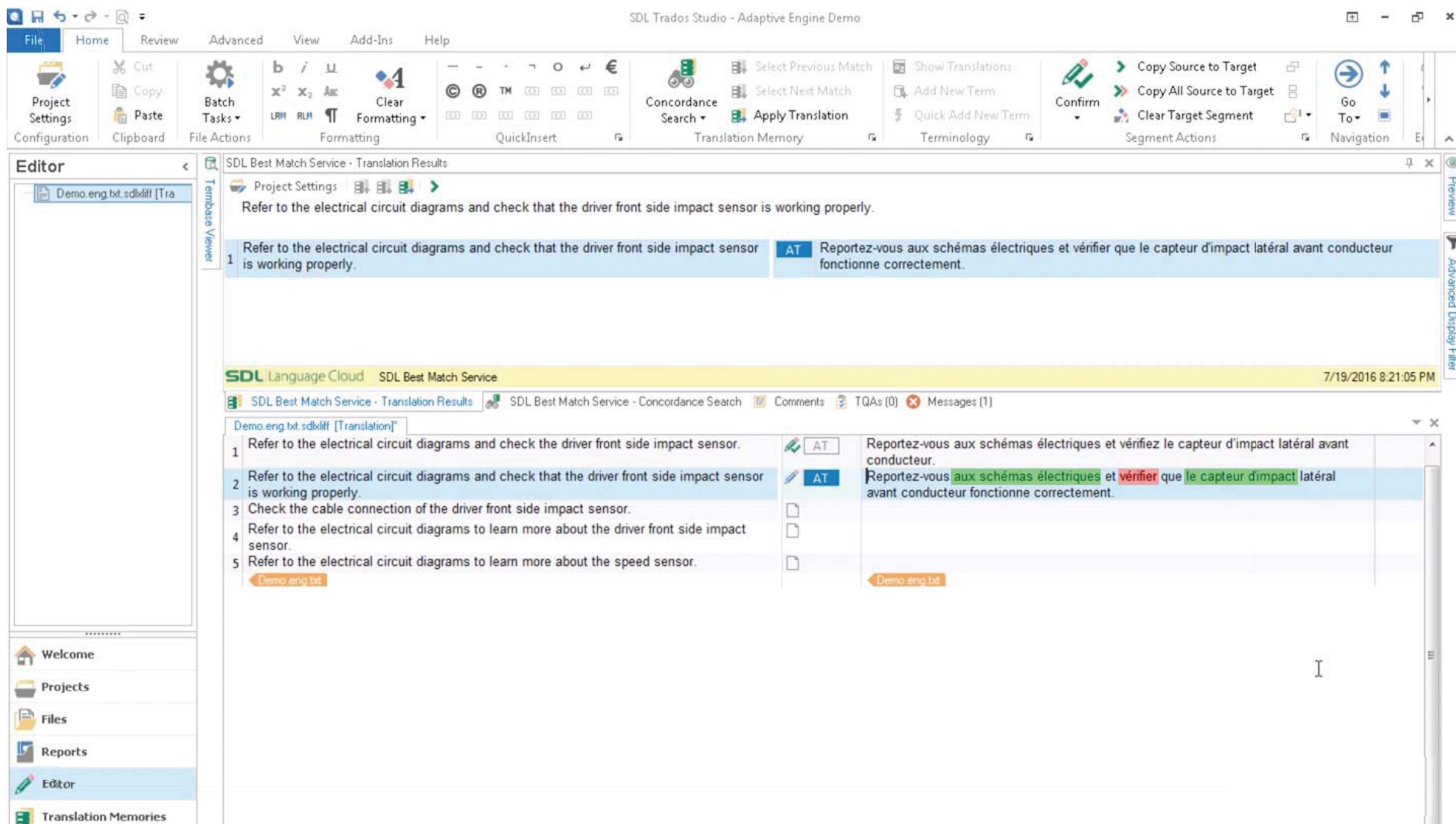


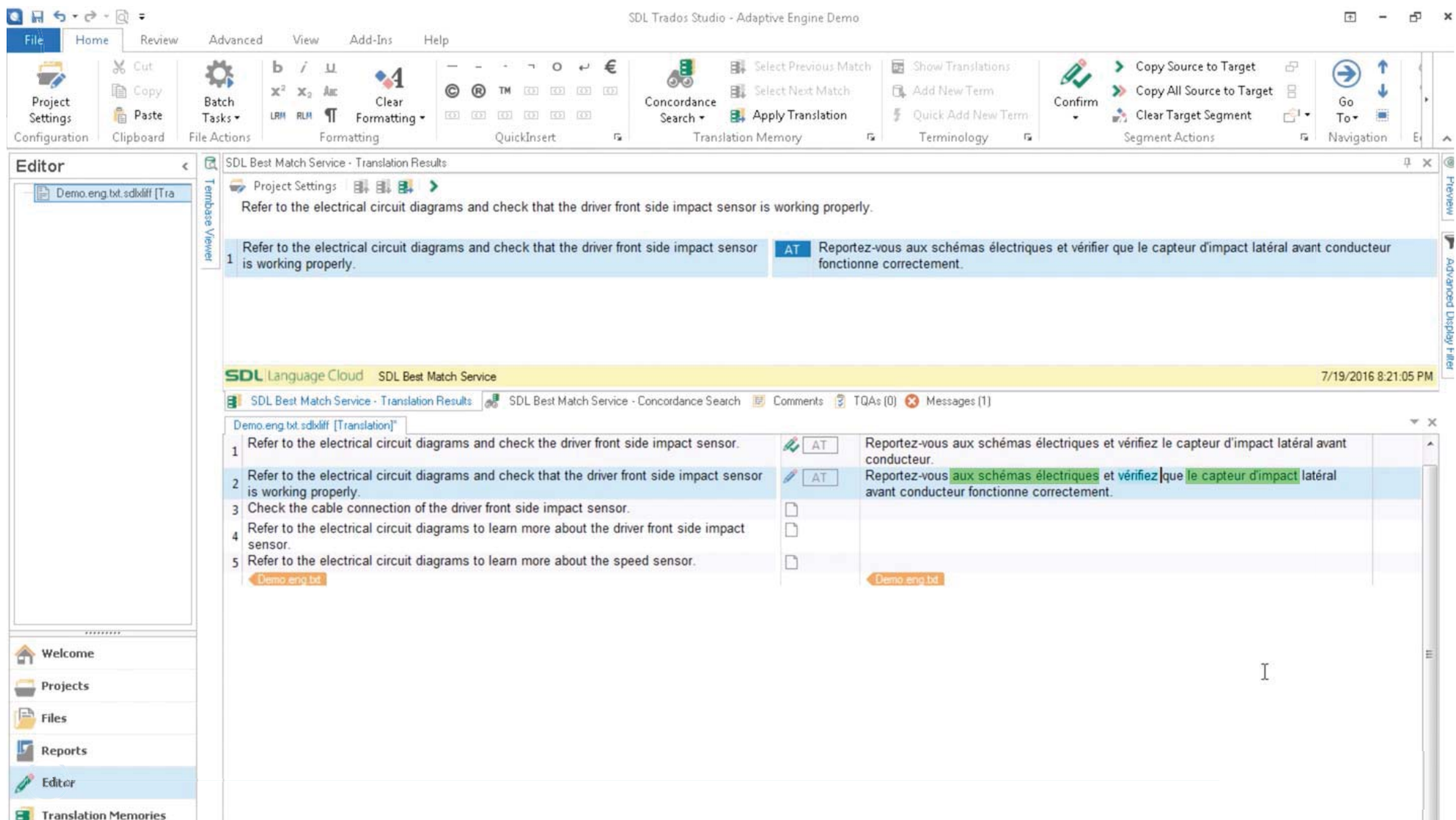


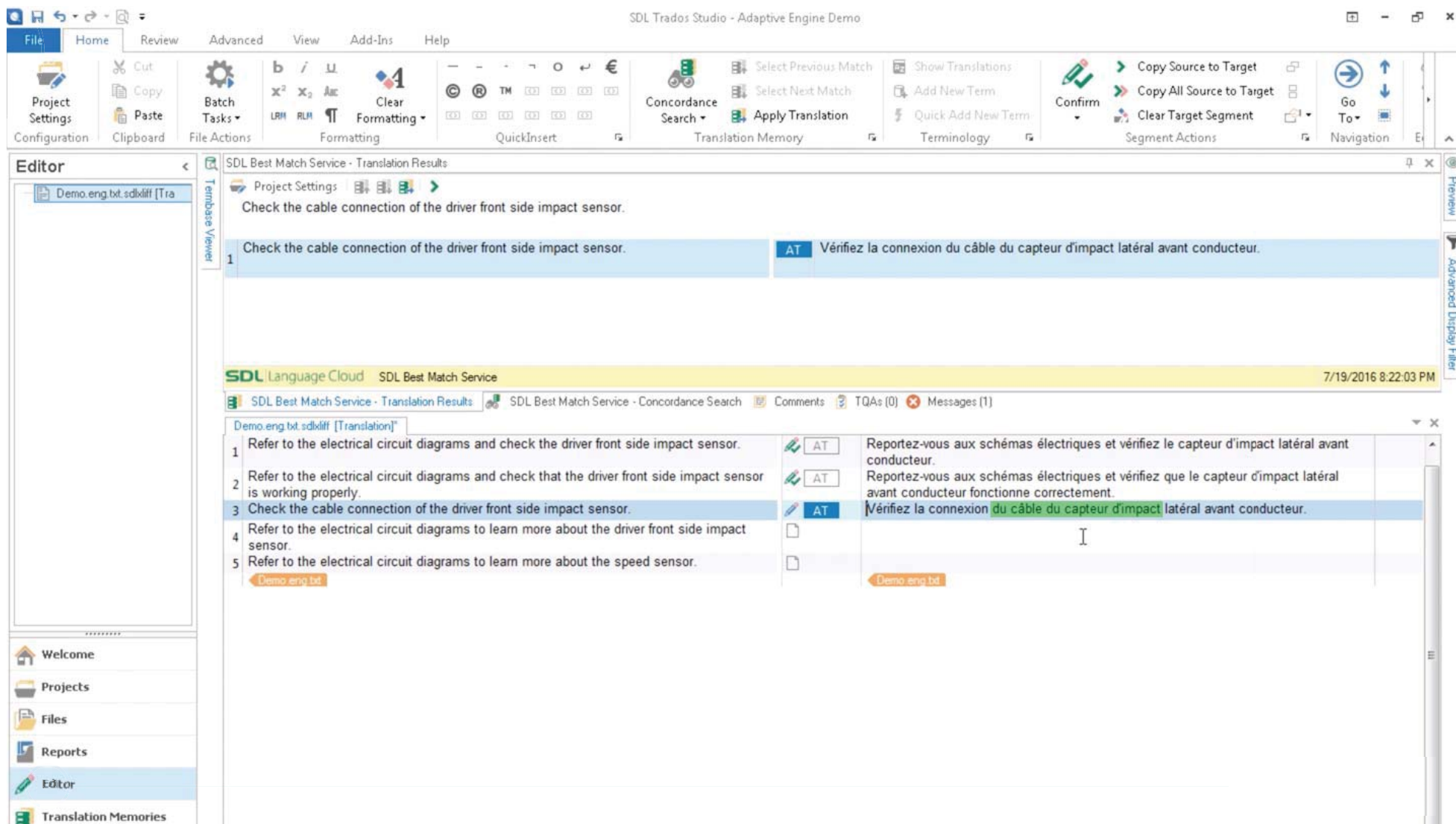


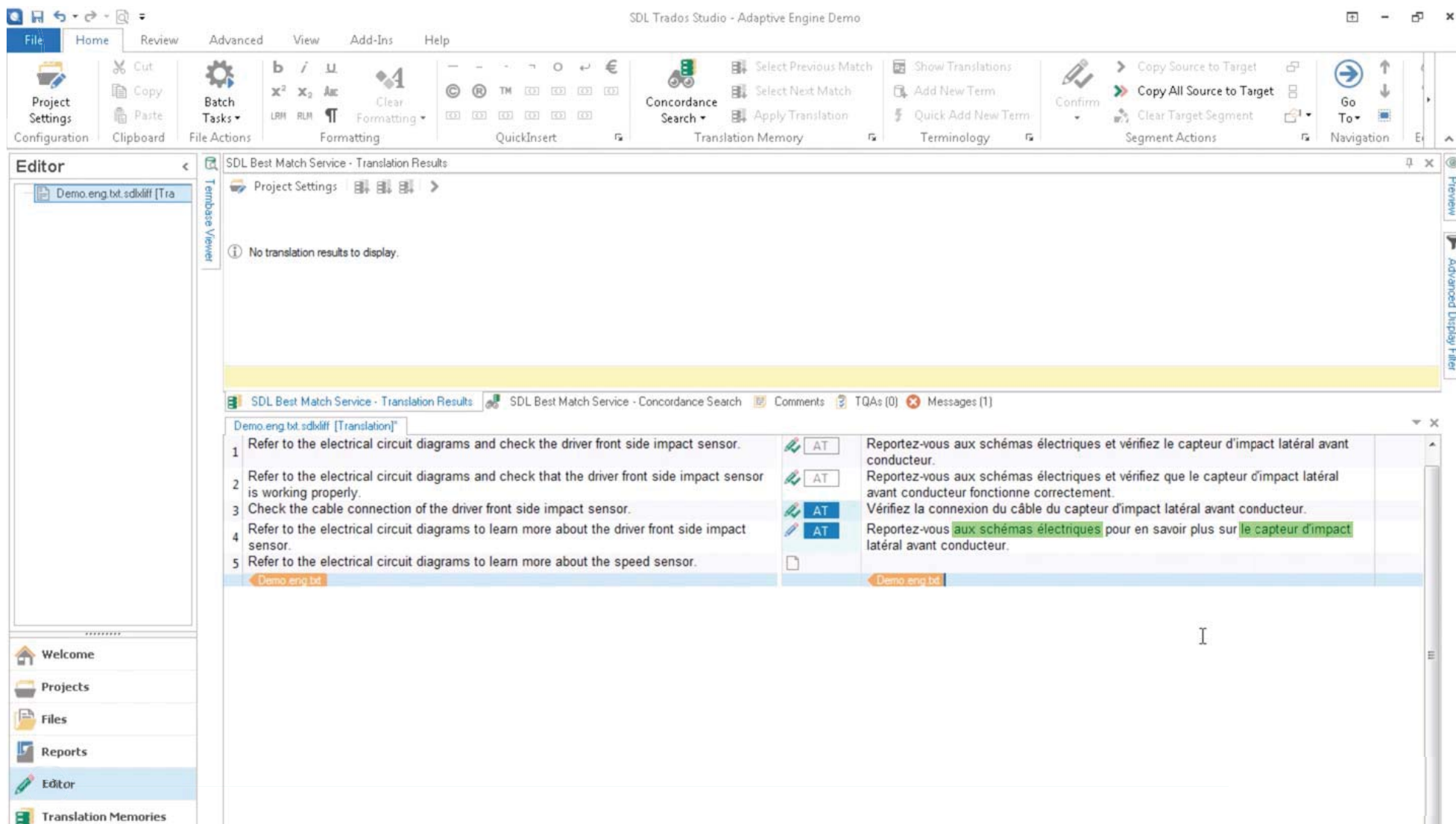


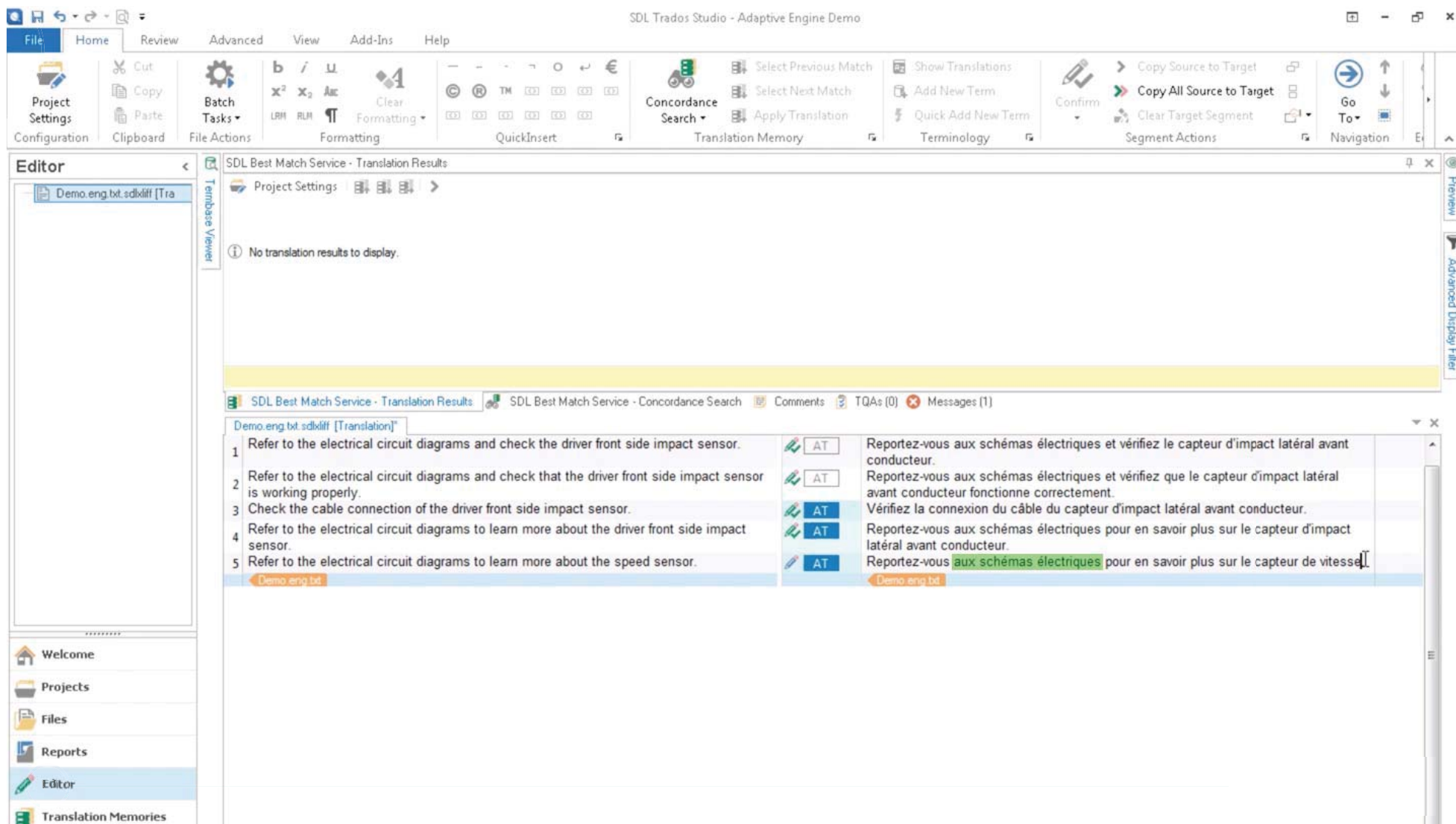












SDL Adaptive MT

- Learn seamlessly and in real time to adapt to the translation preferences of the users
- Reduce repetitive edits, allowing translation professionals to focus on high-value aspects of the translation process
- Significantly improve translation productivity and translation quality
- Give an unprecedented level of control to users

- Available in SDL Trados Studio, via SDL Language Cloud



Global Customer Experience Management

Copyright © 2008-2015 SDL plc. All rights reserved. All company names, brand names, trademarks, service marks, images and logos are the property of their respective owners.

This presentation and its content are SDL confidential unless otherwise specified, and may not be copied, used or distributed except as authorised by SDL.

Improving KantanMT Training Efficiency with `fast_align`

Dimitar Shterionov KantanLabs, Dublin, Ireland	dimitars@kantanmt.com
Jinhua Du ADAPT Centre, DCU, Dublin, Ireland	jinhua.du@adaptcentre.ie
Marc Anthony Palminteri KantanMT.com, Dublin, Ireland	marcp@kantanmt.com
Laura Casanellas KantanMT.com, Dublin, Ireland	laurac@kantanmt.com
Tony O'Dowd KantanMT.com, Dublin, Ireland	tonyod@kantanmt.com
Andy Way ADAPT Centre, DCU, Dublin, Ireland	andy.way@adaptcentre.ie

Abstract

In recent years, statistical machine translation (SMT) has been widely deployed in translators' workflow with significant improvement of productivity. However, prior to invoking an SMT system to translate an unknown text, an SMT engine needs to be built. As such, building speed of the engine is essential for the translation workflow, i.e., the sooner an engine is built, the sooner it will be exploited.

With the increase of the computational capabilities of recent technology the building time for an SMT engine has decreased substantially. For example, cloud-based SMT providers, such as KantanMT, can build high-quality, ready-to-use, custom SMT engines in less than a couple of days. To speed-up furthermore this process we look into optimizing the word alignment process that takes place during building the SMT engine. Namely, we substitute the word alignment tool used by KantanMT pipeline – `Giza++` – with a more efficient one, i.e., `fast_align`.

In this work we present the design and the implementation of the KantanMT pipeline that uses `fast_align` in place of `Giza++`. We also conduct a comparison between the two word alignment tools with industry data and report on our findings. Up to our knowledge, such extensive empirical evaluation of the two tools has not been done before.

1 Introduction

In recent years, statistical machine translation (SMT) systems have been widely deployed in translators' workflows with significant improvements in productivity. KantanMT is a cloud-based SMT platform that allows its clients to train SMT engines that are customized for their specific translation tasks by using their own data. Many factors contribute to the quality of service provided by KantanMT, e.g. the speed for training an SMT engine with KantanMT is an essential factor as it determines how fast clients can commence translating.¹

¹To date, SMT engines built using 250 million words of training data can be built with KantanMT in about 3 days.

The KantanMT platform employs a cloud-based architecture that has two main components: (i) an interface component to process, coordinate and distribute job requests and (ii) a collection of processing steps to execute build, translation or analysis jobs. The architecture of KantanMT is based on the Amazon Web Services (AWS)². For any job request the interface allocates a machine from the cloud. The machine is set-up to comply with KantanMT requirements and used afterwards to execute the specific job. For a build job, the KantanMT training pipeline is composed of 14 steps.

Crucial for the efficiency of the KantanMT training pipeline is word alignment. Word alignment is the task of identifying word-level translation relations between a source text and its translation. Naturally, to date the KantanMT pipeline has been using Giza++ (Och and Ney (2003)) – the most common word-alignment tool used by the SMT community – for word alignment. An alternative to Giza++ is `fast_align` (Dyer et al. (2013)), a simple, fast, yet effective tool to perform word alignment. Dyer et al. (2013) show that `fast_align` is about 10 times faster than IBM Model 4 (Brown et al. (1993)). Moreover, `fast_align` leads to translation performance comparable to MT engines trained using Giza++ (Dyer et al. (2013)). Accordingly, with the aim of reducing the training time of KantanMT engines, we introduced `fast_align` into the KantanMT pipeline in place of Giza++. Improved training times would lead to better quality of service as well as reduced resource allocation, an important issue for any cloud-based system.

In this work we present the integration of `fast_align` into the KantanMT training pipeline. We focus on (i) our collaborative approach to integrating the `fast_align` tool into a live production system; (ii) the improvements in training time; and (iii) a comparison of the translation quality between MT engines built with Giza++ and with `fast_align`.³

2 Training KantanMT engines with Giza++

2.1 KantanMT training pipeline

Once a building request has been received and a dedicated machine has been allocated, there are 14 processing steps that take place in order to train a KantanMT engine. Each processing step applies on the output from the preceding step and provides input for the next⁴. These steps can be divided into 5 major stages:

1. **Instance setup.** Required software is downloaded and installed; bilingual and monolingual data is retrieved and verified.
2. **Data preprocessing.** Once the data is downloaded and verified it is subjected to preprocessing, cleansing and partitioning. The bilingual data is divided into three sets – a training, a tuning and a test set – that are used for training and optimizing the engine.
3. **Building.** Three models are built during this stage: (i) a language model that captures the linguistic aspects of the target language and aims at improving MT output; (ii) a recaser model to set the correct letter casing in the MT output and (iii) a translation model used for decoding unseen text. Monolingual data (in the target language) is often used to improve the quality of the language model. The KantanMT platform employs the open-source toolkit Moses (Koehn et al. (2007)) to train the language, the translation and the recaser models.

²<https://aws.amazon.com/>

³To the best of our knowledge, such an extensive empirical evaluation of these two word-alignment approaches with industry data has not been performed to date.

⁴That is why we refer to the KantanMT architecture as a *pipeline* architecture.

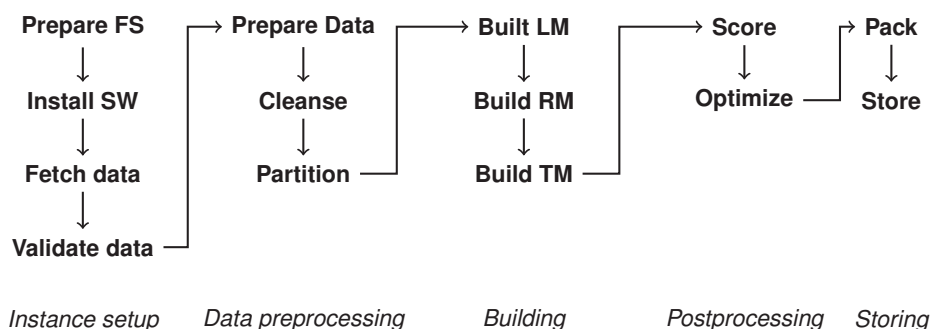


Figure 1: KantanMT training pipeline.

4. **Engine postprocessing.** After the engine is built it is scored by calculating three evaluation metrics: (i) BLEU score (Papineni et al. (2002)); (ii) F-Measure (van Rijsbergen (1979); Melamed (1995)) and (iii) Translation Error Rate (TER) (Snover et al. (2006)). If required, the engine is also optimized by using the tuning subset from the training data.
5. **Storing.** The models, configuration files and scores are packed and stored for future use.

Word alignment is invoked during building the translation model. To compute the word alignment is one of the computationally most expensive tasks in the building step. Up to date, KantanMT was using *Giza++* to perform word alignment during this step.

Figure 1 shows the original (i.e., using the *Giza++* word-alignment tool) KantanMT pipeline for training an engine.

2.2 Faster word-alignment for faster end-user delivery

Giza++ implements IBM models 1 to 5 (Brown et al. (1993)) as well as an HMM word alignment model (Och and Ney (2003)). IBM models 1 and 2 are computationally inexpensive, however, higher IBM models increase the complexity of training the models by adding additional components. For example, IBM Model 3 introduces a fertility model which can address the issue of an input word producing multiple target words or zero words⁵.

A software platform oriented to provide MT services to industry, such as KantanMT, needs to meet its clients requirements for quick service delivery. Moreover, there is a twofold gain in improving engine build time: (i) faster delivery to the end-user – the faster MT engines are trained, the sooner the client may start exploiting them; and (ii) optimized resource allocation – allocated hardware is released sooner and can be ready to use for new tasks faster.

A recent project, conducted by the ADAPT centre⁶ (Du et al. (2015)), implements an SMT pipeline where *Giza++* is substituted by *fast_align* (Dyer et al. (2013)). Motivated by the results from their experiments and aiming to achieve faster delivery to the end-user as well as improved resource allocation, we integrate the *fast_align* word alignment tool in the KantanMT platform.

3 KantanMT pipeline with *fast_align*

In order to integrate *fast_align* into the KantanMT training pipeline, an absolute prerequisite is that it needs to be 100% compatible with the already used MT pipeline. This has two require-

⁵For more information about the IBM models and their complexity as well as about HMM models we refer the interested reader to (Koehn (2010)).

⁶<http://adaptcentre.ie>

ments: (i) *no* task that KantanMT performs should see *any* degradation in performance; and (ii) the user experience should *not* be negatively impacted.

3.1 A collaborative approach for industry-standard software development

In order to ensure the aforementioned requirements, we devised a systematic development approach that teamed state-of-the-art academic knowledge and experience with industry-leading software development and quality assessment (QA). We developed our solution following an AGILE-based methodology⁷ that involved three main stages:

1. **Design.** During the design stage we first analysed the current KantanMT pipeline and identified the software requirements towards the pipeline for integrating a new word alignment tool. We then identified the software requirements of *fast_align*. Joining the industry experience and the academic know-how we formulated the system design, i.e., the functional and technical specifications of the modified KantanMT pipeline⁸. Furthermore, we reviewed and accepted any licences of additionally required software.
2. **Implementation.** Upon approval of the system design we implemented the new pipeline according to the steps defined in the technical specification document. As a first part of the quality assessment (QA) strategy we performed a series of tests to verify that each component of the pipeline, including the newly introduced ones, works as designed (i.e., alpha testing).
3. **Quality Assessment.** Our QA strategy involved alpha testing performed in parallel with the initial implementation; beta testing and life testing. While the alpha testing aimed to verify the coherence of the modified pipeline, the latter two parts aimed to ensure that the integrity of the whole system is intact. In addition, during the life testing we focused on the efficiency and user experience of the platform, i.e., empirical evaluation.

The duration of the project for integrating *fast_align* into the KantanMT training pipeline is 4 weeks plus 2 additional weeks for empirical evaluation.

3.2 System requirements

In order to incorporate *fast_align* into the KantanMT pipeline we first ensure that all software requirements of *fast_align* are met (see <http://www.cdec-decoder.org/guide/> for details). Next we need to ensure that the input data requirements are met as well. In the original KantanMT pipeline (i.e., using Giza++) we use two files that represent the parallel corpus: one for the source part of the corpus and another for the target part⁹. The *fast_align* tool uses a different input format. It requires a single file in which each line contains both the source and target parts for one sentence, separated by a triple pipe (|||)¹⁰. Example 3.1 shows such formatting for German source sentences and its English translation¹¹.

Example 3.1 *doch jetzt ist der Held gefallen . ||| but now the hero has fallen .*

We use the `paste_files.pl` script from the `cdec`¹² tool collection to join the two files in one and therefore ensure the correct formatting for *fast_align*. We invoke `paste_files.pl` right after the data cleansing (see Figure 1).

⁷<http://agilemethodology.org/>

⁸The functional specifications of the pipeline were outlined in a functional specification document; the technical specifications – in a technical specification document.

⁹During the data preparation step (see Figure 1) the corpus is encoded in UTF-8.

¹⁰http://www.cdec-decoder.org/guide/fast_align.html

¹¹http://www.cdec-decoder.org/guide/fast_align.html

¹²<http://www.cdec-decoder.org/>

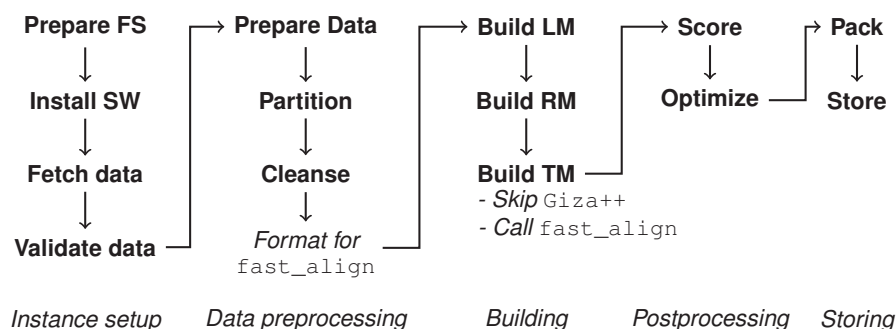


Figure 2: KantanMT training pipeline modified to invoke `fast_align`.

We run the `fast_align` tool in forward (source–target) and reversed (target–source) directions. Each direction generates asymmetric alignments, i.e., by treating either the source or target language in the parallel corpus as primary language being modeled. These two directional `fast_align` will generate slightly different alignments and need to be then symmetrized. We use the `atools` script from the `cdec` collection. The output from `atools` produces a word alignment in the widely-used $i - j$ Pharaoh format (Koehn (2003)). In this format, the pair $i - j$ indicates that the i^{th} word (zero-indexed) of the source language is aligned to the j^{th} word of the target language. Example 3.2 shows the word alignment in Pharaoh format for the sentences in Example 3.1.

Example 3.2 *0-0 1-1 2-4 3-2 4-3 5-5 6-6*

The word alignment in the Pharaoh format is compatible with the steps in the building process that follow the word alignment step.

3.3 System architecture

In order to integrate `fast_align` we modify the original KantanMT pipeline as follows:

1. Incorporate additional data processing step where we invoke `paste_file.pl` to join the two files of the parallel corpus in one according to the requirements of `fast_align` (see Section 3.2).
2. During the build of the translation model we skip `Giza++` and invoke `fast_align`.
3. Invoke `atools` to symmetrise the output of `fast_align`.

The rest of the pipeline remains the same. Our modifications are outlined in Figure 2.

The modularity of our pipelines – the original KantanMT training pipeline and the modified one – allow an easy switching between the two, i.e., between training an engine with `fast_align` and with `Giza++`. To do so we use a global variable that acts as a switch between `fast_align` and `Giza++`. This ensures (i) higher control on the pipeline; (ii) quick and easy rollback in case of unforeseen system issues.

4 Empirical evaluation

In order to determine the effectiveness of the `fast_align` word alignment tool, a series of experiments were conducted in a closed, controlled system separate from the normal workflow of KantanMT. The goal of running these experiments was to quantify the performance of the KantanMT training pipeline as well as the translation quality of engines trained using `fast_align` as compared to the original pipeline that uses `Giza++`.

<i>Language pair</i>	<i>Relative size</i>	<i>Word count</i>	<i>Unique word count</i>	<i>Engine reference</i>
English-French	Small	781 075	42 563	EN-FR-small
English-French	Large	109 379 800	1 008 696	EN-FR-large
English-German	Small	786 981	42 648	EN-DE-small
English-German	Large	138 119 563	1 084 485	EN-DE-large
English-Spanish	Small	861 557	44 375	EN-ES-small
English-Spanish	Large	154 169 102	1 119 475	EN-ES-large
English-Italian	Small	924 331	38 506	EN-IT-small
English-Italian	Large	104 196 079	914 889	EN-IT-large
English-Chinese	Small	810 134	33 281	EN-ZH-small
English-Chinese	Large	58 274 131	550 862	EN-ZH-large

Table 1: Data sets used in our comparison experiments.

4.1 Set-up

Our tests involved building 10 engines from 10 different data sets with both `Giza++` and `fast_align`. These data sets are of varying sizes and language pairs. The data is from legal and financial domain; it is part of the KantanLibrary^{TM13}. At the end of the experiments, speed, performance, accuracy, and stability of each engine were compared to their counterpart for a direct `Giza++-to-fast_align` assessment. Details about the used data sets are presented in Table 1.

We decided to use small data sets in order to test whether the alignment tool would perform better when given a small data set as compared to larger data sets. The engines built with the larger data sets were used to derive close to realistic estimates on the overall engine performance – i.e., automatic evaluation metrics and building time – when comparing the two alignment tools. Typically, a specialised engine is built on around 10 million words. The language pairs were selected based on the availability of professional linguists, translators or native speakers who can evaluate the quality of translated files.

4.2 Experiments

4.2.1 Experiment 1 – Time consumption

The objective of *Experiment 1* is to judge the speed of the KantanMT training pipeline. First we broke down the steps to building an engine from the moment the training data is prepared to when the final package is completed (see Figure 1 and Figure 2). Each step was launched manually, monitored, and timed for all 10 engines for both pipelines – with `Giza++` and with `fast_align`. We then sum the time of each individual step in order to compute the total training time. Table 2 summarises our results from timing each individual step. It shows the time gain (T^+) for building an engine with `fast_align` (T_{fa}) as compared to an engine built using `Giza++` (T_{ga}): $T^+ = \frac{T_{ga} - T_{fa}}{T_{ga}}$.

Table 2 reveals that each engine experienced a decrease in building time when using `fast_align`, with the time gain being between 48% and 73%. We ought to noted that these times do not account for the initial training data upload time, the instance setup, the data preparation, or the job clean-up which, when dealing with large engines, can add a significant amount of time to a build, i.e., can take around 40 – 60% of the total job time. In order to estimate the impact of the new word alignment tool under life conditions we also measured the time for training an engine launched from the online interface of the platform. Table 2 summarizes our results from timing the engine build including the data upload time, instance setup, etc.

¹³KantanLibraryTM is the collection of industry-standard data that KantanMT provides to their clients.

Engine	Specific tasks			Complete pipeline		
	<i>Giza++</i> (hh:mm:ss)	<i>fast_align</i> (hh:mm:ss)	Time gain	<i>Giza++</i> (hh:mm:ss)	<i>fast_align</i> (hh:mm:ss)	Time gain
EN-FR-small	00 : 09 : 23	00 : 03 : 49	59.00%	00 : 25 : 00	00 : 18 : 00	28.00%
EN-FR-large	10 : 35 : 11	04 : 02 : 14	62.00%	24 : 49 : 00	09 : 04 : 00	63.00%
EN-DE-small	00 : 10 : 06	00 : 03 : 57	61.00%	00 : 28 : 00	00 : 17 : 00	39.00%
EN-DE-large	15 : 33 : 43	04 : 13 : 57	73.00%	26 : 44 : 00	10 : 28 : 00	61.00%
EN-ES-small	00 : 10 : 21	00 : 04 : 20	58.00%	00 : 27 : 00	00 : 20 : 00	26.00%
EN-ES-large	14 : 07 : 21	04 : 54 : 12	65.00%	26 : 04 : 00	07 : 58 : 00	69.00%
EN-IT-small	00 : 11 : 03	00 : 04 : 32	59.00%	00 : 29 : 00	00 : 22 : 00	24.00%
EN-IT-large	11 : 09 : 32	05 : 46 : 41	48.00%	19 : 27 : 00	06 : 47 : 00	65.00%
EN-ZH-small	00 : 10 : 07	00 : 04 : 35	55.00%	00 : 20 : 00	00 : 16 : 00	20.00%
EN-ZH-large	10 : 08 : 16	03 : 34 : 13	65.00%	13 : 18 : 00	06 : 55 : 00	48.00%
	Average:		60.50%	Average:		44.30%

Table 2: Summary of the results from *Experiment 1*: time comparison.

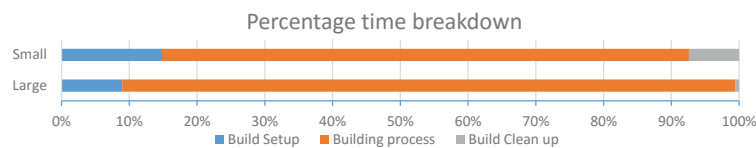


Figure 3: Percentage time taken to complete each building phase.

These results show a more accurate interpretation as to how long a build job would take for a user. For the smaller engines, the time gain from using *fast_align* has a smaller impact due to the fact that the setup and shutdown times do not change, taking a larger proportion of the total job time. In Figure 3 we show an time for initialization, training and cleanup for a random small and a large engines.

Table 2 and Figure 3 indicate that the higher the word count of the training data the more influenced the training is by the specific training steps of the pipeline, and in particular by the word alignment. That is why, for larger engines *fast_align* typically leads to higher time gain.

4.2.2 Experiment 2 – Automatic quality estimation

One of the final step in KantanMT training pipeline is to score the engines (see Figure 1). We compare the *F-Measure*, *BLEU* and *TER* scores computed at that step for each engine in order to determine the relative quality between engines built with *Giza++* and engines built with *fast_align*. We summarize our results in Table 3.

For engines that were built with *fast_align* we notice maximum score decrease of 4.5 points (in the EN-DE-large engine) and maximum score increase of 2.2 points (in the EN-ES-large engine). The average difference is 1 point in favour of engines built with *Giza++*.

Regarding, *BLEU* score, we notice a maximum decrease of 2.8 points (in the EN-DE-small engine) and a maximum increase of 3.6 points (in the EN-DE-large engine). The average

Engine	F-Measure				BLEU				TER			
	Giza++	fast_align	Absolute	Relative	Giza++	fast_align	Absolute	Relative	Giza++	fast_align	Absolute	Relative
EN-FR-small	63.0	61.8	-1.2	1.90%	62.7	60	-2.7	4.31%	51.9	53.5	1.6	-3.08%
EN-FR-large	70.4	69.5	-0.9	1.28%	61.8	62.2	0.4	-0.65%	42.7	43.5	0.8	-1.87%
EN-DE-small	57.3	58.6	1.3	-2.27%	55.6	59.2	3.6	-6.47%	58.9	55.1	-3.8	6.45%
EN-DE-large	70.8	66.3	-4.5	6.36%	66.2	63.4	-2.8	4.23%	43.7	49.6	5.9	-13.50%
EN-ES-small	69.5	67.1	-2.4	3.45%	59.2	56.9	-2.3	3.89%	44.9	48.6	3.7	-8.24%
EN-ES-large	73.7	75.9	2.2	-2.99%	60.5	63.5	3.0	-4.96%	40.2	37.2	-3	7.46%
EN-IT-small	61.9	61.0	-0.9	1.45%	54.2	53	-1.2	2.21%	52.6	54.4	1.8	-3.42%
EN-IT-large	71.0	66.6	-4.4	6.20%	60.5	61.3	0.8	-1.32%	41	44.6	3.6	-8.78%
EN-ZH-small	75.4	76.5	1.1	-1.46%	44.2	45.3	1.1	-2.49%	43.9	41.5	-2.4	5.47%
EN-ZH-large	74.7	74.4	-0.3	0.40%	53.7	52.2	-1.5	2.79%	48.7	48.8	0.1	-0.21%

Table 3: Summary of the results from *Experiment 2*: F-Measure, BLEU and TER scores.

difference is 0.16 points in favour of engines built with Giza++.

The last three columns in Table 3 refer to the TER score. It shows a maximum increase of 5.9 points (in the EN-DE-large engine) and maximum decrease of 3.8 points (in the EN-DE-small engine). On average there is an increased by 0.83 points for engines built with `fast_align`

Although we notice a general tendency of automatic quality measurements to decrease for engines built with `fast_align` (when compared to Giza++) we ought to point: (i) the fluctuation we observe in Table 3 indicates that there is no sensible change in quality and (ii) the fact that for different language pairs and data set sizes the scores alternate between `fast_align` and Giza++ indicates that under specific conditions `fast_align` is better and under other conditions Giza++ is better (the specifics of these conditions are out of the scope of this project and remain a topic for future research).

Experiment 3 In order to be get a more specific estimate of the quality of engines built with `fast_align` as compared to ones built with Giza++ we involved professional translators and native speakers (members of the KantanMT Professional Services department as well as the ADAPT centre at DCU) for human evaluation. We performed a blind comparison of 4 documents translated by engines built with `fast_align` and with Giza++ for each of the language pairs and data set sizes. The human evaluators would indicate which of the documents they considered to be better in terms of accuracy, fluency, and overall quality. The objective was to determine which engines translation would read easier when assessed by a person who is fluent in the target language.

For the small engines, all of translators independently said that the translation quality from engines built with `fast_align` is the same as from engines built with Giza++. In practice, due to the quantity of training data used by the smaller engines (being too small to translate to a high standard) the overall quality for all pairs was rather low. Therefore, we consider the results from the larger engines of practical value and focus on their analysis.

The translators had to answer two questions:

1. Which document has higher language quality?
2. Does the quality respond to their requirements as translators/native speakers?

All translators stated that both sets of documents were of a very high quality and that they cannot make a distinctive decision. With one exception: for the EN-FR-large engine, had translators disagreed as to which set of documents were better.

All translators consider both groups of translation documents to match their expectations. They also claimed that there was very little difference between documents, with mainly minor grammatical errors, e.g., two words in a long segment being in the wrong order. The results of the blind test were satisfactory to our requirements for the project.

5 Conclusions

In this work we substituted the word alignment tool used in the KantanMT pipeline – `Giza++` – with the more efficient and yet effective `fast_align`. The latter has already been successfully incorporated in other industrial products and has shown promising results. In our design and implementation strategy we combined industry established software development practices with academic know-how to effectively modify a large-scale online platform such as KantanMT with **no** downtime or decay in KantanMT services.

To assess the impact of the new word alignment method into the efficiency of the KantanMT pipeline as well as the quality of the built engines we performed a series of tests with industry based data. Our results show an average speed-up of 60% and comparable quality to engines built with `Giza++`.

Our experiments also show that for languages that differ substantially in the word order (such as English and German) `fast_align` may lead to a slight decrease in quality (according to automatic measures). In the future, we aim to carry out more investigation on word alignment results from `fast_align` and examine possible solutions of improving SMT performance.

References

- Brown, P. F., Della-Pietra, S. A., Della-Pietra, V. J., and Mercer, R. L. (1993). The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311.
- Du, J., Moorkens, J., Srivastava, A., Lauer, M., Way, A., Maldonado, A., and Lewis, D. (2015). D4.3: Translation project – level evaluation distribution: Public federated active linguistic data curation (falcon).
- Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia, USA.
- Koehn, P. (2003). Pharaoh: a beam search decoder for phrase-based statistical machine translation models: User manual and description for version 1.2.
- Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1st edition.
- Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics: Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.
- Melamed, I. D. (1995). Automatic evaluation and uniform filter cascades for inducing n-best translation lexicons. In *Proceedings of the third Workshop on Very Large Corpora*, pages 184–198, Cambridge, Massachusetts, USA.
- Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA.
- Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231, Cambridge, Massachusetts, USA.
- van Rijsbergen, C. J. (1979). *Information Retrieval*. Butterworth.

Speech translation user experience in practice

Building a speech translation feature for Skype

Chris Wendt
Will Lewis
Tanvi Surti
Microsoft Research, Redmond, Washington, USA

christw@microsoft.com
wilewis@microsoft.com
tsurti@microsoft.com

Abstract

Human conversations, especially between humans who are untrained in interpreted conversations, do not fit an utterance-translation-utterance-translation temporal sequence. Speech, especially speech by a human to another human, is burdened with many artefacts that are undesirable in translation. As implementers, we have a number of methods available to us, to alleviate these artefacts and provide a text to the translation system that looks more like written text than spoken text. We'll discuss the state and the limits of that approach, and the challenges on the journey of providing a representation of the speaker's intent.

1. Introduction

Skype is in the business of breaking down barriers to long-distance communication, crossing geographical distance. Automatic translation extends the concept to language, breaking down those barriers faster and cheaper, opening up more and more scenarios where communication across language boundaries becomes possible.

In the Star Trek TV series, the Universal Translator appears to be such an ordinary device that it almost seems surprising that translation of speech is actually new. Since its inception in 2003, Skype has been a primary vehicle for long-distance communication that easily bridges continents. Automatic translation of speech in a human-to-human fashion using a conversational domain is now extending the concept of removing barriers to communication to language.

2. Three Factors Coming Together

Why introduce wide domain conversational speech translation now? Technology for recognizing speech has been around since the 1960s. Machine translation exists since the late 1950s. Putting them both together produced a multiplication of each other's mistakes, resulting in useless garbage. Speech translation is difficult because a single misrecognized word in the original language easily makes the target translation completely unusable. Looking at the output of automatic speech recognition (ASR) in the same language as spoken, mistakes don't appear so bad: the recognizer is likely to produce a homophone of the intended word or phrase, and, with some imagination, a human can revert the misrecognized word back to the similar-sounding original, and make sense of the whole. The translation system, as designed for text translation, cannot recover from homophones, because after translation of the text, the words in the target language are not phonetically related to the correct alternative anymore. Once translated, the homophones don't sound alike at all anymore – the translation becomes incomprehensible.

Like speech recognition, deep neural networks were originally proposed in the 1950s as well. But, due to the lack of processing power, they didn't show their revolutionary ability until recently. Over the last 10 years, deep neural networks have been successfully applied in speech recognition, in OCR and in image recognition. In speech recognition the introduction of deep neural networks gave us a 33% quality boost, reducing the word error rate from 30% to 20%, on an unbound domain English test set, with varying recording quality. A relative 33% reduction of the error rate is a truly drastic, enormous improvement [5], [6].

Automatic translation of text already had a major breakthrough in terms of quality and language coverage, within the last 5-10 years, due to the advances brought to the field by statistical machine translation (MT). Statistical MT gave us wide domain coverage and easy trainability for any language pair that has enough bilingual training material, and is producing good results in translating professionally or casually authored text in the major language pairs. Deep neural networks are on the verge of bringing another big boost in text translation, most visible in language pairs that have significantly different language structure, for instance when translating between English and Japanese.

Skype is the primary network for people wanting to connect across the earth's borders, oceans and continents, using their voice and video. This made Skype the perfect medium to help these people communicate across language barriers as well.

Three factors have now come together to create the opportunity for developing a speech translation system for conversational speech in an unbound domain: 33% quality improvement from deep neural networks for speech, statistical MT being reasonably good, and the border-crossing network of Skype.

3. Human-to-human Conversation

Our mobile devices and gaming consoles, our cars, our banks, and Siri, Cortana, Google Now, act on voice commands quite well, even complex commands and queries involving names of people or locations. With digital assistants, users are talking to a machine with a clearly limited scope at any given point in time, not another person. The systems are trained and optimized to perform well in a restricted scenario. A conversation between two humans doesn't follow any rules or defined vocabulary. It is unbound in terms of what the conversations participants say and how they say it. The domain of what the participants are talking about is unrestricted.

This of course poses a challenge. The larger the vocabulary, the more inherent acoustic ambiguities are present in the vocabulary, as close and not-so-close homophones, which drastically increase the set of alternatives the system needs to consider. When you say the word "wreck", for instance, the acoustic signal doesn't know word boundaries, so "wreck" can be the first part of "recognize" or the last or middle part of some other words. The acoustic model and the language models have consumed many thousands of hours of audio and transcripts from real people talking to each other, in order to build their knowledge of context, and the suitable alternatives within the context.

Deep neural networks in speech recognition, due to their depth, can consume a wide variety of accented speech, and different ways to say a word, without degrading the experience for other accents. They are better suited to remember the context as seen during training, and can reproduce the context at runtime. As a result, Skype doesn't need different settings for English spoken on different continents – it uses the same large models that are trained with British, Canadian, American, Australian and New Zealand speakers, together. The same is true for French and Spanish. The resulting system may still do better with one accent or another accent,

but that's only due to the amount and acoustic variations within the training material. The training cycle can add this material as it becomes available, without degrading the experience for anyone.

4. Intent vs. Received Audio

When humans speak to each other, they are typically not very cognizant of grammar, fluency, clarity, casing and punctuation, far less so than when putting down the written word.

What he thought he said:

Yeah. I guess it was worth it.

What he actually said:

Yeah, but um, but it was you know, it was, I guess, it was worth it.

When the MT system translates what he really said, it doesn't get better or easier to understand.

Microsoft uses a component called TrueText [1] to remove disfluencies and other undesired artefacts of human speech: Grunts and coughs, ahs and ems, false starts and repeats. TrueText is a translation system by itself – it translates spoken language to a series of tokens that look more like written language within the same language. Written language is what the actual translation system is good at. TrueText is trained on pairs of exact transcripts and fluent transcripts of the same spoken utterance. A typical human transcript is in fact a fluent, clean transcript. The exact transcript can be synthesized from the original audio and the fluent transcript, thereby creating the parallel training data for the TrueText translation system.

5. Design Considerations

5.1. Interaction with Humans

Does an interpreter have a persona? Should the interpreter have a persona, or rather be transparent, in the background, not an actor in the conversation? We can observe that a novice at interpreted conversations tends to perceive the interpreter as an actor, a person who can be addressed directly, maybe even as a cultural consultant. Experienced participants in interpreted conversations will remain focused on the foreign-language conversation partner, and let the interpreter be the medium, the conveyor of information, visible and audible, but not interfering in the conversation. In settings like the European Parliament, this understanding of the hidden interpreter is helped by the fact that the interpreter is in fact far away and invisible.

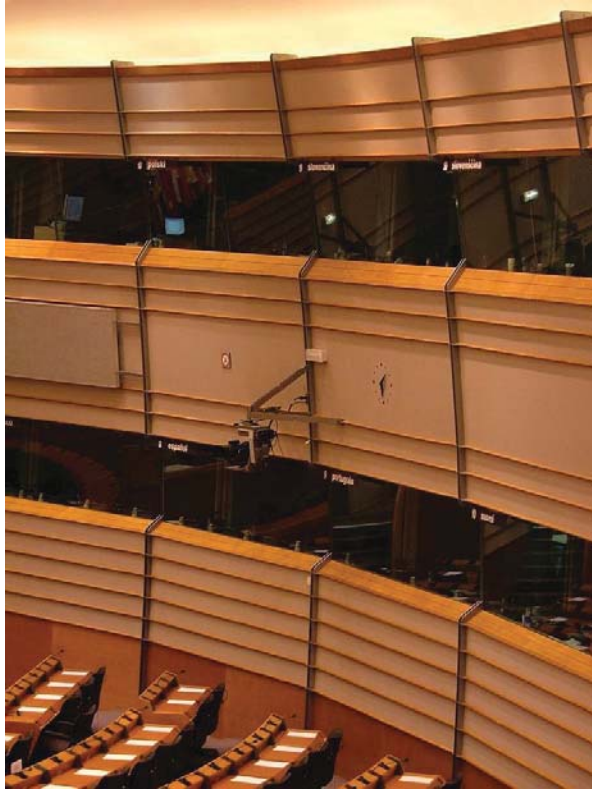


Figure 1. At the European Parliament, the interpreters are mostly invisible

With Skype Translator, Microsoft had a choice to model the translation function with an acting persona, or transparent like the EP interpreter. The acting persona implementation has the psychological advantage that there is someone to blame for misinterpretation and misunderstanding. “No, I didn’t say that, it is the interpreter’s fault. “. If the system remains persona-less, the utterance is more directly attributable to the person who spoke the original, even if it was the recognition or translation that introduced the mistake. The Skype Translator team decided to go persona-less, and just put the conversations between the humans in the foreground. It is not only easier to implement in a functional fashion, it is also more conducive to productive conversations.



Figure 2. Chinese President Xi Jinping visits Microsoft, with Microsoft CEO Satya Nadella and interpreter

Maybe there is a time when the interpreter can get a persona and interact with the participants, in a fun and entertaining way.

5.2. Handling the Audio Stream

In the physical world, an interpreted conversation goes something like this:

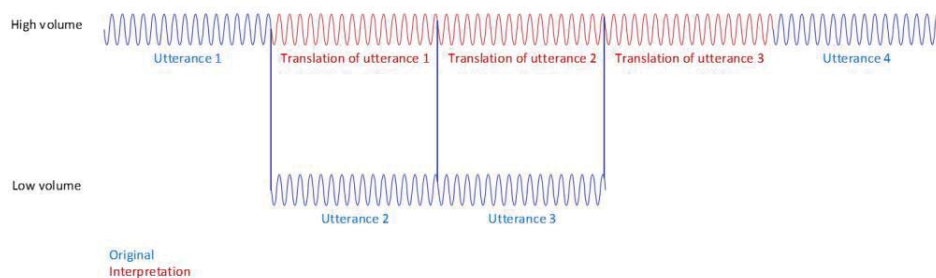
- A speaks, and then A stops speaking.
- Interpreter speaks what A just said, in B's language, and then stops speaking.
- B speaks, and then B stops speaking.
- Interpreter speaks what B just said, in A's language, and then stops speaking.

That is a fairly long and drawn-out process, in fact twice as long as a direct, same-language conversation. It is complicated by the fact that B won't know when A's interpretation is finished. B doesn't know what A said, so B won't know the logical end of the utterance, there would need to be some kind of signal indicating the end of the utterance, so that B knows when to start speaking. Humans can use gestures and facial expressions to indicate when they are done.

Skype can do away with all this, making use of the fact that Skype controls what each participant hears. After all, they are on Skype, and not in the same room. Skype uses a broadcasting technique called "ducking", which ducks the volume of the original voice under the voice of the interpreter. This is commonly used in broadcasts of interviews with foreign dignitaries: First you hear a few words of the foreign dignitary speaking in the foreign language at full volume, then the interpreter kicks in at full volume, and the volume of the dignitary's voice becomes low enough to duck under the interpreter's voice. Skype does the same: You hear your

partner at full volume first, then the interpretation kicks in at full volume. Your partner can continue to speak and you'll hear her voice ducked under the interpreter's voice, as long as the interpretation goes on, and then it comes back at full volume.

Ducking: Varying the volume of the original audio



Speaker hears the other person ducked during interpretation.
Speaker hears his own translation always at low volume.

Figure 3 Ducking: Varying the volume of the original audio

You will hear the interpretation of what you yourself said, ducked under everything, always faint. It shows that the system is working, explains why your partner didn't laugh yet about your super funny joke – she didn't hear it yet. Skype could have chosen a different indication that “interpretation of what you just said is in progress”, but the team figured the actual ducked audio of your interpretation is the best indicator. Speaking over your own interpretation is entirely possible and liberating: you can tell the whole long story of your weekend adventures, without having to pause.

In addition to the translated audio, Skype also shows you a transcript of what you said, and the translation to your language of what your partner said. In the usability tests [3], about half of the participants preferred to turn the translated audio off and just read the transcripts. Indeed, it makes for a more fluent conversation – most of us are faster reading than we are listening. However, the translated audio works better on small-factor devices and in situations where you don't want to, or can't look at the screen the whole time.

6. Observations in Practice

6.1. Named Entities

While Skype Translator is designed with a large vocabulary, allowing you to speak about anything and everything, the vocabulary does have limits, which is most painful in the area of named entities. Names of places, people or businesses fail to be recognized correctly, in the majority of cases. With a few exceptions: at the start of the call, Skype builds a mini-language model from your personal Skype address book, and adds it to the vocabulary, in the same way your phone does, to allow you to call the people in your address book by voice command. A

difficult problem to overcome is that these added names are synthesized in the pronunciation of your language. If you are an English speaker, you'll have to pronounce all names in an anglicized manner, for the people's names to be recognized. That's probably not how you would properly pronounce foreign names in your address book, which most likely happen to be the ones you will be mentioning during a translated call. Adding the entire set of the world's first and last names, all place names and all business names to the vocabulary would degrade the accuracy so much, the system would become unusable if that happened. For the foreseeable future the system will need to employ selective methods of dynamically adding just the right amount of phrases to the vocabulary, for the given situation, by taking intent and location into account.

6.2. Accented Speech

Skype Translator's training material mostly comes from native speakers of their language. The audio characteristics of second language speakers significantly differ, and second language speakers will find a noticeable drop in recognition quality, compared to native speakers.

6.3. Audio Quality

Speech recognition in a wide domain requires a microphone that sits close to your mouth, in order to pick up the nuances of your voice and to reduce the impact of background noise. Such a microphone comes with a headset that also gives you better audio for your own listening. Skype users are used to having Skype conversations simply using the laptop's or tablet's built-in microphone and speakers.

6.4. Pronunciation

Practicing better pronunciation helps a lot. Some users expressed that using Skype translator helps them to become better native speakers – forcing them to exercise more careful pronunciation. It also helps to know what you want to say before you say it – a feat that the non-politicians among us sometimes have trouble with.

7. Conclusion

Are we done with automatic interpretation? Not by far. There is lots of room for growth: Recognition accuracy and spoken language artefact removal, while maintaining and growing the very wide domain coverage, will get better drastically.

Translation itself benefits from the introduction of neural networks, which have introduced a qualitative jump in other areas of machine learning. Neural networks have the ability to remember any number of factors that influence a particular translation, much better than today's statistical systems do. That helps translating between languages of different language structure as well as lifting the quality of the lower-resourced languages, making better use of smaller amounts of parallel material.

User experience benefits from adjustment to the usage scenario: While the experience in Skype benefits the use in long-distance video calls, it will evolve for in-person meetings or group calls.

Microsoft makes the API that powers translation in Skype [4] available for your own communication applications as well. You can use it to build close-to-real-time speech recognition and translation solutions for your own scenarios.

References

This paper cites large sections from [2] in unchanged form.

- [1] Lee Schwartz, Dilek Hakkani-Tür, Gokhan Tur, Hany Hassan Awadalla, “Segmentation and Disfluency Removal for Conversational Speech Translation” Proceedings of Interspeech, ISCA - International Speech Communication Association, September 1, 2014.
- [2] Chris Wendt, “Behind Skype’s machine interpreting”, Multilingual Magazine, September 2016, https://multilingual.com/all-articles/?art_id=2373
- [3] Kotaro Hara, Shamsi Iqbal, “Effect of Machine Translation in Interlingual Conversation: Lessons from a Formative Study”, ACM Conference on Human Factors for Computing Systems (CHI), April 1, 2015.
- [4] Microsoft Corporation, “Microsoft Translator”, <http://www.microsoft.com/translator>.
- [5] Li Deng, Jinyu Li, Jui-Ting Huang, Kaisheng Yao, Dong Yu, Frank Seide, Michael Seltzer, Geoff Zweig, Xiaodong He, Jason Williams, Yifan Gong, Alex Acero, Mike Seltzer, “Recent Advances in Deep Learning for Speech Research at Microsoft”, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), May 1, 2013.
- [6] Dong Yu, Frank Seide, Gang Li, “Conversational Speech Transcription Using Context-Dependent Deep Neural Networks”, ICML 2012, June 1, 2012.



Booking.com

Maxim Khalilov

Evaluation of machine translation quality in e-commerce environment

AMTA 2016
Commercial MT Users and Translators

October 31, 2016
Austin, TX

Booking.com story.

The world's #1 website for booking hotels and other accommodations.

- Founded in 1996 in Amsterdam
- Part of the Priceline Group (NASDAQ: PCLN) since 2005
- 1,000,000+ properties in more than 200 countries including 490,000+ vacation rental properties
- Over 1,100,000 room nights every 24 hours
- Accommodation available in: 220+ countries and territories
- Number of unique destinations worldwide: 94,000+
- 42 languages
- 184 offices worldwide in 50+ countries
- More than 13,000 employees



Booking.com

Outline.

- Why? MT for e-commerce.
- MT at Booking.com: timeline and approach
- Why MT can be dangerous and how can we deal with it?
- Evaluation of MT quality: methods
- Quality assurance system
- Benefits
- Q&A



Market strategy

Growth is driven by the localization of global market strategies



Technology

Significant impact on the e-commerce business in general by enabling global customers reach the product description content across language borders

Industry

Travel industry was one of the first to come to the Web



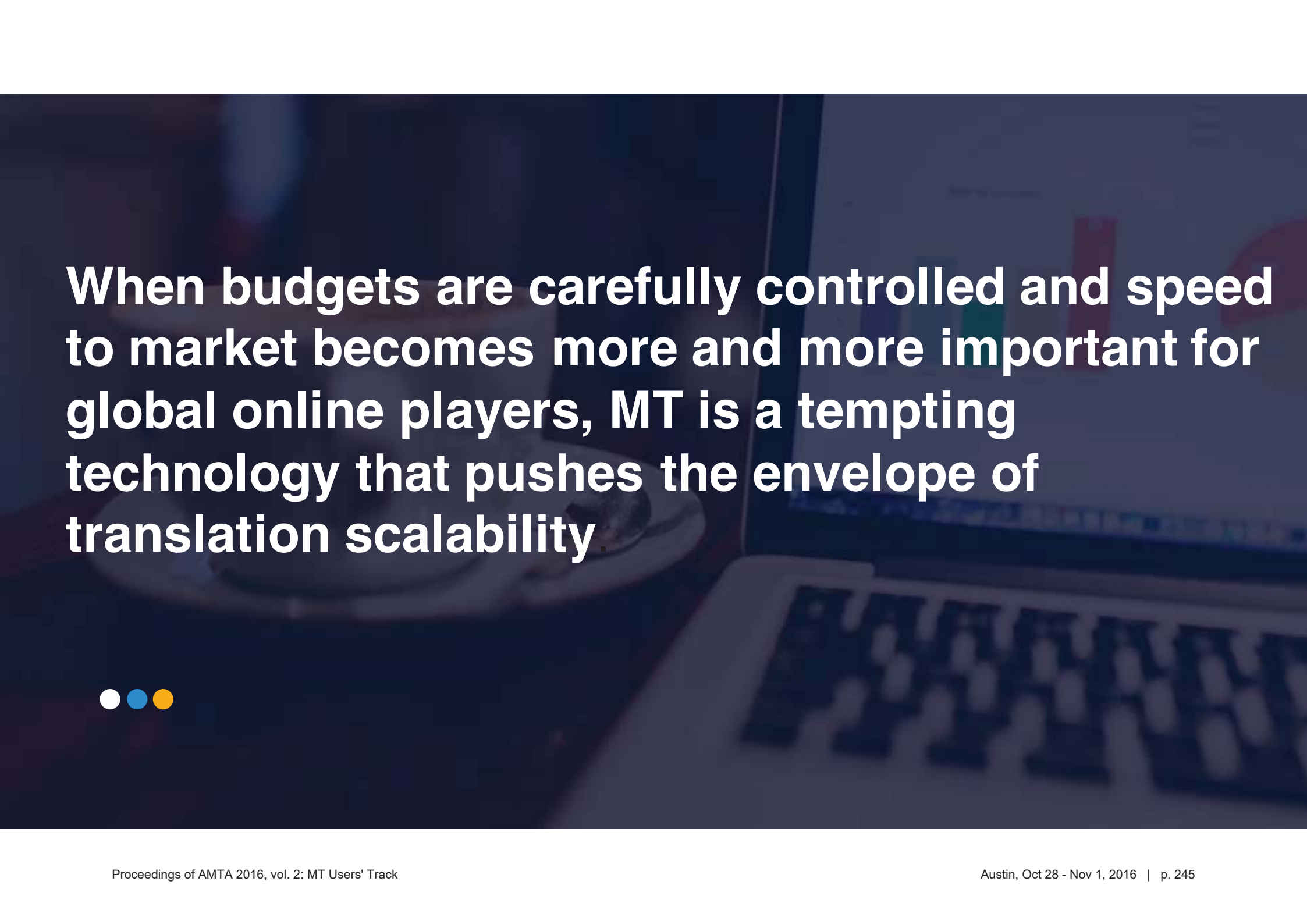
Languages

Demand-driven language selection



Why Machine Translation?





When budgets are carefully controlled and speed to market becomes more and more important for global online players, MT is a tempting technology that pushes the envelope of translation scalability.



Four business cases. (for Booking.com)

1. Property descriptions

MT helps to scale property (hotels, apartments, hostels, etc.) descriptions into 42 languages.

2. Translation support

In-house MT supports translation department of the company to increase the productivity of translation process.

3. Customer experience applications

Our guests and partners consumer and produce content in different languages in various forms from USGs to CS tickets.

4. Other cases

Various other MT use cases, including Big Data applications and user feedback processing.

Booking.com

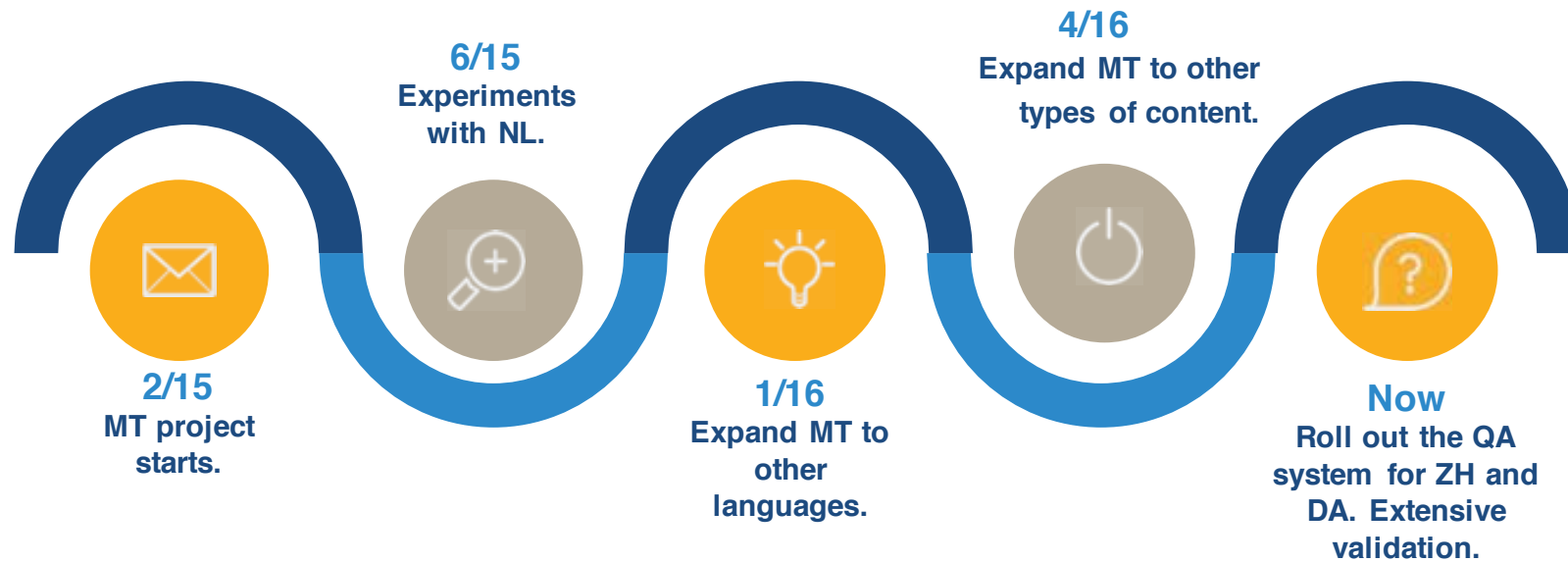
Property
descriptions.

Translation
support.

Customer
experience
applications.

Other cases.

MT timeline.



Booking.com



Why MT can be dangerous?

The imperfection of MT might mislead users, have legal consequences for the company or damage brand's reputation and customer's confidence of translated content.

Booking.com

Challenges

- A lack of objective and comprehensive methodology of MT quality evaluation that would be flexible enough to give reliable results for different types of content.
- An absence of the clear link between a fully automatic metric or a set of metrics with business impact of translation quality.



How can we control MT quality in e-commerce environment?



Human evaluation.



Error analysis

Useful for regular checks of translation quality from the linguistic perspective



Usability assessment
Adequacy/Fluency scoring

Rough assessment of the MT-ed content in terms of its publishability



User feedback
User behavior analysis

Both methods are mostly used in the customer-facing scenarios



Productivity measurement

Post-editing MT scenario

Booking.com

Automatic methods.



Conventional
automatic metrics



Applicable to make sure
there are no new bugs
introduced as the result
of the MT engine
retraining.



Number of business-sensitive
errors



Methods that link MT
quality with potential
threats for the business

Booking.com



Problem.

- **Automatic metrics for MT quality evaluation, such as BLEU, METEOR and GTM are not capable of reflecting the usability of MT-ed content**
- **They also can not distinguishing harmful errors from those that do not have critical impact on the overall translation understandability and adequacy.**
- **Human quality assurance is time consuming and non-cost effective.**
- **Human evaluation is subjective.**

Booking.com

Quality assurance system



How?



Solution.

Quality heuristics: enhance the quality evaluation model to ensure the MT system is not making sensitive errors like offering free facilities that aren't actually free, or mistranslating distances.

Implementation.

Sensitive words.

-
-
-
-

Manually created multilingual list of sensitive words, expressions and values that can prejudice the accuracy of translation

Matching module.

-
-
-
-
-
-
-

Compares factual and numerical discrepancies between source copy and MT output

Reaction.

-
-
-
-
-
-

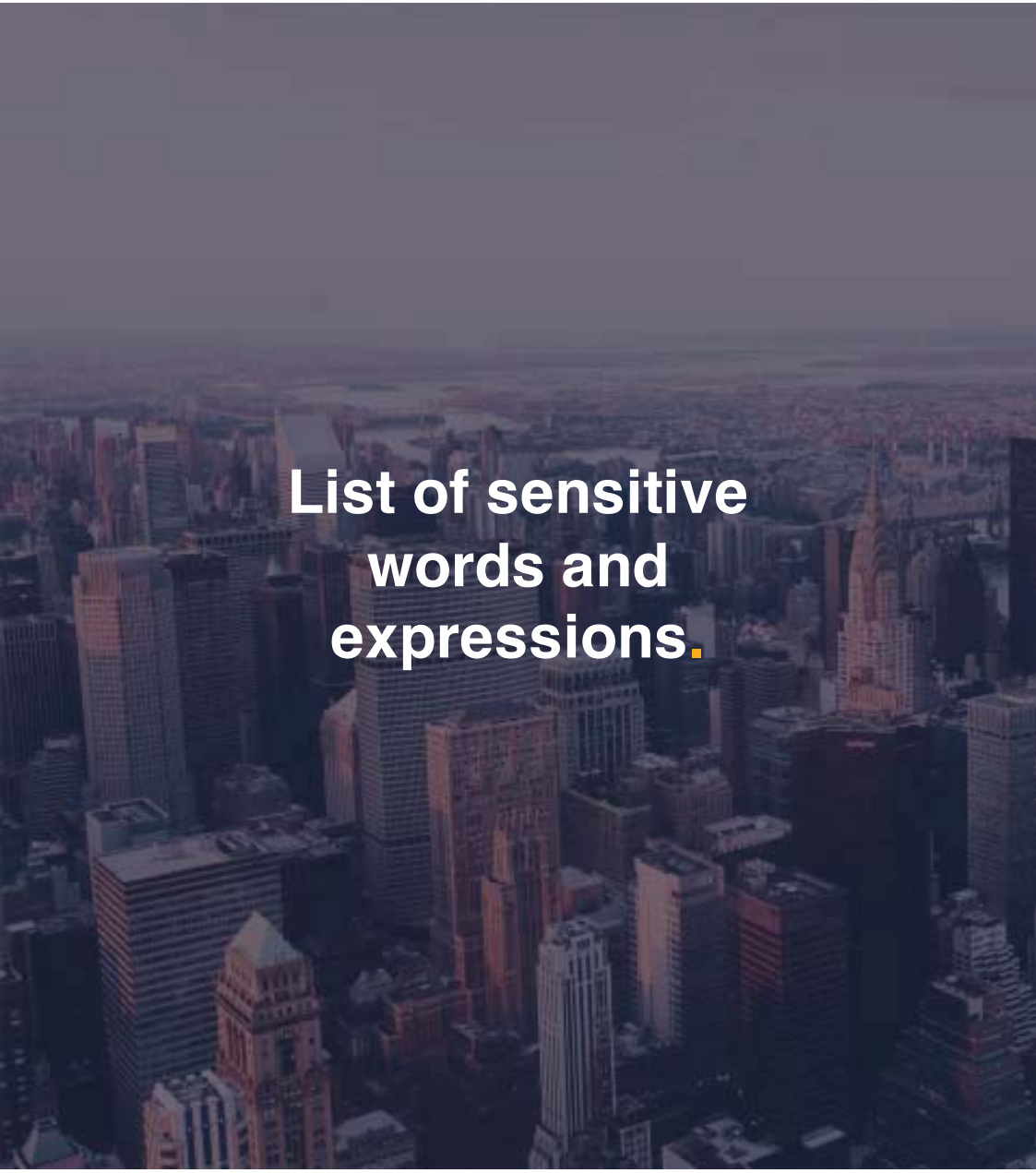
Actions in return to a “defective” translation

Feedback loop.

-
-
-
-

Process of multilingual list correction and modification

Booking.com



List of sensitive words and expressions.

Several categories:

- Free/non-free
- Available/unavailable
- Payment and price
- Location and distance
- Time
- Units of measure
- Other

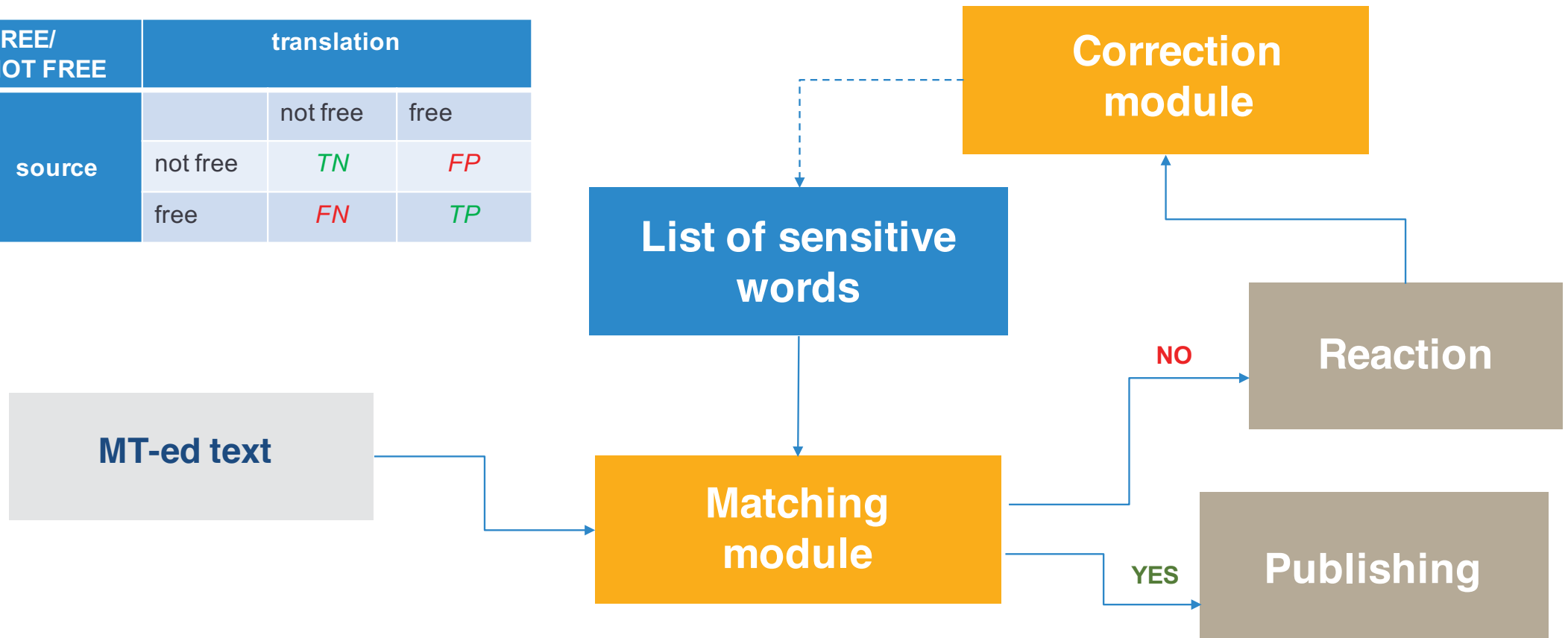
Statistical testing:

- False positives and false negatives

[Booking.com](https://www.booking.com)

Quality assurance system

FREE/ NOT FREE	translation		
	not free	free	
source	not free	<i>TN</i>	<i>FP</i>
	free	<i>FN</i>	<i>TP</i>



Advantages?





Benefits.

1. **Measure the quality of MT for web publishing in a meaningful way**
2. **Minimize subjectivity of MT quality evaluation**
3. **Ensure the accuracy of MT-ed content (with a certain level of confidence)**
4. **Gives a quantifiable measure of business impact caused by some of MT failures**
5. **Increases translators engagement**

Booking.com



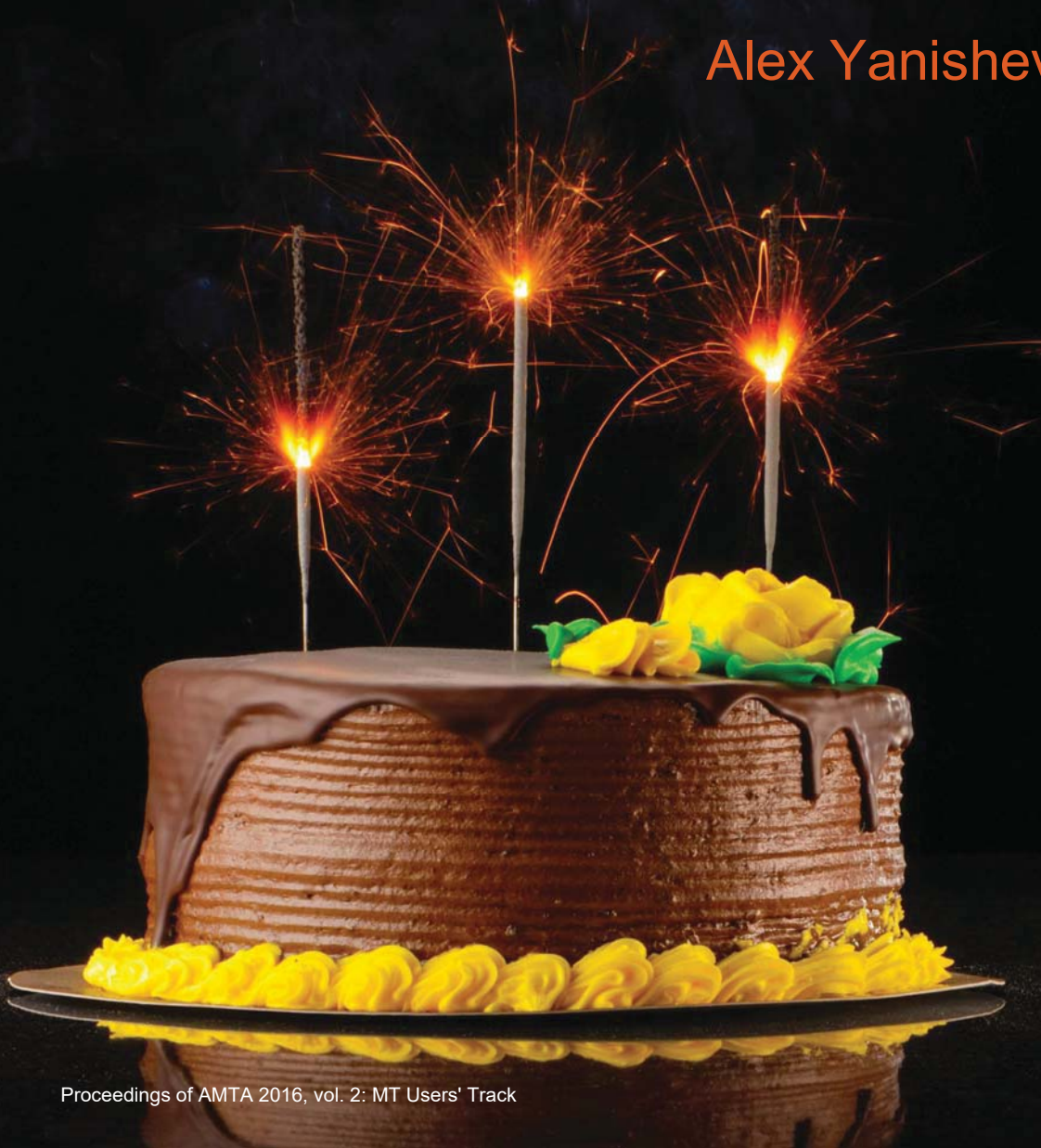
Thank you!

Booking.com

All references to "Booking.com", including any mention of "us", "we" and "our" refer to Booking.com BV, the company behind Booking.com™

I Ate Too Much Cake: Beyond Domain-Specific MT Engines

Alex Yanishevsky



How Much Cake Is Too Much Cake?



What is the Tipping Point?



AGENDA

Recap of Previous Experiments
Challenges for Mature MT Programs
Opportunities for Mature MT Programs

welocalizeo
doing things differently

Recap of Previous Experiments // Part 1

- Criteria for training domain-specific engines
 - ✓ Environment: elegant deployment?
 - ✓ Cost
 - ✓ How different are they from each other
 - ✓ Maintenance (engineering and linguistic feedback implementation)
- Trained and deployed over 50 engines in 13 languages (to and from English)
- Corrected over 300 linguistic issues



Recap of Previous Experiments // Part 2

Implementing Linguistic Feedback

Machine Translation
Incorrect term

[Edit](#)
[Comment](#)
[Assign](#)
[Start Progress](#)
[Resolve Issue](#)
[Close Issue](#)

Details

Type: PE Feedback Status: **OPEN**

Priority: P1 Resolution: Unresolved

Labels: None [+](#)

MTPE State: Newly logged

Target Language: English (US)

Severity: 1

MTPE Error Type: Terminology

Source Text:

Translated Target:

Suggested Target: Table of Contents

Key	Summary & Description	Target Language	Source Text	Translated Target	Suggested Target	Translation from New Engine	Comments
	term (Portable devices)						Term translation correct do NOT add to tune or UD: checked with lead translator, and both translations are correct depending on context
MTR-136	Portable device = terminal mobile term & gender (appliance)	French	Portable devices	Appareils portables	Terminaux mobiles	Terminaux mobiles	
MTR-143	appliance = appliance (masculin)	French	appliance	dispositif	appliance	appliance	add to UD.



Recap of Previous Experiments // Part 3

Savings on MT per quarter



Recap of Previous Experiments // Part 4

MT Usage Per Month



Challenges for Mature MT Programs

- Engage in only those activities that can have an objective, measurable value and/or ROI to the program
- Be wary of not making the engines worse - a threshold beyond which re-training may not be optimal
- Less concerned with automatic scoring as the overriding benchmark for quality since the engines are already at a high quality level
- Stress should be diverted to greater lexical coverage and to fixing high priority and/or high severity linguistic issues that occur numerous times in a corpus

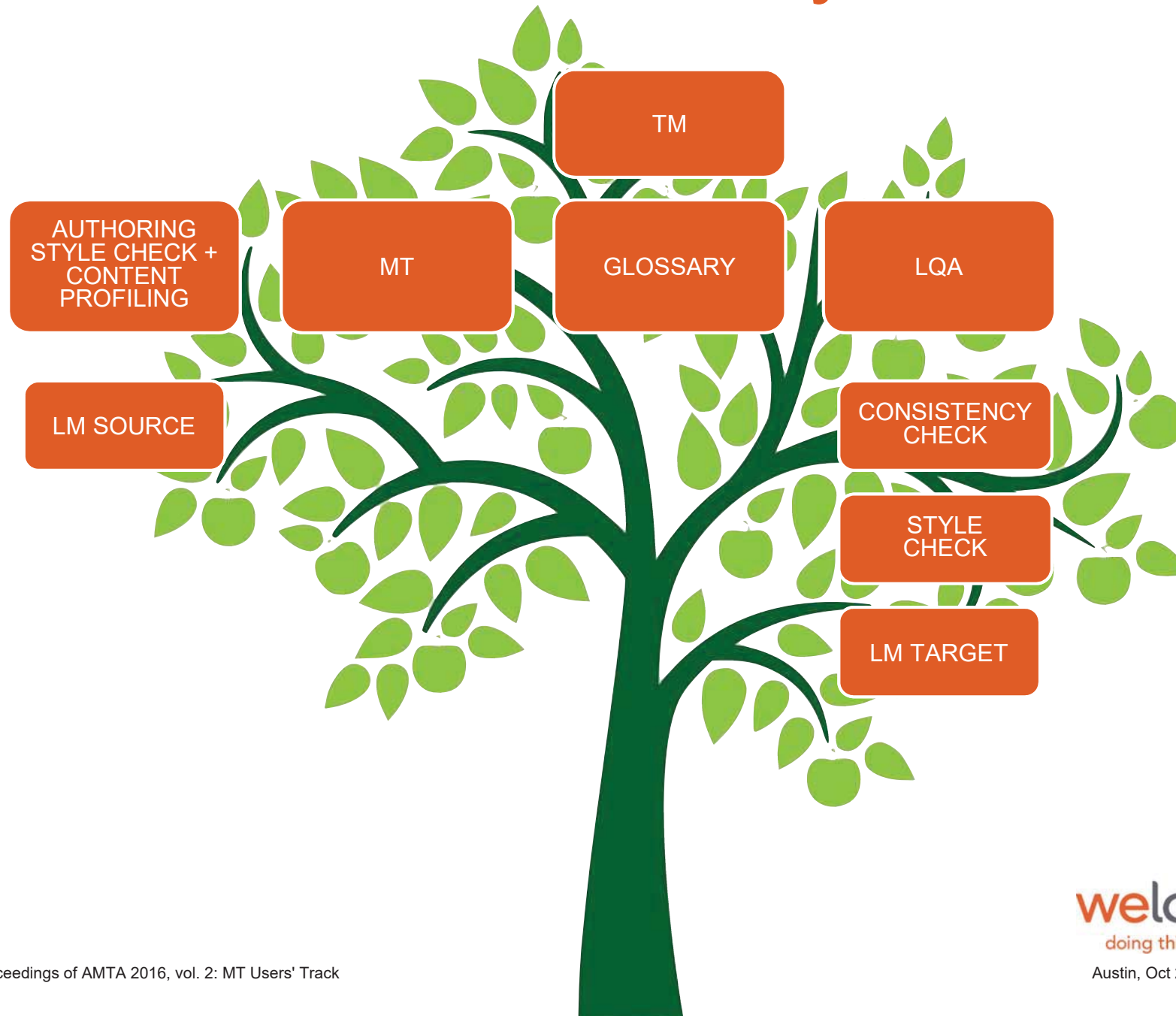


Opportunities for Mature MT Programs

- Pushing the MT engagement upstream
- Analyzing the source content for suitability
- A correlation between the quality of source and the quality and efficacy of MT
- Forecast an MT program, including expected productivity and discounts and make data-driven decisions about the source and its impact before any MT even takes place



The TM Family Tree



Workflow



What is Style?

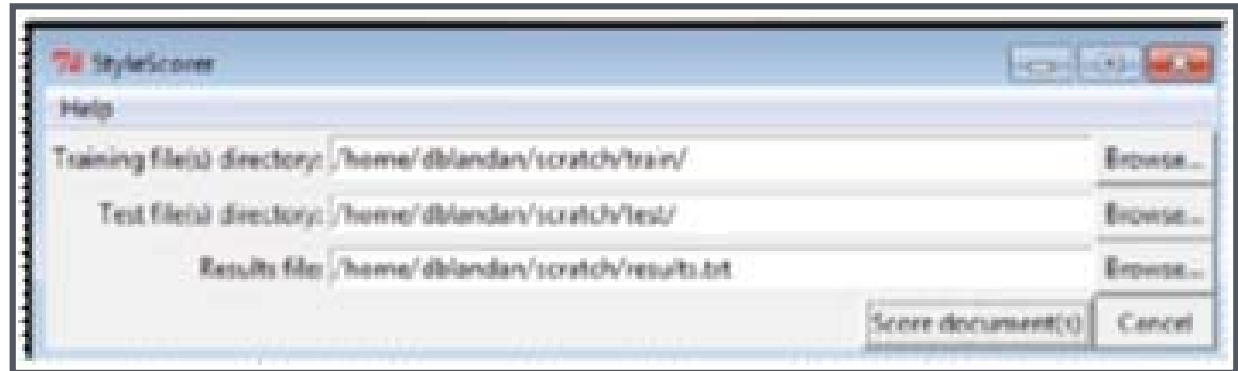
- *Style Can be Formally Defined in a Style Guide That Authors & Translators are Requested to Adhere to (But it Doesn't Have to Be)*
- *Style is a Consistency of Voice Across Multiple Documents*
- *Style Tells us Something About the Target Audience*
- *Style Tends to Reflect Patterns of Conscious Grammatical Decisions*
- *For the Purposes of Style Scorer, the Documents Define the Style, Rather than the Style Defining the Documents*



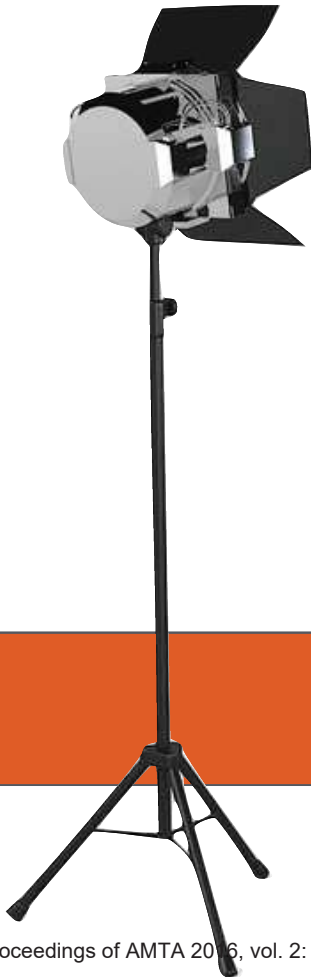
Source: TAUS 2016, Dave Landan, Welocalize

Style Scorer Overview

Combines PPL Ratios,
Dissimilarity Score +
Classification Score



EN-US	OLH	SCORE
DOCUMENT	SM_MANAGER_TAILO RING	3.98
DOCUMENT	_Marketing_whitep aper_4aa5-7132enw	0.74



Style Scorer: Under the Hood

- Score between 0 and 4, with higher score indicating better style match.
- Dissimilarity
Use character n-gram frequency to generate dissimilarity scores. For each document in the Gold Standard, find its maximum dissimilarity compared with all other documents in the Gold Standard. Let G be the set of Gold Standard documents, and g be a document in G . For each g_i in G , calculate $D_{\max}(g_i, G)$. For a document t in the set of Test Documents, calculate $D(g_i, t)$ for all g_i . We want to find the average of the ratio of $D(g_i, t) / D_{\max}(g_i, G)$ across all g_i . That average is the dissimilarity score component.
- Classification
Using a one-class classifier, return 1 if the Test Document is in the Gold Standard class; otherwise return -1.
- Perplexity
Build a language model from the Gold Standard, and get perplexity score for each document within the Gold Standard to establish PPL_{\min} , the theoretical floor for perplexity. For each document in Test Documents, calculate PPL. $PPL_{\min} / PPL_{\text{Test}}$ will be in the range (0,1].

Why Use Style Scorer?

Source

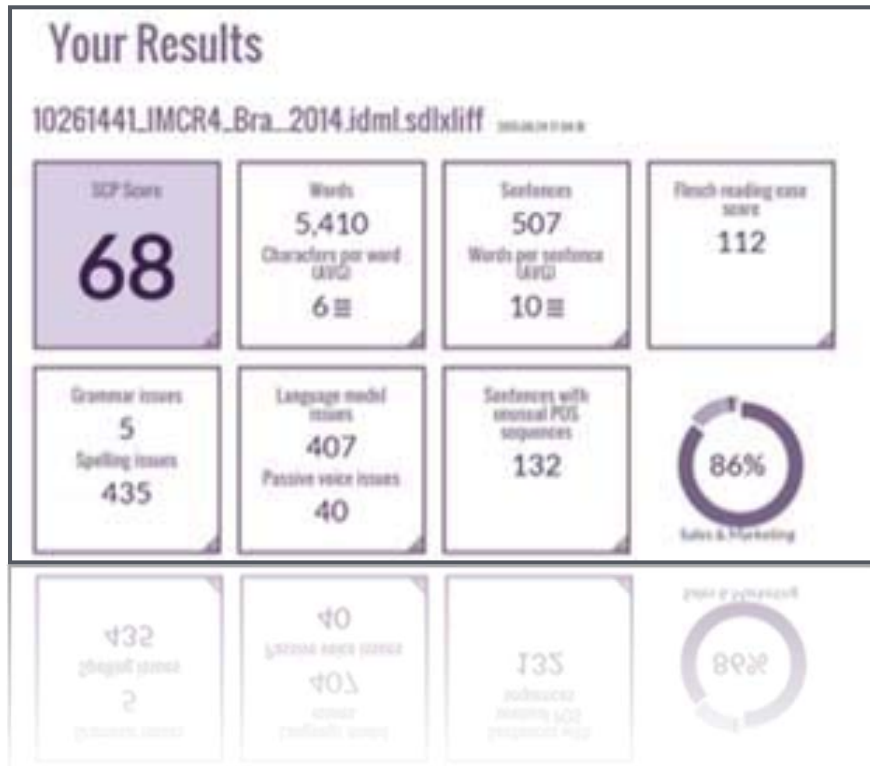
- ✓ Is this really a support document? To what degree is it similar to other support documents, tech doc documents, etc.?
- ✓ Dissimilarity can point to worse quality for raw MT and/or reduced post-editing productivity
- ✓ Find supplemental training data

Target

- ✓ Does this target match the style that the client found to be acceptable in the past?
- ✓ Dissimilarity can point to worse quality and reduced post-editing productivity



Source Content Profiler // Part 1



Source Content Profiler: Part 2

- SCP helps you classify a document
- SCP only works on English source

Words per sentence	Occurrence
1	33
2	30
3	31
10+	3
20+	0
50+	0
10+	3
3	31



Source Content Profiler: Part 3

SCP Highlights Source Issues on a Segment Level

- ✓ *Difficult constructions (e.g. noun phrases)*
- ✓ *Very short or very long sentences*
- ✓ *Passive constructions*



Productivity Metrics

Segment Level

- Source
- Pre-edit Target
- Post-edit Target
- Time to edit (overall, keystroke, pause)
- Number of visits
- Source word count
- Target word count
- Total character inserted
- PE Distance as %



src	preeditText	segTargetFinal	totalVisitCount	totalPeTimeMiliseconds	totalInsertCharCount	srcWordcount
Optional storage bays (used with 3.5" and mixed configuration arrays) house up to 12 DAEs.	Las bahías de almacenamiento opcionales (utilizadas con 3,5" y arreglos combinados de configuración) contienen hasta 12 DAE.	Las bahías de almacenamiento opcionales (utilizadas con arreglos de 3,5" y de configuración combinada) contienen hasta 12 DAE.	4	372413	25	13
Source	Hypothesis	Reference	Lev. Dist. ▾	PE Dist. (% of ref. le ▾	Words ▾	
Contact support	Contactez l'assistance technique	Contacter l'assistance	11	47.83%	2	



Automatic Scoring

File or Project Level

- BLEU
- Meteor
- GTM
- Precision
- Recall
- TER
- PE Distance as %



BLEU	NIST	METEOR	GTM	Avg. PE	TER	Precision	Recall	Length (Hyp./Ref.)	Segs.	Words
46.90	9.86	62.51	68.86	31.80%	40.71	0.69	0.69	1.01	13797	148862
40.30	3.80	05.21	08.80	31.80%	40.11	0.03	0.03	1.01	13131	148805



Goal

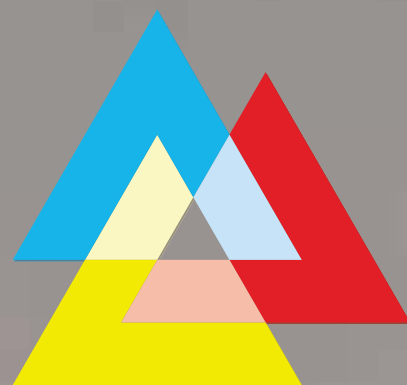


Next Steps

- Build LMs per domain for English source for Style Scorer
- Build LMs per domain for target languages (tier 1 and subsequently tier 2) for Style Scorer
- Build LMs per domain for English source for Source Content Profiler
- Calculate auto-scoring including PE distance for before_PE, after_PE, after_client_review
- Find how strong the correlation is between all the metrics above
- What can be done to prevent some of the issues?



Thank
You!



Iconic
Translation Machines

What? Why? How?
Factors that impact the success of
commercial MT projects

John Tinsley
Iconic Translation Machines

▶ “Why would I need MT?”

What’s the MT value proposition?

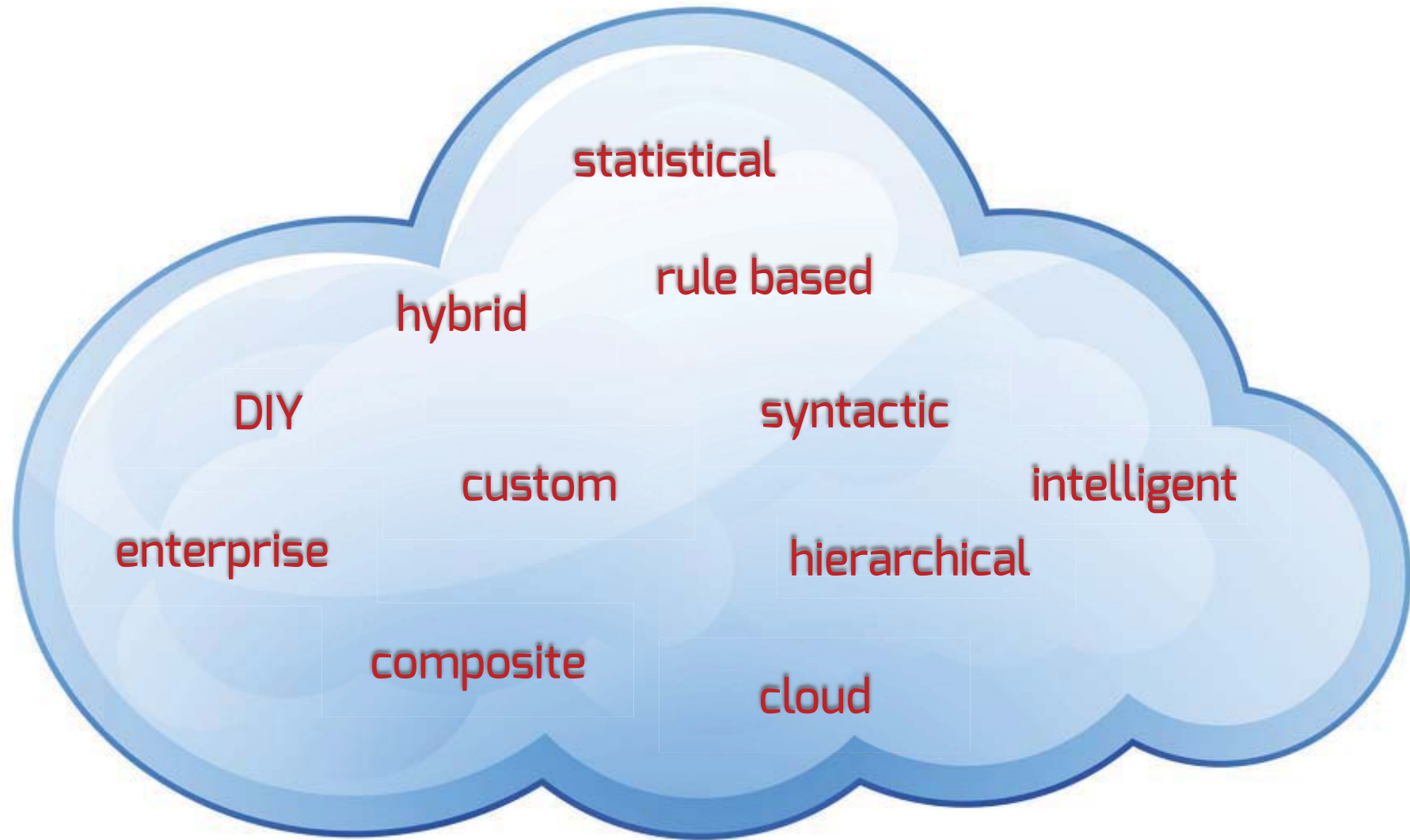
Why MT?

- Speed
- Cost savings
- Time to market
- Your competitors are doing it!

Why Now?

- volume of content is growing
- demand, more words less time
- growth facilitator
- #FOMO – you’re missing out on business

▶ “What type of MT is it?”



▶ How are companies using MT?

What are the use cases for MT?

Translator productivity through post-editing

- The goal of the MT here is to be good enough so that - on the whole - with TMs, translators are faster post-editing some segments
- Challenges
 - development has to focus on reducing needs for edits, not necessarily anything else
 - translator acceptance always a big barrier
 - evaluation can take time and has many factors
 - pricing models



▶ How are companies using MT?

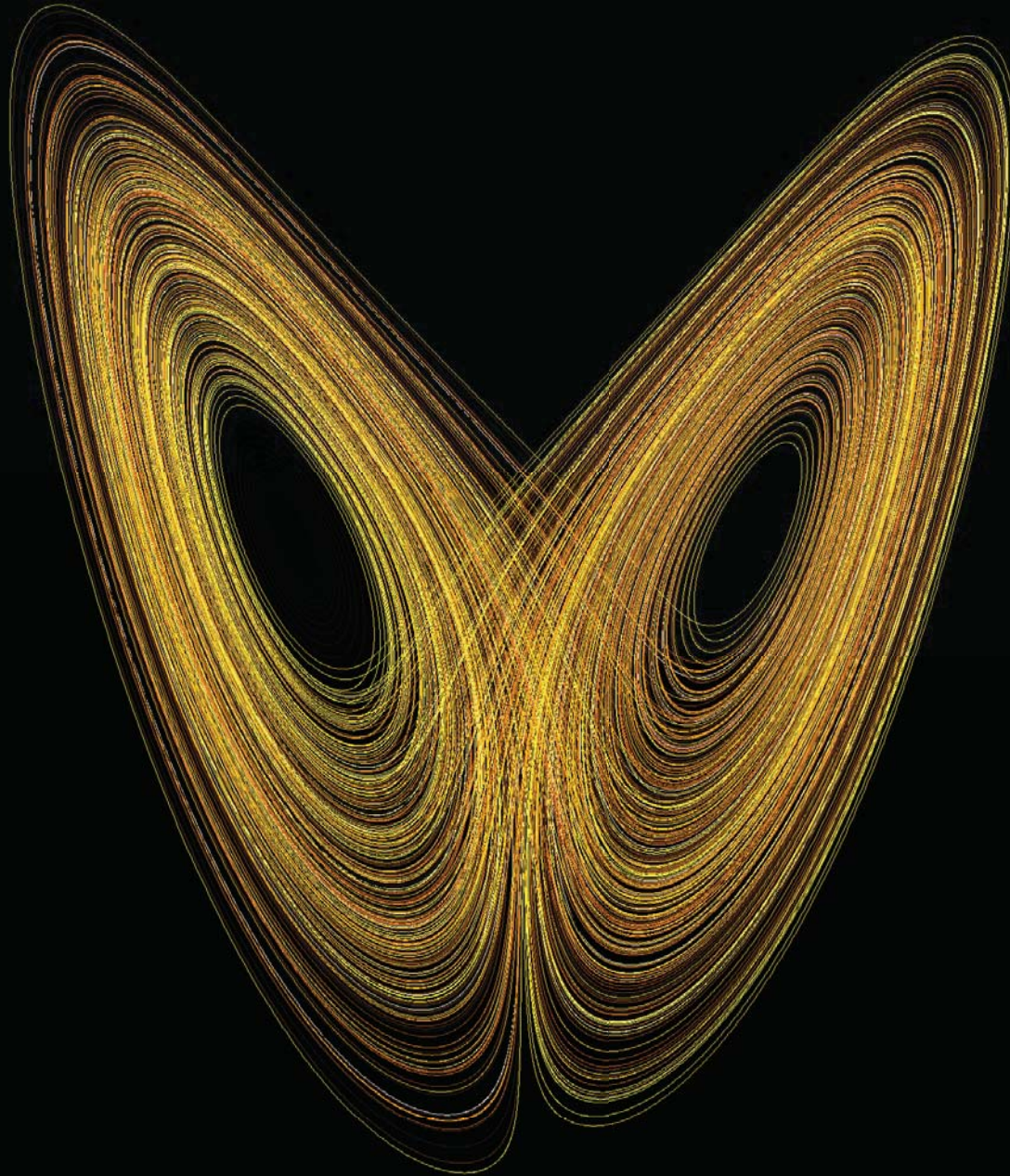
What are the use cases for MT?

▶ MT for information

- The goal is to produce MT that's fit for a particular purpose *as is*
- Arguably easier from an MT development perspective
- Often high-volumes = more achievable



Butterfly Effect



▶ The 8 Factors influencing MT suitability





Not all languages are created equal



Social Media



 @Wimbledon
Wimbledon ✓

We know UR all waiting for Friday's schedule at #Wimbledon but it won't be out until we see how far we get with the matches outside 2nite

20 hours ago via TweetDeck ☆ Favorite ↻ Retweet ↩ Reply

Highly Technical

Chem. Abs. Vol. 66, 1967 Page 9799

104825b Methyl 3,4-dihydroxybenzyl ketones. Merck & Co., Inc. (by David F. Hinkley and John Budavari). Fr. 1,450,200 (Cl. C 07c, A 61k), Aug. 19, 1966; U.S. Appl. Oct. 21, 1964; 6 pp. The title compds. are prepd. and can be used as chem. intermediates. Thus, a soln. of 60 g. 3,4-(MeO)₂-C₆H₃CHO in 500 ml. C₆H₆ is cooled to 0°, a mixt. of 30 g. NaO-Me and 44.1 ml. MeCHClCO₂Me added in 30 min. at 2-5°, and the mixt. agitated ~1 hr. at ~20° to give Me α-methyl-α,β-epoxy-β-(3,4-dimethoxyphenyl)propionate (I). A concd. soln. of I in C₆H₆ is treated with 400 ml. MeOH, the mixt. heated to 75°, 240 ml. 10% NaOH added in 30 min., and the mixt. heated to 82° and treated with 50 ml. water to give Na α-methyl-α,β-epoxy-β-(3,4-dimethoxyphenyl)propionate (II). An aq. soln. of II is heated to 100° for 30 min., 70 ml. concd. HCl added, and the mixt. heated 30 min. at 100° to give Me 3,4-dimethoxybenzyl ketone, which can be used in the prepn. of 3,4-(HO)₂C₆H₃CH₂-CMe(NH₂)CO₂H. Similarly prepd. are 3,4-(HO)₂C₆H₃CH₂-COMe (b.p. 165-8°), 4,3-HO(MeO)C₆H₃CH₂COMe, and methyl piperonyl ketone. BDPF

User Generated Content

9 March 2016

James
USA
3 reviews

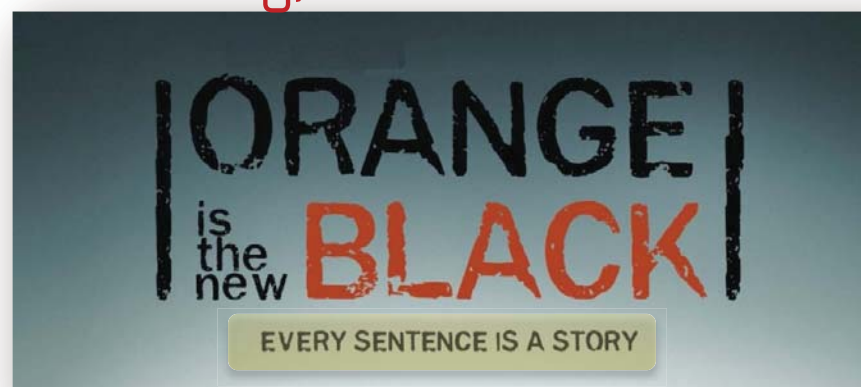
8.8 "Great trip to Dublin"

Leisure trip Couple Standard Guest Room
Stayed 2 nights Submitted via mobile

- The bed was two twin beds put together and me and my girlfriend kept fallin in the middle (since we like to cuddle) and that was irritating

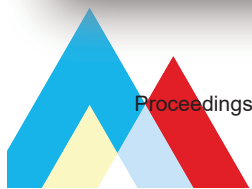
+ Late nite room service was awesome

Marketing, Nuanced



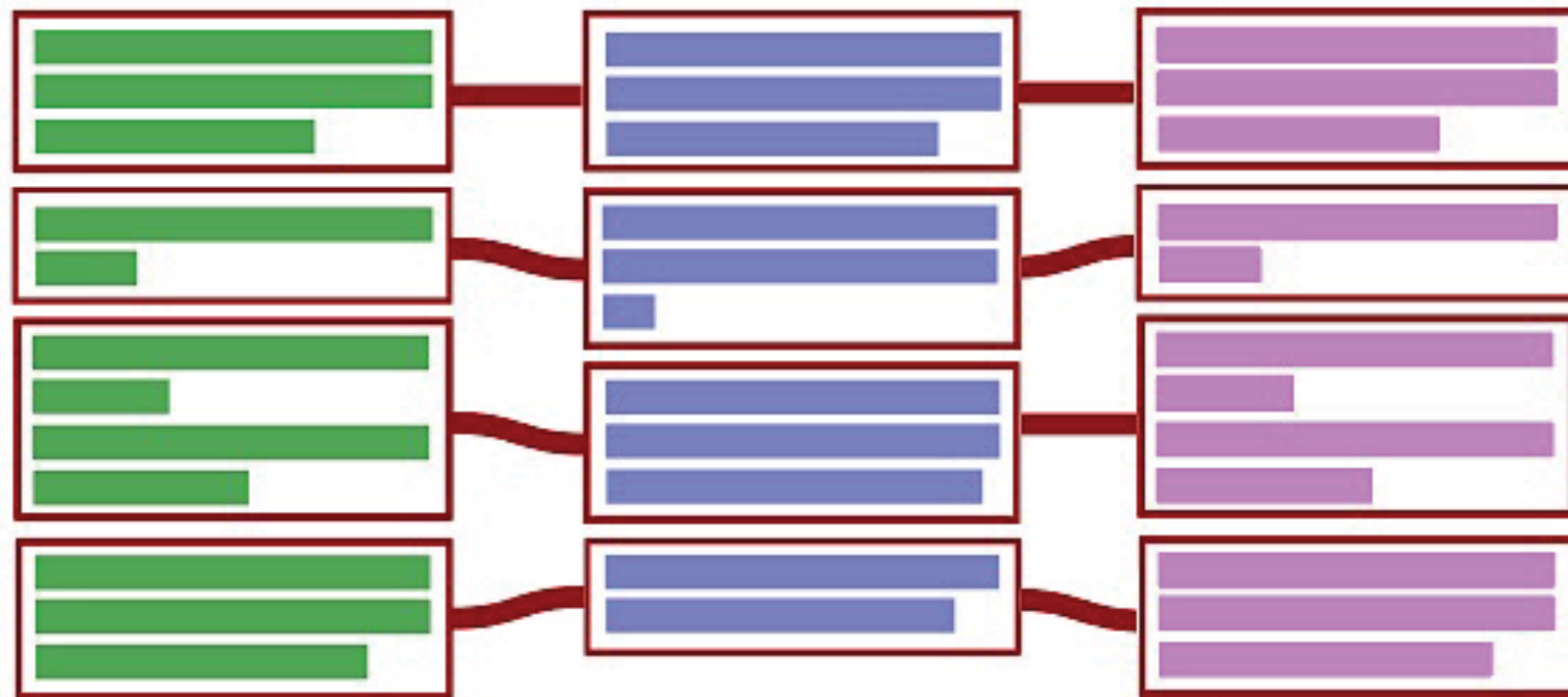
ORANGE
is the new BLACK

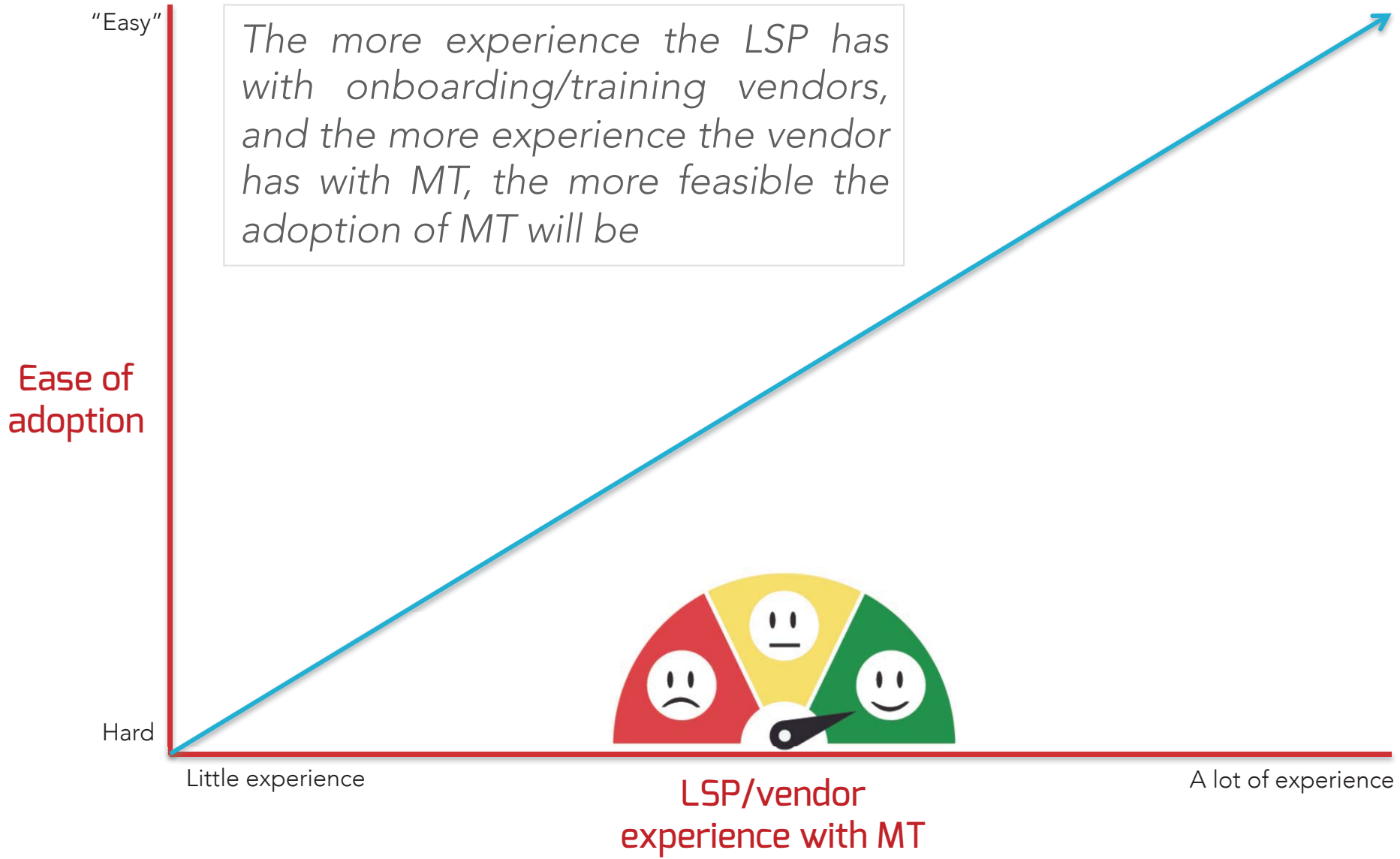
EVERY SENTENCE IS A STORY





Corpora. Dictionaries. Terminology.





Standard vs Custom Integration

SDL | Trados
Studio 2015

memoQ

WF WORDFAST

m matecat™



“instant” solution costs rise proportionality with the number of languages and the throughput needs



 **High TM
Leverage**

 **Low MT
Effectiveness**

Matches	# words
Context	403,803
100%	585,459
95-99%	50,366
85-94%	41,604
75-84%	32,319
50-74%	18,972
No Match	81,119
Total	1,213,643

Only 8% of
all words go
to MT



REALITY-CHECK

- Fully automatic human quality
- 300% post-editing productivity
- French to Spanish == English to Korean
- Best performance out of the box



► The 8 Factors influencing MT suitability



▶ What questions should YOU be asking?

- “What volume of words do you estimate for the project?”
- “Do we have translation memories, glossaries that are relevant? Can we create them?”
- “If so, what leverage are we getting?”
- “Do we have post-editors? Access to a supply chain?”
 - “what experience do they have?”
- “Where will MT fit in the workflow (depending on the use case)?”
- “What variety is there in the content that the MT will be processing?”
- “Why aren’t you using Google Translate?”
- “Is there sufficient budget for this project?”



▶ john@iconictranslation.com
www.iconictranslation.com
twitter.com/iconictrans



"How good is the quality?"

"How much training data do I need?"

"How frequently can I retrain the engine?"

"Do you do language X?"

"How do you measure performance over time?"

"What happens to my data?"



Assessing Translation Quality Metrics

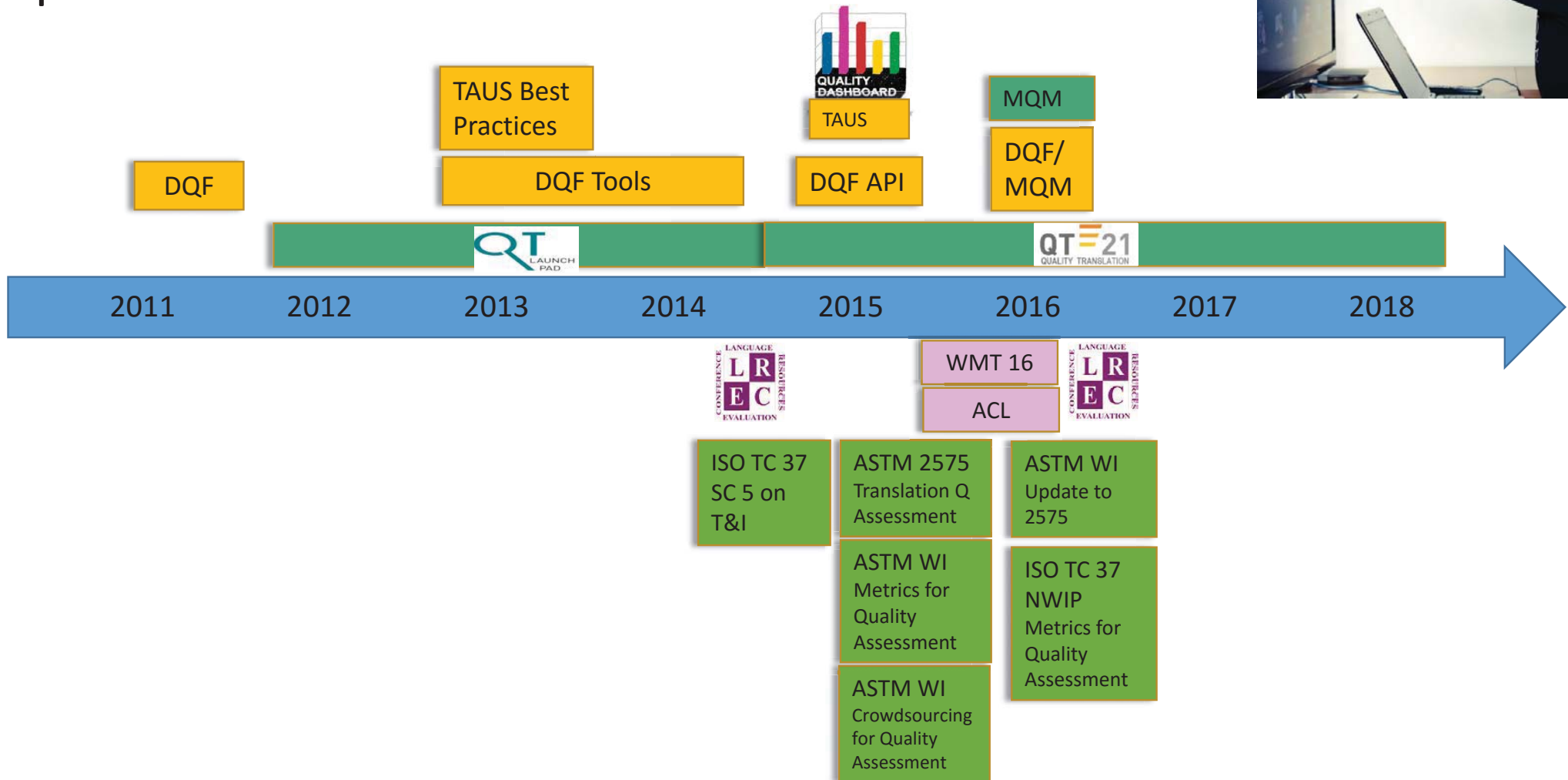
29 October 2016

Jennifer DeCamp



©2016 The MITRE Corporation. All rights reserved.

What is going on with assessing quality of production translation?



©2016 The MITRE Corporation. All rights reserved.

Why is this important to us?

- These standards could have a real impact
 - Contracting organizations
 - LPTA (Lowest Price/Technically Acceptable)
 - Best Value
 - Potential for being used as policy
 - We use tools from TAUS integrators
- But TAUS, MQM, and ASTM based on industry surveys, requirements, practices, and efforts (e.g., TAUS, GALA, SAE)
 - Need government requirements and review



What do we do?

- Build our awareness of what is happening in this space
 - Understand marketing vs. reality
 - Assess impact
- Define our user requirements and make those requirements known in ASTM and ISO, and to companies implementing TAUS software
- Find ways we can help
 - Definitions of metrics, measures, etc. compatible with current systems
 - Definitions of specific metrics and error types
 - Contacts with the research community and their extensive findings
 - Coordination between standards efforts (e.g., AMTA workshop)
 - Close review and possible testing of tools and standards

©2016 The MITRE Corporation. All rights reserved.

What are we doing in this presentation?

- Raising awareness
 - What is the state of assessment production translation quality?
 - What are these standards and efforts?
 - What are other approaches?
 - What is a first cut at decisions?
 - What was said in the workshop on October 28?
- Discussing next steps

©2016 The MITRE Corporation. All rights reserved.

What is the state of approaches to MT or professional translation evaluation?

“Current approaches to Machine Translation (MT) or professional translation evaluation, both automatic and manual, are characterized by a high degree of fragmentation, heterogeneity, and a lack of interoperability between methods. As a consequence, it is difficult to reproduce, interpret, and compare evaluation results.”

Rehm, G., A. Burchardt, O. Boja, C. Dugast, M. Federico, J. van Genabith, B. Haddow, J. Hajič, K. Harris, P. Koehn, M. Negri, M. Popel, L. Specia, M. Turchi, and H. Uszkoreit, (2016). Workshop on Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem. In *Language Resources Evaluation Conference Proceedings*. Retrieved June, 2016, from <http://lrec2016.lrec-conf.org/en/about/conference-proceedings/>

©2016 The MITRE Corporation. All rights reserved.

How well do these measures apply to human-translated material?

“Quality measured by BLEU, NIST, METEOR etc. does not indicate the type of quality problems,” and that “these metrics are also better suited for measuring progress in the ‘ugly’ or ‘bad’ sectors of the quality spectrum....Even the human evaluations usually by ranking, often done by CS researchers and students, do not help the human translators....and the LISA [Quality Assessment] QA model, EN-15038 and current International Organization for Standardization (ISO) work on a successor, are not known and not used in MT research.”

Uszkoreit, H. and A. Lommel. (2014). Multidimensional Quality Metrics: A New Unified Paradigm for Human and Machine Translation Quality Assessment. Retrieved October, 2016 from <http://www.qt21.eu/launchpad/sites/default/files/MQM.pdf>

©2016 The MITRE Corporation. All rights reserved.

What are some of the problems?

“Humans differ in their understanding of quality problems, their causes, and the way to fix them.”

Factors impacting human identification and classification of errors, include:

- Disagreement as to the precise spans that contain an error
- Errors whose categorization is unclear or ambiguous
- Differences of opinion about whether something is or is not an error or how severe it is.”

Lommel, A., M. Popović, A. Burchardt (2014a). Assessing Inter-Annotator Agreement for Translation Error Annotation. Workshop on Automatic and Manual Metrics for Operational Translation Evaluation. *Proceedings of the Language Resources Evaluation Conference*, 2014. Retrieved June, 2016 from <http://www.lrec-conf.org/proceedings/lrec2014/index.htm>

©2016 The MITRE Corporation. All rights reserved.

What is standard practice in industry?

- Methods for assessing production translation quality
 - Extent to which the product met customer specifications
 - BLEU scores
- Demand for more consistent means of conducting assessments

O'Brien, S. (2012). Towards a Dynamic Quality Evaluation Model for Translation. *The Journal of Specialized Translation*, 17:January, pp. 55-77.

©2016 The MITRE Corporation. All rights reserved.

What is industry doing? TAUS

- Translation Automation User Society
- Dynamic Quality Framework (DQF), 2011
- DQF tools, 2013-2014
- Quality Dashboard, 2015
 - Some capabilities available to anyone; others available only to members
 - Developed and copyrighted by TAUS
- DQF API and member integrators, 2015
- Also have a Productivity Dashboard



Measure and benchmark your translation quality

©2016 The MITRE Corporation. All rights reserved.

What else is industry doing? MQM

- Multidimensional Quality Metrics
- Developed and copyrighted by DFKI and QT LaunchPad
- Based on the following definition by Melby
 - “A quality translation (1) demonstrates required accuracy and fluency (2) for the audience and purpose and (3) complies with all other negotiated specifications, taking into account end-user needs”
- Provides
 - A hierarchical catalog of issue types
 - Dimensions (based on ISO/TS-11669) to guide users in selecting appropriate issue types
 - A method for declaring/describing a particular metric
 - An inline format for tagging issues in XML files
 - A reporting format with scoring formula for determining scores/acceptance
- Error typology integrated with the one from DQF
- Study indicated “*Low inter-rater reliability but better with classifying errors using MQM than with identifying errors*” (Snow 2015).
- Basis for ASTM WK 46397 *Language Quality Assurance*

©2016 The MITRE Corporation. All rights reserved.

What are these standards?



F15.48 Committee on Language Services and Products/ Subcommittee on Language Translation

ASTM F2575 <i>Standard Guide for Quality Assurance in Translation</i>	2014
ASTM Work Item (WK) 47362 <i>Standard Practice for Quality Assurance in Translation</i>	2015
ASTM WK 46397 <i>Language Quality Assurance</i>	2016
ASTM WK 46396 <i>New Practice for the Development of Translation Quality Metrics</i>	2016

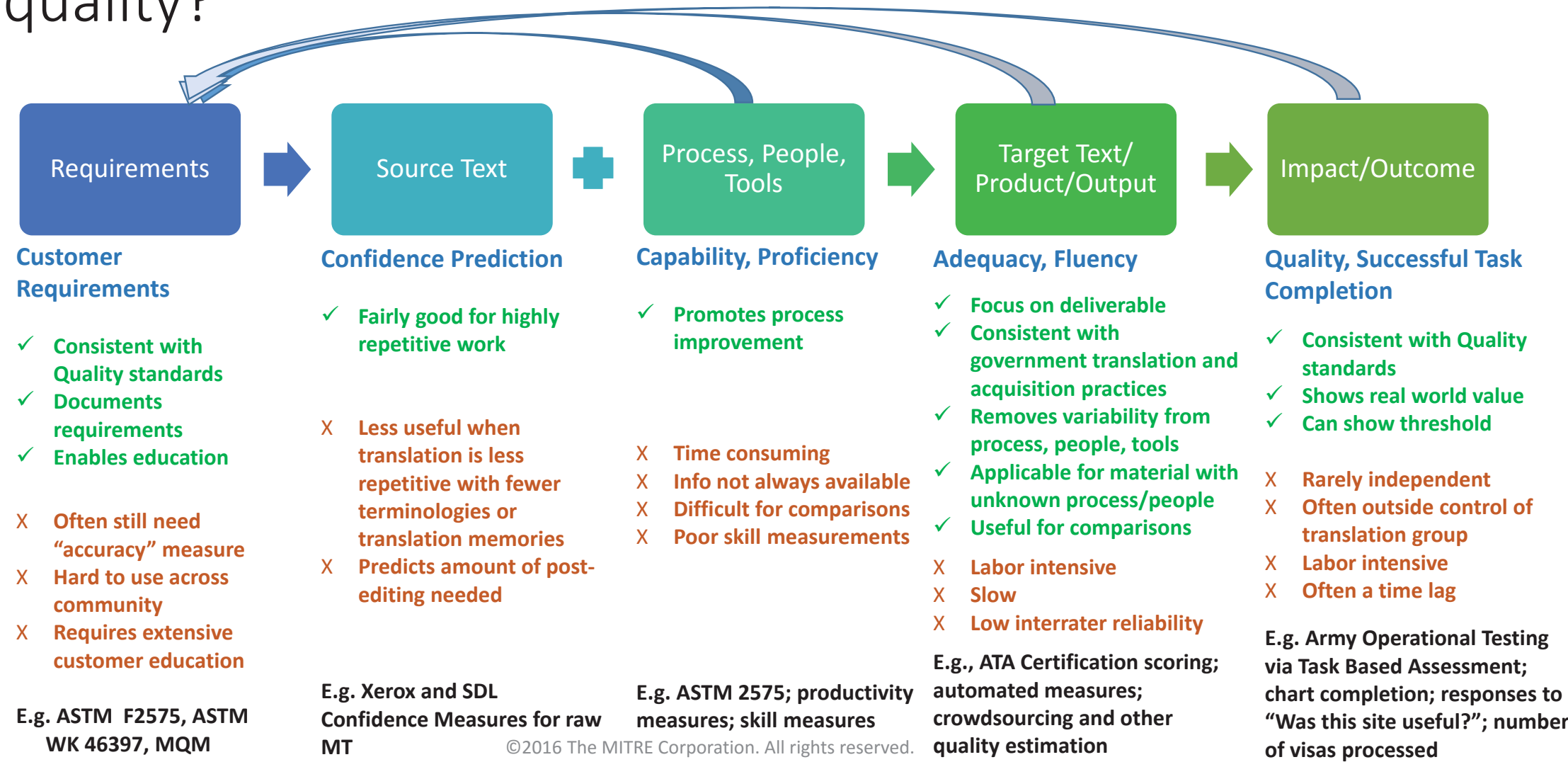


Technical Committee 37 on Terminology and Other Language and Content Resources/ Subcommittee on Translation, Interpreting and Related Technology

ISO/NP 21999 <i>Translation Quality Assurance and Assessment – Models and Metrics</i>	2016
---	------

©2016 The MITRE Corporation. All rights reserved.

What are options for assessing production translation quality?



©2016 The MITRE Corporation. All rights reserved.

What is a first pass at decisions needs?

Population	Decision
Customer	Is the translation ready for use?
	Does additional work needs to be negotiated?
	Should the provider be paid?
	Did the translator or LSC provide good value?
	Should pricing be adjusted for future contracts?
Translator or LSC	Is the translation is ready for delivery to customer?
	If not, what else needs to be done?
	Does additional work needs to be negotiated?
	Should pricing be adjusted for future contracts?
	How do current practices and workflows compare to proposed ones?
	How do tools compare?
	How does the translator or the company compare to others?
LSC	Are the specific translators doing a good job in this language pair, domain, etc.?
	Is translator performance being affected by stress, fatigue, or other factors?
	Should the translators receive pay increases?
	Are some translators better than others at certain types of work?
End user	How reliable is the translated information?
	Is a re-translation warranted?
Researchers & Developers	How can we improve the tools?
	Does one tool or process work better than another?
	Can we improve the translation (e.g., through annotated data sets)?

What are government decisions needs?

©2016 The MITRE Corporation. All rights reserved.

What happened at the AMTA Workshop on Assessing Production Translation Quality?

- Facilitator

- Jennifer DeCamp Chair, ATA Standards Committee; member ASTM, ISO, ILR

- Industry

- TAUS

- Achim Ruopp TAUS Director of Research and Development

- SDL

- Daniel Brockman SDL Director of Product Management

- MQM

- Alan Melby Co-Author of MQM

- Standards Groups

- ASTM

- Amanda Curry Chair, Translation Subcommittee

- ISO

- Sue Ellen Wright Head of U.S. Delegation; member ASTM
- Monika Popiolek Chair, WG on ISO/NP 21999

- Interagency Language Roundtable (ILR)

- Maria Brau Chair, Translation Subcommittee

©2016 The MITRE Corporation. All rights reserved.

Summary

- Different populations with different decisions, decision factors, and metrics
- Inconsistent practice
- Efforts that are “fragmented, heterogeneous, and non-interoperable” (Rehm et. al., 2016)
- Conflicting terminology
- Still a divide between the MT research and HT practitioners
- Promising efforts to improve evaluation of production translation quality (TAUS Quality Dashboard, MQM, standards)
 - Extraordinary outreach
 - But based on industry practice, surveys, requirements, and needs

©2016 The MITRE Corporation. All rights reserved.

Recommendations

- Negotiate and document a common terminology and structure for evaluation (e.g., metrics, measures, methods, tools) and for the specific metrics (e.g., adequacy, fluency) and error types.
- Analyze the relationship between decision needs, metrics, and measures.
 - Do the metrics (e.g., fluency and adequacy) meet the needs of the decision-maker?
 - What do specific measures actually say about the metrics, and how may that be presenting an inaccurate or incomplete picture for the decision-maker?
- Leverage and/or conduct analysis on how standards, methods, measures, and tools support specific translation requirements.
- Provide resources for language service providers and others to easily access information on relevant metrics and measures.
- Provide standards within a framework, and reference that framework at the beginning of each standard along with the audience of the standard and the part of the problem that it is addressing.
- Develop more best practices documents and other guidelines for developing tools; encourage adoption through funding organizations and through professional organizations.
- Continue to bring together the research and language services communities.
- Increase outreach from the standards committees to researchers, decision-makers, and other users.

©2016 The MITRE Corporation. All rights reserved.

References

- Agarwal, A. and A. Lavie (2008). METEOR, M-BLEU, and M-TER: Evaluation Metrics for High Correlation with Human Rankings of Machine Translation Output. In *Proceedings of the Third Workshop on Statistical Machine Translation*, Columbus, June 2008, pp. 115-118.
- ASTM International (2016). *ASTM 2575 Standard Guide for Quality Assurance in Translation*. Retrieved February, 2016, from <http://www.astm.org/Standards/F2575.htm>.
- Bojar, O., C. Federmann, B. Haddow, P. Koehn, M. Post, and L. Specia (2016). Ten Years of WMT Evaluation Campaigns: Lessons Learnt. In *Language Resources Evaluation Conference Proceedings*. Retrieved June, 2016, from <http://lrec2016.lrec-conf.org/en/about/conference-proceedings/>
- Lommel, A., M. Popović, A. Burchardt (2014a). Assessing Inter-Annotator Agreement for Translation Error Annotation. Workshop on Automatic and Manual Metrics for Operational Translation Evaluation. *Proceedings of the Language Resources Evaluation Conference, 2014*. Retrieved June, 2016 from <http://www.lrec-conf.org/proceedings/lrec2014/index.html>.
- Lommel, A., Uszkoreit, H., and Burchardt, A. (2014b). Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Revista tradumàtica: technologies de la traducció*, 12, December 2014. ISSN 1578-17559.
- Melby, A. (2016). A Spectrum from All MT at the Left to All HT at the Right End. Two-page Spec-Oriented Description for Spectrum v 4b.
- O'Brien, S. (2012). Towards a Dynamic Quality Evaluation Model for Translation. *The Journal of Specialized Translation*, 17 January, pp. 55-77.
- Rehm, G., A. Burchardt, O. Boja, C. Dugast, M. Federico, J. van Genabith, B. Haddow, J. Hajič, K. Harris, P. Koehn, M. Negri, M. Popel, L. Specia, M. Turchi, and H. Uszkoreit, (2016). Workshop on Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem. In *Language Resources Evaluation Conference Proceedings*. Retrieved June, 2016, from <http://lrec2016.lrec-conf.org/en/about/conference-proceedings/>
- Snow, T. (2015). *Establishing the Viability of the Multidimensional Quality Metric*. Dissertation, Brigham Young University. Paper 5593.
- Translation Automation User Society (2016). The TAUS Dynamic Quality Dashboard: An Industry Collaborative Platform for Translation Quality and Tracking. Retrieved August, 2016 from <http://www.slideshare.net/TAUS/quality-dashboard-an-industry-collaborative-platform-for-translation-quality-measurement-and-tracking-achim-ruopp-and-jaap-van-der-meer-taus/>
- U.S. Government Interagency Language Roundtable (2016). ILR Skill Level Descriptions for Translation Performance. Retrieved May, 2016 from <http://www.govtilr.org/skills/AdoptedILRTranslationGuidelines.htm>.
- Uszkoreit, H. and A. Lommel. Multidimensional Quality Metrics: A New Unified Paradigm for Human and Machine Translation Quality Assessment. Retrieved October, 2016 from <http://www.qt21.eu/launchpad/sites/default/files/MQM.pdf>
- Van Ess-Dykema, C., J. Phillips, F. Reeder, L. Gerber (2011). Paralinguist Assessment Decision Factor for Machine Translation Output: A Case Study. Retrieved October, 2016 from <http://www.mt-archive.info/AMTA-2010-VanEss-Dykema.pdf>

©2016 The MITRE Corporation. All rights reserved.

OFFICE OF THE DIRECTOR OF NATIONAL INTELLIGENCE



MATERIAL: MAchine Translation for English Retrieval of Information in Any Language (Machine translation for English-based domain-appropriate triage of information in any language)

Dr. Carl Rubino
IARPA
September 27, 2016



Disclaimers

- This Proposers' Day Conference is provided solely for information and planning purposes.
- The Proposers' Day Conference does not constitute a formal solicitation for proposals or proposal abstracts.
- Nothing said at Proposers' Day changes the requirements set forth in a Broad Agency Announcement (BAA).



Proposers' Day Goals

- Familiarize participants with IARPA's interest in human language technology.
- Familiarize participants with IARPA's mission and how to do business with IARPA.
- Provide answers to participants' questions.
 - This is your chance to provide input to the program plan.
- Foster discussion of synergistic capabilities among potential program participants, i.e., facilitate teaming.
 - Take a chance – someone might have a missing piece of your puzzle.



Important Points

- Proposers' Day slides will be posted on iarpa.gov
- Please save questions for the end; write on notecards
- Posters are available for browsing during break/lunch
- Government will not be present during the poster/teaming session
- Discussions with PM allowed until BAA release
 - Once BAA is published, questions can only be submitted and answered in writing in accordance with the BAA guidance.



MATERIAL's Motivating Scenario





Navigating our multilingual world



ಅಂಕಾರಾ: ಟರ್ಕಿ ರಾಜಧಾನಿ ಅಂಕಾರದಲ್ಲಿ ಮಿಲಿಟರಿ ಪಡೆಯನ್ನು ಗುರಿಯಾಗಿರಿಸಿ ಬುಧವಾರ ಸಂಭವಿಸಿದ ಕಾರು ಬಾಂಬ್ ದಾಳಿಯಲ್ಲಿ 28 ಮಂದಿ ಸಾವಿಗೀಡಾಗಿದ್ದು, 61 ಮಂದಿಗೆ ಗಾಯಗಳಾಗಿವೆ. ಮಿಲಿಟರಿ ವಾಹನಗಳು ಹಾದು ಹೋಗುತ್ತಿರುವ ದಾರಿಯಲ್ಲಿ ಈ ಬಾಂಬ್ ಸ್ಫೋಟ ಸಂಭವಿಸಿದ್ದು, ಇದರ ಹಿಂದೆ ಯಾರ ಕೈವಾಡವಿದೆ ಎಂಬುದ ಸದ್ಯ ಪತ್ತೆಯಾಗಿಲ್ಲ ಎಂದು ಟರ್ಕಿ ಡೆಪ್ಯೂಟಿ ಪ್ರಧಾನಿ ನುಮಾನ್ ಕುರ್ತುಲ್ಮುಸ್ ಹೇಳಿದ್ದಾರೆ.

Military convoy

Car Bomb

ಪಾರ್ಲಿಮೆಂಟ್ ಮತ್ತು ಮಿಲಿಟರಿ ಕೆಲವು ಸ್ಥಳಗಳನ್ನು ಒಳಗೊಂಡು ನಡೆದಿತ್ತು. ಈ ಸ್ಫೋಟದಿಂದ ಸಾವಿರಾರು ಜನರೇ ಆಗಿದ್ದರೂ ಅವರ ವಿರುದ್ಧ ಕ್ರಮ ಕೈಗೊಳ್ಳಲಾಗುವುದು ಎಂದು ಅಧ್ಯಕ್ಷ ರೆಸೆಪ್ ತಯ್ಯಿಪ್ ಎಂದಿದ್ದಾರೆ. ಈ ದಾಳಿಯ ಹಿಂದೆ ಜಿಹಾದಿ ಇಲ್ಲವೇ ಕುರ್ದಿಶ್ ಬಂಡುಕೋರರ ಕೈವಾಡವಿದೆ ಎಂದು ಟರ್ಕಿ ಸರ್ಕಾರ ಆರೋಪಿಸಿದೆ.

Jihadist

New, unexpected languages do appear in an analyst's dataset which would require much time and investment to deploy automated methods for triage.



Working with a New Language

- Data in a new language may contain information critical for intelligence analysis but:
 - Many/most domain experts do not speak the language.
 - Few/no analysts speak the language.
 - No machine translation (MT) systems from that language to English are available to aid in analysis.
 - Only small amounts of new language to English bitext training data are available to build an MT-based system, and no domain-matched adaptation bitext is available to customize engines to support English-speaking analysts.
- If **Human Language Technology (HLT)** for the target language could be quickly deployed with output displayed in **English**, it would enable the domain experts to focus their efforts on the most relevant portions of the data



MATERIAL Goal

- Revolutionize multilingual triage by enabling rapid development of language-independent methods to field systems capable of fulfilling domain-specific cross-language information retrieval tasks over both text and speech data, with:
 - Limited bitext and transcribed speech training data
 - English domain-specific queries as input
 - English summaries of retrieved results as output
 - Methods for domain adaptation and portability to new languages
 - Assessment of the technology via a resonating end-to-end use case



The MATERIAL System

- An “English-in, English-out” information retrieval system that, given a domain-sensitive English query, will retrieve relevant data from a large multilingual repository and display the retrieved information in English as summaries that reflect the document relevance:











Query Format

- **Domain-specific (e.g., Agriculture, Government, Business, Entertainment)**
- **Address domain-restricted information need**

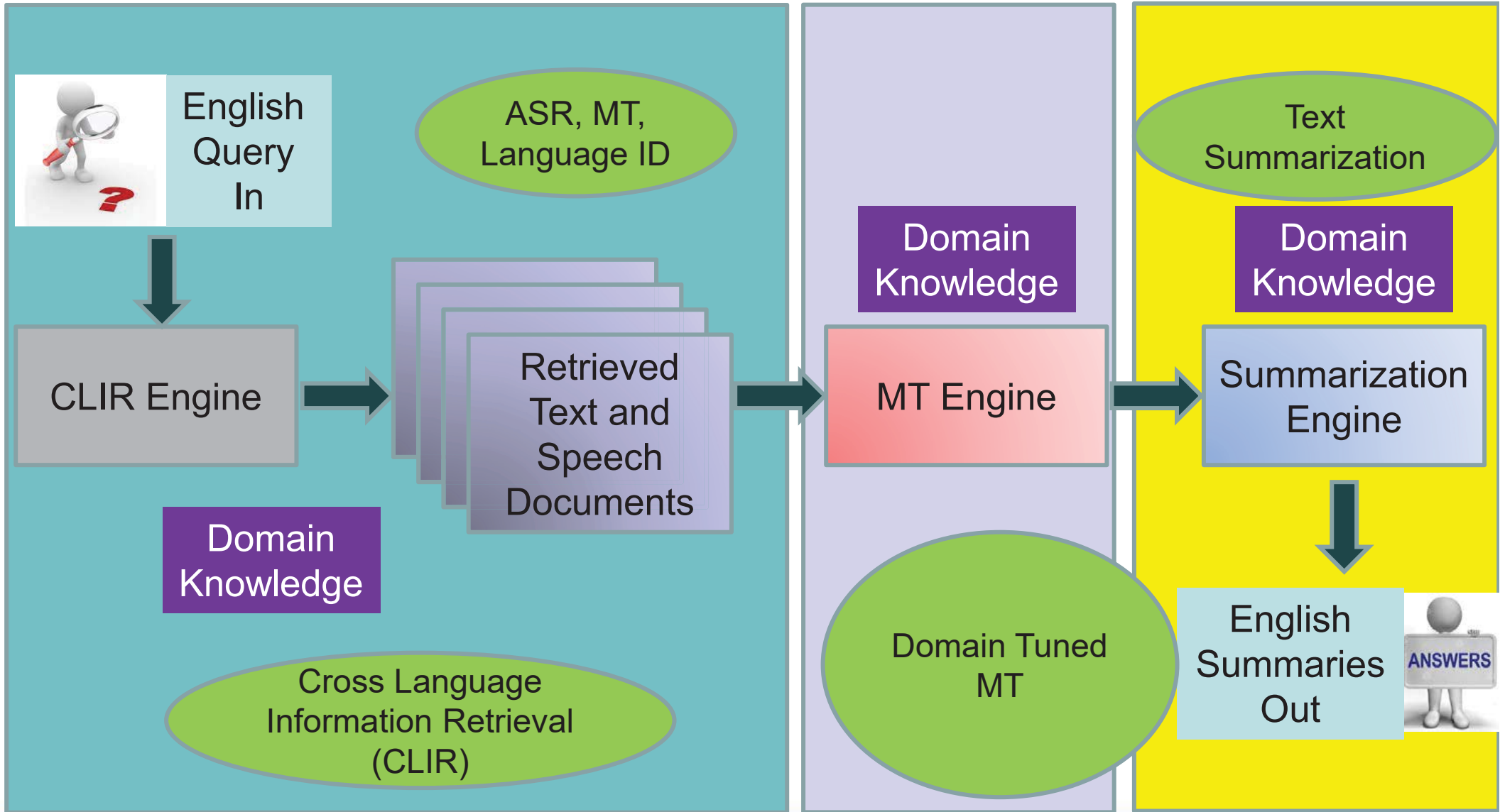
 “polio vaccine”
Domain: Government

Subject Domain

- | | | |
|---|---|---|
|  |  | ...In response, the Armenian Ministry of Health urged all Syrian Armenians under age 15 to get the polio vaccination ... |
|  |  | ...Severe adverse reactions to this vaccine are rare.... |
|  |  | ...The oral vaccine was made by weakening the three strains of poliovirus that caused disease by growing them in monkey kidney cells... |



Technology Areas in Notional End-to-End System





Key Technical Challenges

- Techniques appropriate for a wide variety of languages
- Performance on formal and informal text and speech
- Development of new methods for domain adaptation without monolingual or parallel training data in that domain
- Limited development time
- Inter-language domain mismatches reflecting a cultural component

MATERIAL will emphasize minimizing training data needs for Automatic Speech Recognition (ASR) and MT.

MATERIAL will not provide domain adaptation data or domain annotated data. Performers will develop language-independent methods that are not data intensive.



MATERIAL Training Data

- Each language will be provided at kick-off to performers in a “pack” from the IARPA T&E Team that will contain training data for MT and ASR as well as relevant language information
- Speech Training data will include:
 - 35 - 45 hours of speech data with transcriptions in a normalized orthography with additional non-transcribed speech
 - Not all genres and domains will be present in the training data
- MT Training data will include:
 - 800K word bitexts in each program language, sentence-aligned
 - Not all genres and domains will be present in the training data
- Language information will include:
 - Description of the language (e.g., dialect regions, phoneme set definitions)
 - Basic information on dialects, spelling and encoding



Domains and Genres Used for Development and Evaluation

Program data will include formal and informal varieties of text and speech, including genres that are not present in the MT or ASR training data.

Mode	% Collect	Genre
Text	~ 75	News
		Topical
		Social Media
Speech	~ 25	Broadcast News
		Topical Broadcasts
		Conversation

Domains (Broad Subject Fields) may include: Agriculture, Science, Law and Order, Military, Sports, Politics, etc.



Test Structure Rationale

- T&E regimen designed to drive R&D towards the program goal, viz:
Language independent methods, tools, and technologies to provide **rapid-deployment** of **domain-adapted** MT for **low-resource** languages effectively integrated in a usable CLIR system
- So:
 - Multiple languages with varying characteristics
 - Only small amounts of IARPA-furnished bitexts for training
 - Domain contextualized queries
 - Decreasing lead-time for development & surprise language evaluation

	Base	Option 1	Option 2
Length (months)	18	16	12
# Practice Languages	2	2	3
Surprise Language Period (months)	6	4	1.5



MATERIAL Phase Structure

- Program Level Evaluations
 - CLIR evaluation (CLE)
 - Tests ability to find query-relevant source language “documents” in the appropriate domain of the query
 - CLIR + Summary evaluation (CLE+S)
 - Tests ability to find query-relevant source language “documents” in the appropriate domain
 - Tests ability to provide summaries that help a human reader identify relevant hits and filter out remaining irrelevant results
- Post-evaluation analysis
 - Bitexts and transcriptions will be available after each testing cycle to allow performers to correlate the performance of their ASR and MT systems with the program metrics. BLEU and WER scores will be provided.



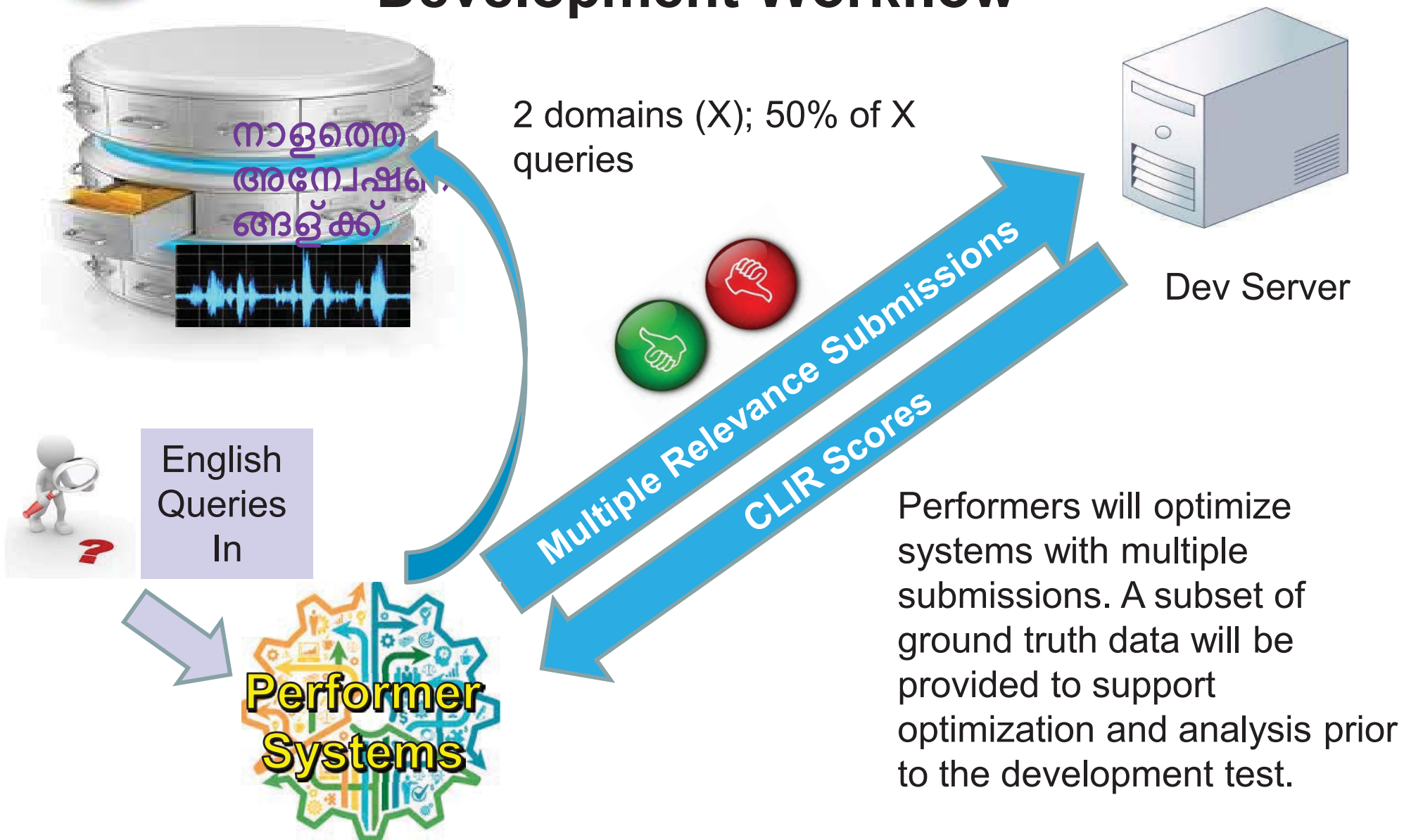
Query Release Schedule (per language)

Dataset Released	Epoch 1	Epoch 1+2	Epoch 1+2+3
# Domains	2	4	5
Query Set Introduced	X	X+Y	X+Y+Z
Practice Language Dev	50% Dev		
CLIR Evaluation		75% Dev 50% CLIR	
CLIR+Summary Evaluation			100% ALL

Performers will learn to handle new queries on new and old data, as well as old queries on new data.

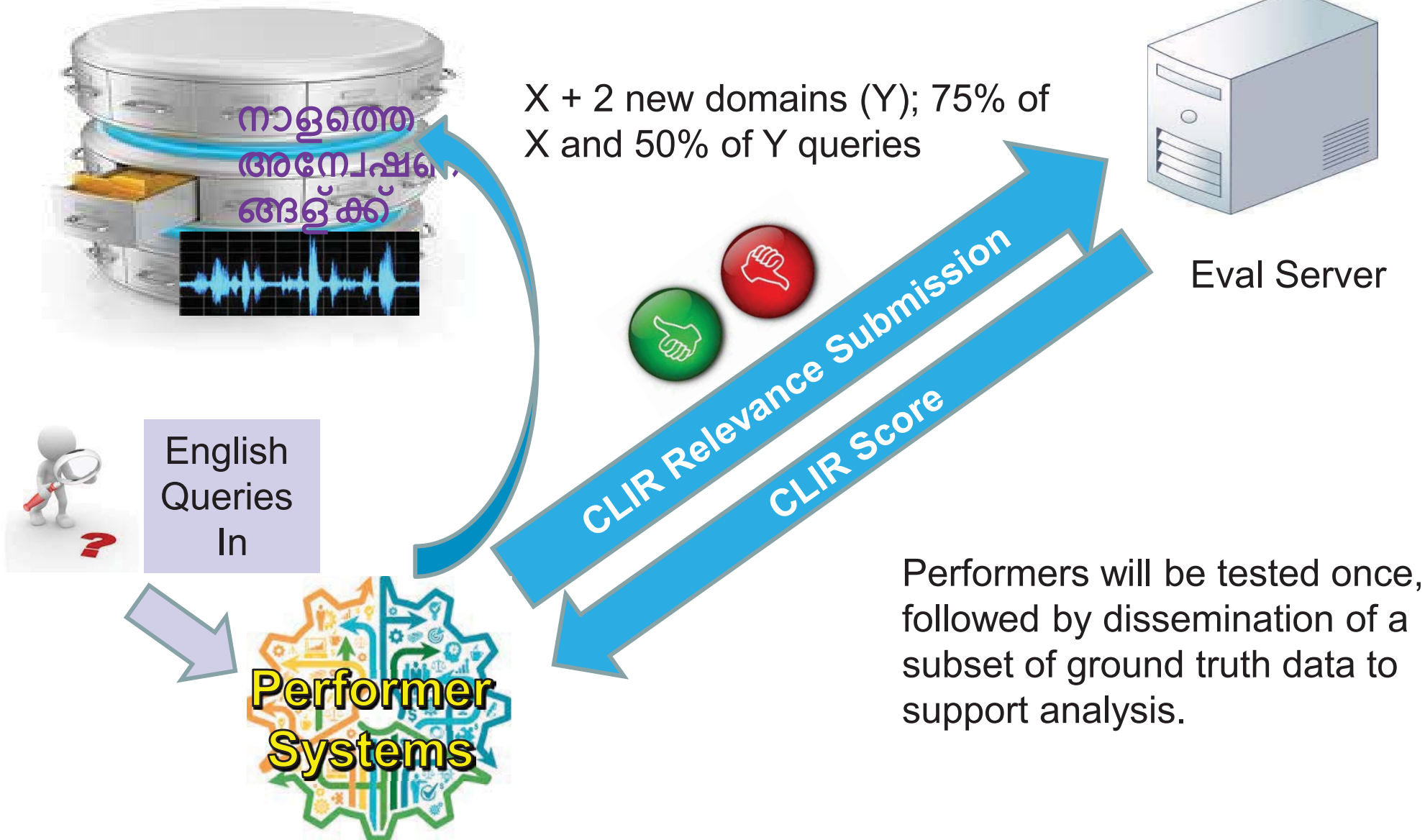


Development Workflow





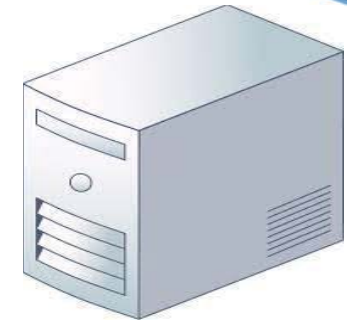
CLIR Evaluation



Performers will be tested once, followed by dissemination of a subset of ground truth data to support analysis.



Final End-to-End Evaluation CLIR+S



Eval Server

Performers will be tested once, followed by dissemination of a subset of ground truth data to support analysis.

X+Y + 1 new domain (Z); All X, Y, and Z queries



Crowdsourced Judgments



English Queries In

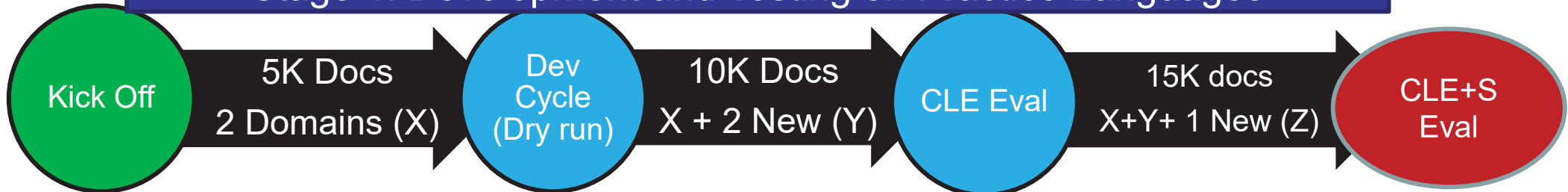




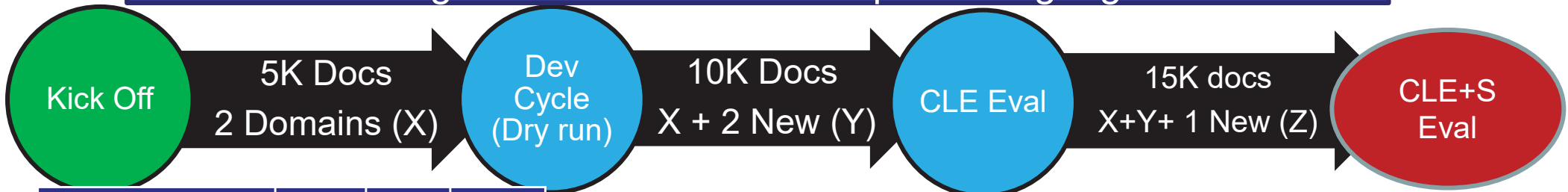
Program at a Glance

Training Data at Each Kickoff Period per language: 800K Words Bitexts; 35-45 Hours of Transcribed Audio

Stage 1: Development and Testing on Practice Languages



Stage 2: Evaluation on 1 Surprise Language



	Base	Opt 1	Opt 2
# Dev Languages	2	2	3
Phase Duration	18	16	13
Practice CLE	7	8	8
Practice CLE+S	10	10	9
Surprise CLE	13	12	10.5
Surprise CLE+S	16	14	11.5
Surprise Duration	6	4	1.5

Staging in of Queries:

Kickoff	Test 1	Test 2
50% X queries	75% X & 50% Y queries	100% X, Y & Z queries

10% responsive data translations provided at each stage for analysis that cannot be used for further training.

*Document sets and queries are reported per language



CLIR Detection Metric: AQWV Actual Query Weighted Value

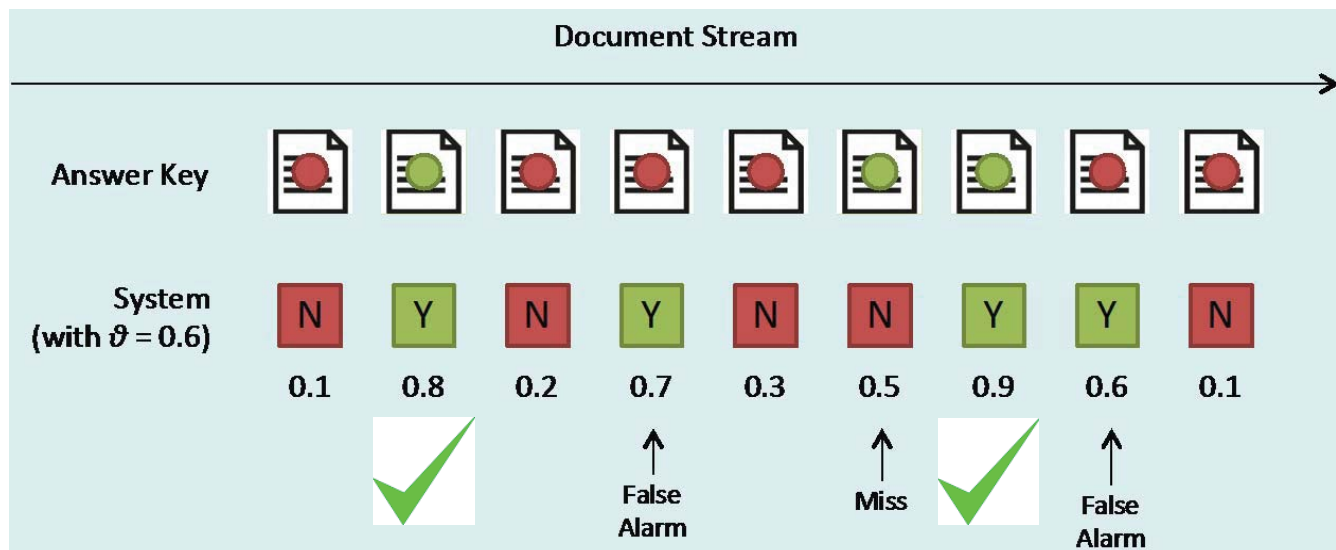
- All queries are treated equally (regardless of whether they generate single or multiple hits).
- Must be able to calibrate the metric against a baseline CLIR system that has two inputs: GOTS MT and human translation.
- Metrics will be reported to performers as an average over the set of queries (not individually for each query).
- Calculated as a representation of error rate taking into account probability of hits, false alarms, and the total number of responsive documents. These parameters will be set by T&E once the data are collected and evaluated.

Actual Query Weighted Value (AQWV)

Developers choose Θ , the detection threshold for their “Actual Decisions”, to optimize query-weighted value

- V is the *a priori* value (benefit) of a correct response
- C is the *a priori* cost of an incorrect response
- P_{rel} is the prior probability that a document is relevant to a query, e.g. 10^{-3}

$$Value_Q(\theta) = 1 - \underset{Q}{\text{average}} \left\{ p_{miss}(Q, \theta) + \frac{C}{V} (p_{rel}^{-1} - 1) \cdot p_{fa}(Q, \theta) \right\}$$



Developers will tune their systems to the threshold that maximizes the AQWV. Note that 1 is a perfect score; that is, error rate is zero.



Evaluating Summarization

CLIR Contingency Matrix

		Y	N
Key	Y	X_1 True Positive	X_2 False Negative
	N	X_3 False Positive	X_4 True Negative

$$QWV = 1 - \frac{X_2}{(X_1 + X_2)} - \beta \frac{X_3}{(X_3 + X_4)}$$

Crowd Summary Judgments

		Y	N
CLIR/ Key	Y/Y (= X_1)	A	B
	Y/N (= X_3)	C	D

CROWD-SOURCED JUDGMENTS:

Y/Y: Retrieved docs that are relevant

Y/N: Retrieved docs that are not relevant

A : # Relevant docs judged relevant

B : # Relevant docs judged non-relevant

C : # Non-relevant docs judged relevant

D : # Non-relevant docs judged non-relevant

A “perfect” summarization capability would hold B at zero and reduce C to zero



Evaluating Summarization (cont.)

CLIR Contingency Matrix

		Y	N
Key	Y	X_1 True Positive	X_2 False Negative
	N	X_3 False Positive	X_4 True Negative

$$QWV = 1 - \frac{X_2}{(X_1 + X_2)} - \beta \frac{X_3}{(X_3 + X_4)}$$

Crowd Summary Judgments

		Y	N
CLIR/ Key	Y/Y (= X_1)	A	B
	Y/N (= X_3)	C	D

A "perfect" summarization capability would hold B at zero and reduce C to zero

Summarization can reduce the false alarm rate ($C \leq X_3$) but cannot reduce the number of missed detections ($B \geq 0$)

End to End Contingency Matrix

		Y	N
Key	Y	A	$X_2 + B$
	N	C	$X_4 + D$

$$QWV = 1 - \frac{X_2 + B}{(A + X_2 + B)} - \beta \frac{C}{(C + X_4 + D)}$$



Summary

- Broad language portfolio:
 - Languages from a variety of language families (e.g., Afro-Asiatic, Niger-Congo, Sino-Tibetan, Austronesian, Altaic)
 - Mixed language typology (i.e., with different phonotactic, morphological, syntactic characteristics)
- Researchers will:
 - work with development languages to create new methods
 - be evaluated annually on a surprise language with development time and training data size constraints
- Evaluation:
 - On the set of development languages and the surprise language
 - Progress will be measured for:
 - [CLIR+S: AQWV](#) (Actual Query Weighted Value) metric and crowd sourced assessment of perceived relevance of delivered summaries
 - [MT, ASR](#): BLEU and WER scores for correlational analysis



Program Roles and Responsibilities

- **Performer R&D**
 - In Scope:
 - Novel methods for developing and adapting CLIR and MT of multi-lingual, multi-genre, multi-domain documents
 - Novel methods for developing summarization methods in English of retrieved documents for display
 - Novel use of machine learning, data resource gathering, and linguistics
 - Computational methods to reduce running time and memory footprint of models
 - Out of Scope:
 - Human User Interface
 - Image data
- **Government Support**
 - Government Furnished Information (GFI):
 - Acquire, organize and disseminate training data
 - Prepare and disseminate development, evaluation, and analysis data
 - Testing and Evaluation:
 - Evaluation framework to measure performer progress on practice and surprise languages
 - Development data to measure interim progress



Eligibility Information

- Other Government Agencies, Federally Funded Research and Development Centers (FFRDCs), University Affiliated Research Centers (UARCs), and any other similar type of organization that has a special relationship with the Government, that gives them access to privileged and/or proprietary information or access to Government equipment or real property, are not eligible to submit proposals under this BAA or participate as team members under proposals submitted by eligible entities.
- Non-US organizations and individuals may be able to participate.
 - Must comply with Non-Disclosure Agreements, Security Regulations, Export Control Laws, etc, as appropriate
 - Specific guidance for non-US participation will be provided in the BAA



Proposal Guidance

- Your proposal should include a full discussion of the technical approach that will be used to meet the program goals.
- Programmatic issues to be addressed in the proposal:
 - Your team's current technical capabilities
 - Key resources needed (not currently available to your team), to include capital equipment and special expertise (teaming will likely play an essential role in providing special expertise). The risk in acquiring these key resources, and mitigation strategies, should be indicated as well.
 - A teaming plan along with the roles and responsibilities of each member of the research team
 - End of phase and some intermediate milestones are set, but it is expected that other intermediate milestones that are on the critical path of the proposed approach will be offered.
 - A schedule of all milestones including a clearly charted description of the various risk mitigation strategies that will be undertaken to achieve program goals



Proposal Evaluation Criteria

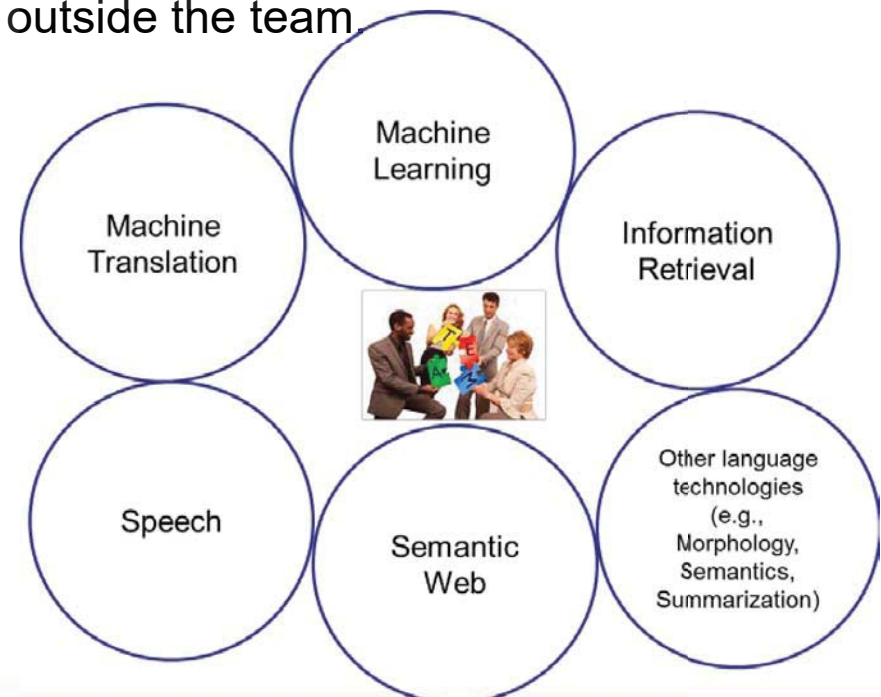
- Overall Scientific and Technical Merit
- Effectiveness of Proposed Work Plan
- Relevance to IARPA Mission and MATERIAL Program Goals
- Relevant Experience and Expertise
- Cost Realism

Evaluation criteria will appear in the BAA.

Teaming

- Because of the many challenges presented by this program, both depth and diversity will be strongly encouraged for overcoming these challenges.
 - Throughput: Consider all that you will need to do, all the ideas you will need to test. Make sure you have:
 - Enough people and expertise to do the job
 - Sufficient resources to follow critical path while still exploring alternatives – risk mitigation
 - Completeness: teams should not lack any capability necessary for success, e.g. should not rely on enabling technology to be developed outside the team.
 - Tightly knit teams
 - Clear, strong, management, single point of contact
 - No loose confederations
 - Each team member should be contributing significantly to the program goals.
Explain why each member is important, i.e. if you didn't have them, what wouldn't get done?

Can your team complete an evaluation well in the time frame required?





Additional Information

- Email dni-iarpa-baa-16-11@iarpa.gov with additional questions.
- MATERIAL BAA will be posted on FedBizOpps website (www.fedbizopps.gov).
- Q&As will appear after the BAA is posted. See http://www.iarpa.gov/solicitations_material.html.



Questions?

A Taxonomy of Weeds: A Field Guide for Corpus Curators to Winnowing the Parallel Text Harvest

Katherine M. Young[†]
N-Space Analysis, LLC

katherine.young.ctr.1@us.af.mil

Jeremy Gwinnup
Air Force Research Laboratory

jeremy.gwinnup.1@us.af.mil

Lane O.B. Schwartz
University of Illinois

lanes@illinois.edu

Abstract

Modern machine translation techniques rely heavily on parallel corpora, which are commonly harvested from the web. Such harvested corpora commonly exhibit problems in encoding, language identification, sentence alignment, and transliteration. Just as agricultural harvests must be threshed and winnowed to separate grain from chaff, electronic harvests should be carefully processed to ensure the quality and usability of the resulting corpora. In this work, we catalog a taxonomy of problems commonly found in harvested parallel corpora, and outline approaches for detecting and correcting these problems.

This work is motivated by the lack of a standardized field guide outlining best practices for curating parallel corpora, especially those harvested from the web. Even the most-well curated parallel corpus is likely to contain some problems; even Europarl (Koehn, 2005), arguably the most widely examined parallel corpus, has undergone eight distinct revisions since its release in 2005. While this work is by no means comprehensive of all problems extant in corpus creation and curation, we nevertheless believe that a practical taxonomic field guide, laying out likely pitfalls awaiting corpus curators will represent an important contribution to our community.

1 Introduction

Statistical machine translation typically requires large amounts of translated parallel text to serve as training data for statistical translation models. End-users of machine translation may use in-house data developed from years of prior human translation efforts (Plitt and Masselot, 2010; Hellstern and Marciano, 2014). A perhaps more common practice, developed over the past fifteen years (Resnik, 1998), involves the automatic harvest of parallel corpora from online

[†]This work is sponsored by the Air Force Research Laboratory under Air Force contract FA-8650-09-D-6939-029.

resources, such as bilingual web sites (Smith et al., 2013) or the crowd-sourced translations of the TED Talk transcripts (Cettolo et al., 2012).

Just as agricultural harvests must be threshed and winnowed to separate grain from chaff, electronic harvests may be carefully processed to ensure the quality and usability of the resulting corpora. Simard (2014) suggested the metaphor of weeds choking out cultivated plants to be more apropos than that of cleaning “dirt” from corpora. We adopt this terminology, identifying a broad variety of such *weeds* found growing wild in online data, potentially degrading the quality of harvested corpora. In keeping with this botanic metaphor, we use *zizania*, a Greek term for a type of weed that grows intermixed with wheat,¹ as a basis for our taxonomic nomenclature.

In this work, we present a taxonomy of weeds commonly found in harvested parallel corpora, and outline approaches for detecting and correcting these problems. At the highest rank, the taxa we present are categorized based on provenance: Do the errors originate from problems during automatic processing of the text (*zizania ex machina*) or from human failure (*zizania ex homine*)? We categorize six major types of the former (§2.1–2.6), as well as six major types of the latter (§3.1–3.6). Throughout this work, we consider weeds that have been previously identified in the established literature, as well as weeds that we have encountered that have not heretofore been described in the literature.

This work is motivated by the lack of a standardized field guide outlining best practices for curating parallel corpora, especially those harvested from the web. Even the most-well curated parallel corpus is likely to contain some weeds; even Europarl (Koehn, 2005), arguably the most widely examined parallel corpus, has undergone eight distinct revisions since its release in 2005. We believe that a practical taxonomic field guide, laying out likely pitfalls awaiting corpus curators will represent an important contribution to our community.

2 *Zizania ex machina*: Weeds of mechanical origin

We now survey various *zizania ex machina*: weeds that originate during automated corpus processing.

2.1 Wrong Language Text

Wrong-language text errors can occur during automatic collection of parallel text from websites. The scraping program may mis-identify similar languages, or the program may fail to notice a section of foreign text within a page produced in the correct language. For example, if the program is scraping an English-language site with hotel reviews, it may pick up some reviews written in French. Alternatively, the program may fail to exclude a section of text that has remained untranslated across pages of a multilingual site. These failures create two types of errors that can be automatically detected, *Source-Source* errors, and *Source-Other* errors.

¹See, for example, the usage of *zizania* in the Greek New Testament (Matthew 13:25).

2.1.1 Source-Source instead of Source-Target

An example of Source-Source error occurred in the initial release of the IWSLT 2014 data (Cettolo et al., 2012), in which some of the parallel English-French text was provided untranslated, creating English-English data. This was subsequently corrected. Source-Source errors can be detected automatically by searching for sentences that are duplicated across parallel text; these are usually untranslated sections. Short duplicate sentences should be examined separately, since there can be some legitimate duplication if the text contains URLs, named entities, borrowed words, or quotations. Legitimate duplication at the token level can also be caused by cognates (for example, the English word *importance* matches French *importance*).

2.1.2 Source-Other instead of Source-Target

Examples of Source-Other errors can be found in the French side of the 10⁹ English-French corpus (Callison-Burch et al., 2009), in which we find paragraphs in Greek, Russian, German, and other languages. Such Source-Other errors can be detected easily if the incorrect language has a different character set than the correct language. For example, a section of Greek within a supposedly French document can be easily filtered out by specifying a desired range of permitted Unicode code points.

For languages with similar alphabets, we apply a simple dictionary-based program to remove sentences with a majority of unknown words. Recent work (Zampieri et al., 2014; Lui et al., 2014) leverages character n-grams, POS sequences, and other features to train language discrimination systems for similar languages.

Depending on the application, thresholding may be desired to allow a specified amount of wrong-language text (for foreign names, borrowed words, quotations, etc.). On the other hand, web-scraped text from multi-lingual sites often contains isolated wrong-language phrases that we may want to remove, such as hyperlinks in multiple languages. Multi-lingual sites can also contain stock phrases like “Click here to login” that may remain untranslated across the site; these might also need to be removed.

2.1.3 An illustration of a specific language identification clean-up process

For languages with similar but not identical alphabets, detection programs can be written that are specific to that language pair. For example, the English-Russian Common Crawl data includes sections which are actually English-Ukrainian. Ukrainian has four characters not found in Russian which can be used to identify unwanted Ukrainian segments: UKRAINIAN I (і І), YI(ї Ї), GHE WITH UPTURN (ґ Г) or IE (є Є). We make an exception to allow UKRAINIAN I in Russian segments when it occurs in a potential context for a Roman numeral (adjacent to Latin X, I, V, x, i, v, or their Cyrillic counterparts).

Second, on the English side of the Russian-English Common Crawl, we find sections of text in other languages such as French. Both English and French use the Latin character set, but French uses special characters not typically found in English such as à é ê î ô œ ç; these could be used to identify the presence of French, with some proportion of exceptions allowed for borrowed words like *café*. However, for the Common Crawl we also want to detect other non-English languages like Spanish. Instead of relying on specific accented characters to detect

Experiment	Corpus Size	Filtered Corpus Size	Avg. Cased BLEU	Avg. Uncased BLEU
Baseline	878386	732129	25.39	26.59
Cleaned	772530	642746	25.73	26.95

Table 1: Before and After Common Crawl experiment results reported in BLEU

non-English text, we apply a spell checker to identify English text. We use the `aspell`² spell-checker to determine the proportion of words that are not recognized as English, and compare this to a set threshold to identify the wrong-language sections. We exclude from consideration words of 3 characters or less, because many short words have false friends in other languages (e.g., *die* in English and German, *on* in English and French).

We demonstrate the effectiveness of these techniques by taking a baseline WMT15 MT system and replacing the phrase and lexicalized reordering tables with ones generated from the Common Crawl corpus in both original and cleaned configurations. Table 1 shows the cleaned corpus yields a +0.34 BLEU improvement over the non-processed baseline even with a 12% reduction in corpus size.

2.2 Historical Encoding Errors

Portions of a corpus are sometimes encoded using a different character encoding scheme than the rest of the document. If not detected and corrected, this leads to an encoding cipher, where sentences appear shifted to an incorrect character range. Encoding errors of this type can also occur when extracting text from a PDF document.

In the Russian-English Common Crawl parallel corpus, a number of Russian source sentences are encoded using the 8-bit Windows-1251 character encoding scheme. Most sentences in this corpus are encoded using UTF-8; when Windows-1251 encoded sentences are interpreted as UTF-8, the Cyrillic characters incorrectly appear as characters from the Latin-1 supplement block. This can be corrected by shifting these characters ahead by `350hex` code points into the correct Unicode Cyrillic character range. An example of this code point shift is shown in Figure 1 below:

- (a) Справка по городам России и мира.
- (b) Ńřđääêà ř āřđřääì Đřññèè è ìèđà.

Figure 1: Russian sentence (a) originally encoded as Windows-1251, interpreted as UTF-8 (b)

Encoding errors may also show up in isolated characters. We see this in some of the Common Crawl data, in which French accented characters have been converted to Cyrillic characters. For example, we find the words *équipe* and *château* written as *ëquipe* and *château*. This is the reverse of the Russian code point shift described above, and these errors can also be corrected automatically if we know that the Cyrillic characters are out of range for our text. The Common Crawl exhibits a variety of code point encoding problems in addition to those shown here. Out of range characters should be examined for code point shifts and encoding problems that could possibly be corrected.

²<http://www.aspell.net>

Lang.	Set	Sentences w. repeat errors	Total sentences
French	dev2010	11	887
	tst2010	87	887
Chinese	tst2010	81	1570
	tst2014	13	1068
Farsi	tst2010	1	885
	tst2011	22	1132
	tst2012	343	1375
	tst2013	187	923
	tst2014	53	1131

Table 2: Number of sentences containing segment-internal repetition errors in IWSLT dev and test sets

Lang.	Year	Sentences w. repeat errors	Total sentences
Arabic	2013	3	155,047
	2014	5	186,467
Chinese	2014	550	177,901
Farsi	2013	5,749	81,872
	2014	8,987	112,704
French	2013	173	162,681
	2014	373	186,510
Russian	2013	109	135,669
	2014	145	185,205

Table 3: Number of sentences containing segment-internal repetition errors in IWSLT training sets

There can also be encoding problems with individual characters. A confusion between UTF-8 encoding and Windows-1252 encoding can lead to a single character such as 0xE28099 (') being interpreted as multiple, single-byte characters: 0xE2 (â), 0x80 (€) and 0x99 (™) Notenbloom (2009). These single-byte/multiple-byte encoding errors can be corrected programmatically with existing tools.

Finally, we note that the character U+FEFF may appear in some files as the residue of a byte order marker at the start of a file; this should be deleted to avoid confusion with the Arabic script character U+FEFF, which is a zero-width non-breaking space.

2.3 Bidirectional Reversal

Adobe's Portable Display Format (PDF) is meant as a display format and does not focus on the orderly layout of data in the document's container. This presents issues when extracting text in an orderly fashion from PDF documents. Extraction issues are compounded when dealing with custom fonts and historical encoding schemes. Additional issues involve the display of Right-to-Left (RTL) text.

Sometimes, extraction of RTL text from PDF creates text in which the line is reversed, character-by-character. We can detect reversals automatically by checking the words against a dictionary or word-frequency list to derive a percentage of unknown words. We then compare that percent unknown against the typical score for text from that language. If the percent unknown is suspiciously high, we can use a program to character-reverse the line, and repeat the dictionary check; a better score on the reversed line confirms the reversal error. In correcting reversed lines, we need to be careful how we handle digits, which run left-to-right within right-to-left text in many Arabic-script languages.

2.4 Automatic sentence alignment errors

When parallel sentences are aligned, typically via automated means, mistakes in sentence alignment lead to mis-aligned sentence pairs that do not represent mutual translations. Many parallel corpora are naturally aligned at the document level: A human translator translates a source document into a target language. However, most statistical methods that make use of parallel data require alignment at the sentence level, and automated sentence aligners may make errors.

Various automated techniques have been proposed to minimize the problem of mis-aligned sentences. Gale and Church (1991) proposed an automated length-based sentence alignment technique that compared the number of words in source and target sentences. Proposed extensions to length-based approaches include the use of cognate frequency (Simard et al., 1992) or other lexical cues (Wu, 1994). Structural tags (such as HTML elements) have also been proposed as an aid to guide sentence alignment (Resnik, 1998).

2.5 Segment-Internal Repetition and Chunking Errors

Processing errors may cause a sentence or sub-sentential fragment to be improperly duplicated within a given line. In many cases, such repetition can be automatically detected and corrected; examination of the corresponding parallel sentence can assist in this process.

The IWSLT 2014 data, for example, contain substantial cases of repetition errors, especially for certain language pairs (see Tables 2 and 3 on the preceding page). An example of a repetition error is shown in Figure 2 below:

Last year I showed these two slides so that demonstrate that the arctic ice cap, <i>which for most of the last three million years has been the size of the lower 48 states</i> , has shrunk by 40 percent.
L'année dernière, je vous ai présenté ces deux diapositives qui montraient que la calotte glacière arctique, <i>qui pendant ces 3 derniers millions d'année avait la taille des Etats-Unis sans l'Alaska</i> , <i>qui pendant ces 3 derniers millions d'année avait la taille des Etats-Unis sans l'Alaska</i> , avait diminué de 40%.

Figure 2: Example of repeated phrase in English-French TED data. Within the French sentence, the words in ***bold italics*** represent an erroneous copy of the words in *italics*.

While some cases of repetition are not errors (the TED Talks in particular may contain repetition for rhetorical effect), the presence of high amounts of repetition errors in training data and development data can degrade machine translation quality; correcting the large number of

repetition errors in the IWSLT 2014 Farsi test file improved Farsi-to-English performance by +1.53 BLEU.

Chunking errors occur when sub-sentential segments are automatically combined without the necessary spacing. For example, a small number of files in the QED Corpus provided to the IWSLT 2016 competition (Abdelali et al., 2014) exhibit a chunking error, in which each line has run-together words in the middle of the line (see Figure 3). This is probably an error in assembly. The QED Corpus derives from the AMARA website, which enables crowd-sourced transcription of video; the AMARA interface presents the worker with 4-second segments of video to transcribe, and these are subsequently assembled into a larger text (Zukerman, 2013). We found 57 files with mid-line chunking, out of 19K total English files.

Chunking errors create unknown words for machine translation. A human looking at these files can analyze the problem easily, based on what is reasonable to expect in the sentence, but automatic, rule-based correction faces some difficulties. A spell checker like `aspell` can be applied to detect and correct run-together words, but we have to protect named entities and technical terms which may not appear in `aspell`'s dictionary. We also have to be careful to split the words in the correct place. Initially, we simply split the unknown word into progressively longer sections of first word vs. second word, until we found two known words. This led to unfortunate splits like *thoughtsand* > *thought sand* instead of *thoughts and* and *monkeysin* > *monkey sin* instead of *monkeys in*. A word frequency list could be applied to select the best split. Alternatively, language modeling could determine which split creates the most reasonable sentence.

It's the difference between divergent thinkingand convergent thinking. You have to separate the two so that you can diverge your thoughtsand come up with this great collection of ideas, and then once you have this great collectionof ideas, you focus on the convergent thinking.
--

Figure 3: An example of chunking errors in the QED Corpus.

2.6 Harvested Machine Translations

When parallel corpora are harvested from the web, there is a danger that some of the parallel content was created by means of machine translation, rather than human translation. Attempts have been made to automatically identify machine-translated content using various machine learning techniques, including decision tree classifiers (Corston-Oliver et al., 2001), SVM classifiers (Gamon et al., 2005), maximum entropy classifiers (Rarrick et al., 2011), watermarking (Venugopal et al., 2011), and identifying the presence of characteristic MT errors (Antonova and Misyurev, 2011). The extent to which the inclusion of machine-translated content in MT training data harms translation quality of the trained system may depend largely on the quality of the harvested machine translations (Simard, 2014).

3 *Zizania ex homine*: Weeds of human origin

In this section we survey weeds of human origin that show up in translated text from online sources. In general, *zizania ex homine* are harder to correct than *zizania ex machina*, but some

automatic correction is possible.

3.1 Mixed Alphabets

Words with mixed alphabets visually resemble correctly spelled words, but are treated as separate tokens in the machine translation process. Such words can be automatically detected and corrected, converting characters to the majority alphabet for that word when they have visually similar counterparts.

Word	Latin (L) or Cyrillic (C)	Meaning
она	LCL	she
сейчас	LCCCCC	now
MP3-плеер	LLL-CCCCC	MP3-player
MP3плеер	LLLCCCCC	MP3player
амазон.com	CCCCCC.LLL	amazon.com
ипациент	LCCCCCCC	iPatient

Figure 4: Examples of Mixed-Alphabet words. In the center column, we annotate each character of the corresponding Russian word as either Latin (L) or Cyrillic (C). For example, in the first row, the Russian word она is encoded such that the Latin characters *o* and *a* are used instead of the more appropriate (but visually indistinguishable) Cyrillic equivalents.

We have encountered mixed alphabet words in the Russian sections of the Russian-English Common Crawl and in the Russian transcriptions of TED Talks. This occurs when the translator uses a combination of Latin and Cyrillic characters to write a Russian word. The reason for these mixed spellings is unknown; perhaps it is due to limitations of the translator’s input method, or perhaps it is influenced by typing both English and Russian words. For example, although the first letter and last letter in the word сейчас appear visually indistinguishable, in this instance we find that the former is U+0063 LATIN SMALL LETTER C and the latter is U+0441 CYRILLIC SMALL LETTER ES. We even find the Russian word она written with U+006F LATIN SMALL LETTER O and U+0061 LATIN SMALL LETTER A instead of the appropriate Cyrillic counterparts (U+043E and U+0430); this word is harder to correct, since the majority favors the wrong alphabet.

Some mixed alphabet spellings are deliberate, combining a borrowed English word with a Russian word. Figure 4 above shows examples of this behavior. Converting punctuated words on a part-by-part basis can protect some but not all of these deliberate mixed spellings from automatic conversion.

In addition to the mixed alphabet spellings in Russian, we find creative spellings in many languages that borrow from other character sets, or repurpose characters within the source alphabet, particularly for punctuation. Some examples are given in Figure 5 on the next page. Determining how to correct such creative spellings generally requires human intervention.

3.2 Mixed Morphology

When a translator brings in a borrowed word through transliteration, he or she may choose to inflect the borrowed word using target language morphology. For example, in Urdu text we

Language	Character Written		Character Intended	
Urdu	U+002D -	LATIN HYPHEN	U+06D4 .	URDU FULL STOP
French	U+00A8 ¨	LATIN DIAERESIS	U+0022 "	LATIN QUOTATION MARK
Russian	U+0431 б	CYRILLIC SMALL LETTER BE	U+0036 6	LATIN DIGIT SIX
English	U+006F o	LATIN SMALL LETTER O	U+00B0 °	LATIN DEGREE SIGN

Figure 5: Examples of Creative Spelling.

find the borrowed English word *leader* with the plural suffix */-wn/*, creating *لیڈروں* /lydrwɒ/, as well as the borrowed word with the original English plural form (*leaders*), *لیڈرز* /lydrz/. Names in particular are subject to variation in the application of target language morphology. An examination of names borrowed into Russian from English in the TED Talk data showed this range of behavior: a) first and last name both uninflected, b) first and last name both inflected, c) last name only inflected. Examples are shown in Figure 6; all three examples are possessive structures which should occur with genitive case.

Russian Text	Phonemes	English Text	Annotation Type
песню Уитни Хьюстон	/uitni x'yuston/	a Whitney Houston song	a) neither name inflected
закон Артура Кларка	/artur+a klark+a/	Arthur Clarke's law	b) both names in genitive case
Книга Эл Гора	/ɛl gor+a/	The Al Gore book	c) last name in genitive case

Figure 6: Examples of Mixed Morphology.

Inflected borrowed words often show up as out-of-vocabulary (OOV) words in MT output. If OOV words are going to be transliterated (see §3.3), it is useful to first apply a stemmer to remove any inflectional endings. Lexical approximation can sometimes rehabilitate inflected borrowed words and allow them to be translated (Mermer et al., 2007). Alternatively, Schwartz et al. (2014) identify inflected OOV words at the start of the decoding process, and replace them with variant inflected forms from the phrase table.

3.3 Transliteration of Names and Borrowings

Borrowed words and names may occur in transliteration, with the original sounds mapped into the characters of the new language. While such coinages are not errors, they are subject to variation that creates problems when an MT system attempts to relate them to the original forms.

Statistical methods may be applied to deal with this variation in transliteration, as for example in Durrani et al. (2014). Our work focuses instead on improving rule-based transliteration, which maps characters into their typical sound values. Because there can be variation in the character-to-sound mapping in both languages, the output of rule-based transliteration is often faulty. This output can be improved by constraining the results to actual English spellings. In particular, we address the recovery of named entities that persist as OOV words in the output of machine translation. We describe two ways to constrain the results of named entity (NE) transliteration into English, one using an English pronunciation dictionary, and another using parallel training data to create a transliteration-based map of NE pairs.

3.3.1 Recovering Names via Transliteration in Conjunction with an English Pronunciation Dictionary

Rule-based transliteration can be improved by leveraging a target language pronunciation dictionary. We adapt the CMU English pronunciation dictionary³ to guide transliteration from Russian into English. Because vowel spellings may be variable, we create a fall-back representation for each word in which all vowels are converted to a placeholder character, @. We derive a word frequency count from the training data and record the frequency count for each dictionary entry. We also supplement the pronunciation dictionary by noting any words in the WMT 2014 Russian data (Bojar et al., 2014) that are not listed in the CMU dictionary, and deriving their phonetic forms via *Sonic* (Pellom and Hacıoglu, 2001).

When we run our transliteration program, we first map the Cyrillic characters into their typical sounds, recording multiple possibilities where appropriate. Next, we compare these phonetic mappings to the phonetic entries in the English pronunciation dictionary. We try to find words which match the sound pattern for both consonants and vowels; failing that, we use the vowel placeholder representations and allow @ to match any vowel or sequence of vowels. If there are multiple candidate words, we select the word with the highest word frequency count. We output the English spelling of the chosen word.

3.3.2 Recovering Names via a List of Transliterated NE Pairs

We apply transliteration and NE tagging to create a list of NE pairs from parallel Russian-English text; this list can subsequently be used to either pre-translate NEs, or to recover OOV names in the MT output. First, we apply the *mystem*⁴ morphological analyzer to tag NE in the Russian text. For each NE, we then use rule-based transliteration to get a phonetic form, from which we identify possible matches in the English sentence. We record the best match along with the Levenshtein edit distance between the phonetic form and the English spelling, normalized for word length. NE pairs with a distance score below 0.66 are stored in a NE list that can be used to translate Russian NEs. When applied to the Russian-English WMT 2014 training data, this method generated a list of 216K potential NE pairs.

3.3.3 Third Language Mappings

Automatic transliteration processes can stumble when dealing with words that derive from languages other than the source or target. In English, for example, the letter *j* usually indicates the affricate sound [dʒ], but in words of Spanish origin, it may represent [h]. This presence of a third-language sound pattern complicates the use of transliteration. Hagiwara and Sekine (2011) and Li et al. (2007) suggest ways to detect alternate languages in statistical transliteration: Li et al. (2007) train with language-tagged word pairs; Hagiwara and Sekine (2011) introduce latent classes to model language origins. For rule-based transliteration, developing programs to detect and correct such third-language spelling differences requires examination of the sound patterns of the various languages; human intervention may be required to decide when to apply the alternate mappings.

Russian provides a particular problem for transliteration due to the presence of third-

³<http://www.speech.cs.cmu.edu/cgi-bin/cmudict>

⁴<https://api.yandex.ru/mystem/>

language sound patterns from Chinese. When referring to Chinese names in Russian texts, Russian writers follow the Palladius mapping (Palladius and Popov, 1888) to transliterate Chinese names into Cyrillic. Many Cyrillic characters generated by this mapping represent different sounds than those Cyrillic characters typically represent in Russian. For example, the Cyrillic character ж typically represents /zh/, but in the Palladius mapping it represents /r/, and the combination of characters Чж is used to represent /zh/. Figure 7 illustrates how applying the typical Russian-to-English transliteration for OOV Russian words will cause errors for Chinese names, unless we first reverse this Palladius mapping (Young et al., 2012).

- (a) 翟志刚
- (b) Чжай Чжиган
- (c) Chzhay Chzhigan
- (d) Zhai Zhigang

Figure 7: Chinese name (a), with transliterations into Cyrillic (b) and Latin using normal Cyrillic-to-Latin transliteration (c) and reverse Palladius transliteration (d). The output in (d) is correct.

3.4 Under-achieving Translation

We use the phrase *under-achieving translation* to designate weeds that result from a lack of attention by the human translator. Sometimes translators leave a word untranslated; this kind of error can be detected by methods discussed above in §2.1, including the detection of out-of-range characters if the languages have different alphabets. More subtle weeds can occur when the translator chooses transliteration in place of translation, as the appropriateness of transliteration depends on context.

3.4.1 Transliteration in Place of Translation

Sometimes the human translator simply transliterates the source word, even when an appropriate translation exists in the target language. This may represent a translator’s decision to preserve the original form in a named entity, or it may reflect a careless translation. For example, the English word *review* has various Russian translations, such as журнал (review, journal) and рецензия (review, critique). However, in the IWSLT 2014 training data we find *review* transliterated in the phrase, *Harvard Business Review*, Гарвард Бизнес Ревью /garvard biznes rev’ju/. This choice preserves the title of the publication; translating *review* to журнал /zurnal/ would have introduced confusion with the English word *journal*. In the Common Crawl, on the other hand, we find an inappropriate transliteration of *review*, in the phrase *Awards and Reviews*, which becomes Награды и Ревью /nagrada i rev’ju/. This instance should probably have been translated. For unfamiliar words, a translator may resort to letter-by-letter spelling, as in the Russian spelling опоссум for *opossum*, which reflects the English spelling rather than the pronunciation [pasəm] or [əpasəm]. The coexistence of translation, sound-based transliteration, and letter-based transliteration creates more variation that must be addressed in machine translation.

3.4.2 Code-switching

The use of transliterated foreign words may also be driven by a form of code-switching (Myers-Scotton, 1993; Diab et al., 2014, 2016) in which the writer deliberately uses foreign words. For example, Urdu writers frequently use transliterated English words, instead of their Urdu counterparts, because the use of English exhibits a level of prestige (Upal, 2008). Hence, we may find transliterated English words in Urdu source text, as well as in Urdu text that has been translated from English. In Table 4, the Urdu writer has used English words in transliteration for four words, in place of using the Urdu words. Such transliterations complicate the machine translation of Urdu by creating variations between transliterated English words and the actual Urdu words.

English	In the top ten, India comes in the last
Urdu	اس سرٹیفکیشن کی ٹاپ ٹین میں بھارت آخری نمبر پر ہے۔
Transliteration	as srtyfkyšn ky tap tyn myn bhart Ajry nmbr pr byn
English words	– certification – top ten – – – number – –

Table 4: Urdu transliteration example. In this example, the author of the Urdu sentence used four English words (transliterating *certification* into *srtyfkyšn*, *top* into *tap*, *ten* into *tyn*, and *number* into *nmbr*) instead of using the corresponding Urdu words.

3.5 Over-achieving Translation (Explication)

Human translators intend to communicate meaning, and so may depart from the source text in ways that improve understanding, but degrade the usefulness of the translation as parallel text. Translators may expand acronyms, add explanation of localized vocabulary, or include the actual source-language words. Translators working on informal speech may remove false starts and clean up awkward sentence structure.

We term this type of explication (Blum-Kulka, 1986) *over-achieving translation*, in contrast to the *under-achieving translation* of the previous section. This type of extra information is difficult to detect and modify for machine translation. If the translator has set off added material in brackets or parentheses this can be detected, but often the additional material is integrated into the translation.

The TED Talks suffer from particular problems with over-achieving translation, since they are spoken presentations supplemented by visual aids. The English transcriptions tend to follow the speaker closely, while the translations often clean up disfluency.⁵ If text appears on the slides, the translators often include a translation of this material in the transcript.

Similarly, sentence alignment problems can also be caused when human translators summarize (Khadivi and Ney, 2005), engage in one-to-many, many-to-one, or many-to-many sentence translations (Gale and Church, 1991), or engage in non-literal free translations (Imamura and Sumita, 2002);⁶ the resulting parallel sentences may be less useful from the perspective of

⁵Cho et al. (2014) suggest handling this issue by tightly integrating disfluency removal into the MT decoding process.

⁶Imamura and Sumita (2002) also identify as problematic to their data-driven rule-based MT technique situations where a given source phrase is translated in multiple different ways throughout the corpus. Modern statistical machine translation techniques tend to be relatively resistant to this variety of weed.

machine translation training than other more literal translation pairs. This problem may be mitigated by removing less literal translation pairs from the parallel corpus (Okita, 2009; Jiang et al., 2010), or by flagging sentence pairs which exhibit atypical length ratios for manual inspection (our tools take the latter approach).

3.6 Translation Directionality

Other researchers have noted that translated text differs in crucial ways from native text, in both general simplification (Lembersky et al., 2013) and by influence from the word order and vocabulary choice of the source language text (Fusco, 1990). Koppel and Ordan (2011) show that classifiers can be trained to distinguish the direction of translation. Translation models are typically built from parallel corpora without regard for which language of the pair is the original source language. Changing this paradigm to one where original source language is taken into account has been shown to improve translation quality (Kurokawa et al., 2009).

4 Conclusion

This work is motivated by the lack of a standardized field guide outlining best practices for curating parallel corpora, especially those harvested from the web. Even the most-well curated parallel corpus is likely to contain some problems; even Europarl (Koehn, 2005), arguably the most widely examined parallel corpus, has undergone eight distinct revisions since its release in 2005. In this work, we categorize six major types of problems that originate in automated processing of corpora, as well as six major types of problems that originate in human translator actions. In this work, we establish an initial taxonomy of weeds. While this work is by no means comprehensive of all problems extant in corpus creation, we nevertheless believe that a practical taxonomic field guide, laying out likely pitfalls awaiting corpus curators will represent an important contribution to our community.

The extent to which various types of weeds are harmful in practice is not fully established. Asia Online (2009) and others have claimed substantial positive results from weeding. Likewise, we found substantial improvement in translation quality when major repetition errors are corrected. On the other hand, Goutte et al. (2012) report that statistical MT systems may be robust to sentence alignment errors as high as 30%. In future work we plan a more thorough empirical examination exploring how sensitive various machine translation systems are to various types of weeds.

References

- Abdelali, A., Guzman, F., Sajjad, H., and Vogel, S. (2014). The amara corpus: Building parallel language resources for the educational domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*.
- Antonova, A. and Misyurev, A. (2011). Building a web-based parallel corpus and filtering out machine-translated text. In *Proc. Building and Using Comparable Corpora (BUCC'11)*.
- Asia Online (2009). Study on the impact of data consolidation and sharing for statistical machine translation. Technical report, TAUS.
- Blum-Kulka, S. (1986). Shifts of cohesion and coherence in translation. In House, J. and Blum-Kulka, S., editors, *Interlingual and Intercultural Communication*. Narr, Tübingen.

- Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., and Tamchyna, A. (2014). Findings of the 2014 workshop on statistical machine translation. In *Proc. WMT*.
- Callison-Burch, C., Koehn, P., Monz, C., and Schroeder, J. (2009). Findings of the 2009 Workshop on Statistical Machine Translation. In *Proc. WMT*.
- Cettolo, M., Girardi, C., and Federico, M. (2012). WIT³: Web inventory of transcribed and translated talks. In *Proc. EAMT'12*, Trento, Italy.
- Cho, E., Niehues, J., and Waibel, A. (2014). Tight integration of speech disfluency removal into SMT. In *Proc. EACL'14*, Gothenburg, Sweden.
- Corston-Oliver, S., Gamon, M., and Brockett, C. (2001). A machine learning approach to the automatic evaluation of machine translation. In *Proc. ACL'01*, Toulouse, France.
- Diab, M., Fung, P., Hirschberg, J., and Solorio, T., editors (2016). *Proceedings of the Second Workshop on Computational Approaches to Code Switching*.
- Diab, M., Hirschberg, J., Fung, P., and Solorio, T., editors (2014). *Proceedings of the First Workshop on Computational Approaches to Code Switching*.
- Durrani, N., Sajjad, H., Hoang, H., and Koehn, P. (2014). Integrating an unsupervised transliteration model into statistical machine translation. In *Proc. EACL'14*, Gothenburg, Sweden.
- Fusco, M. (1990). Quality in conference interpreting between cognate languages: A preliminary approach to the Spanish-Italian case. *Interpreters' Newsletter*, (3).
- Gale, W. A. and Church, K. W. (1991). A program for aligning sentences in bilingual corpora. In *Proc. ACL'91*, Berkeley, California.
- Gamon, M., Aue, A., and Smets, M. (2005). Sentence-level MT evaluation without reference translations: Beyond language modeling. In *Proc. EAMT'05*, Budapest, Hungary.
- Goutte, C., Carpuat, M., and Foster, G. (2012). The impact of sentence alignment errors on phrase-based machine translation performance. In *Proc. AMTA'12*, San Diego, California.
- Hagiwara, M. and Sekine, S. (2011). Latent class transliteration based on source language origin. In *Proc. ACL'11*, Portland, Oregon, USA.
- Hellstern, A. and Marciano, J. (2014). Two sides of a coin: Machine translation and post-editing projects from the perspectives of the client and language services provider. *Proc. ATA'14*.
- Imamura, K. and Sumita, E. (2002). Bilingual corpus cleaning focusing on translation literality. In *Proc. INTERSPEECH'02*, Denver, Colorado.
- Jiang, J., Way, A., and Carson-Berndsen, J. (2010). Lattice score based data cleaning for phrase-based statistical machine translation. In *Proc. EAMT'10*, Saint-Raphaël, France.
- Khadivi, S. and Ney, H. (2005). Automatic filtering of bilingual corpora for statistical machine translation. In *Proc. NLDB'05*, volume 3513 of *Lecture Notes in Computer Science*.
- Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proc. MT Summit X*, Phuket, Thailand.

- Koppel, M. and Ordan, N. (2011). Translationese and its dialects. In *Proc. ACL'11*, Portland, Oregon.
- Kurokawa, D., Goutte, C., and Isabelle, P. (2009). Automatic detection of translated text and its impact on machine translation. In *Proc. MT Summit XII*, Ontario, Canada.
- Lembersky, G., Ordan, N., and Wintner, S. (2013). Improving statistical machine translation by adapting translation models to translationese. *Computational Linguistics*, 39(4).
- Li, H., Sim, K. C., Kuo, J.-S., and Dong, M. (2007). Semantic transliteration of personal names. In *Proc. ACL'07*, Prague, Czech Republic.
- Lui, M., Letcher, N., Adams, O., Duong, L., Cook, P., and Baldwin, T. (2014). Exploring methods and resources for discriminating similar languages. In *Proc. Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 129–138.
- Matthew (1st century). Gospel of Matthew. Greek New Testament.
- Mermer, C., Kaya, H., and Doğan, M. U. (2007). The TÜBİTAK-UEKAE statistical machine translation system for IWSLT 2007. In *Proc. IWSLT'07*, Trento, Italy.
- Myers-Scotton, C. (1993). *Social Motivations for Codeswitching: Evidence from Africa*. Oxford Studies in Language Contact.
- Notenbloom, L. (2009). Why do i get odd characters instead of quotes in my documents? <http://askleo.com>.
- Okita, T. (2009). Data cleaning for word alignment. In *Proc. ACL-IJCNLP'09*, Singapore.
- Palladius and Popov, P. S. (1888). *Китайско-русский словарь (Chinese-Russian Dictionary)*. Beijing, China. <https://archive.org/details/11888>.
- Pellom, B. and Hacıoglu, K. (2001). Sonic: The University of Colorado continuous speech recognizer. Technical Report TR-CSLR-2001-01, University of Colorado, Boulder, Colorado.
- Plitt, M. and Masselot, F. (2010). A productivity test of statistical machine translation post-editing in a typical localisation context. *Prague Bulletin of Mathematical Linguistics*, 93.
- Rarrick, S., Quirk, C., and Lewis, W. (2011). MT detection in web-scraped parallel corpora. In *Proceedings of the 13th Machine Translation Summit (MT Summit XIII)*, Xiamen, China.
- Resnik, P. (1998). Parallel strands: A preliminary investigation into mining the web for bilingual text. In *Proc. AMTA'98*, volume 1529 of *Lecture Notes in Artificial Intelligence*. Springer.
- Schwartz, L., Anderson, T., Gwinnup, J., and Young, K. (2014). Machine translation and monolingual postediting: The AFRL WMT-14 system. In *Proc. WMT'14*, Baltimore, Maryland.
- Simard, M. (2014). Clean data for training statistical MT: The case of MT contamination. In *Proc. AMTA '14*, Vancouver, Canada.
- Simard, M., Foster, G. F., and Isabelle, P. (1992). Using cognates to align sentences in bilingual corpora. In *Proc. Theoretical and Methodological Issues in Machine Translation (TMI'92)*, Montréal, Canada.
- Smith, J. R., Saint-Amand, H., Plamada, M., Koehn, P., Callison-Burch, C., and Lopez, A. (2013). Dirt cheap web-scale parallel text from the common crawl. In *Proc. ACL'13*, Sofia, Bulgaria.

- Upal, M. A. (2008). Personal correspondence.
- Venugopal, A., Uszkoreit, J., Talbot, D., Och, F., and Ganitkevitch, J. (2011). Watermarking the outputs of structured prediction with an application in statistical machine translation. In *Proc. EMNLP'11*, Edinburgh, Scotland.
- Wu, D. (1994). Aligning a parallel English-Chinese corpus statistically with lexical criteria. In *Proc. ACL'94*, Las Cruces, New Mexico.
- Young, K. M., Gwinnup, J., and Reinhart, J. (2012). Reversing the Palladius mapping of Chinese names in Russian text. In *Proc. AMTA'12*, San Diego, California.
- Zampieri, M., Tan, L., Ljubešić, N., and Tiedemann, J. (2014). A Report on the DSL Shared Task 2014. In *Proc. Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 58–67, Dublin, Ireland.
- Zukerman, E. (2013). Review: Amara is a web-based service that lets anyone transcribe and translate online video. <http://www.pcworld.com/article/2032787/review-amara-is-a-web-based-service-that-lets-anyone-transcribe-and-translate-online-video.html>.

Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government. Cleared for public release on 4 Dec 2014. Originator reference number RH-14-113337. Case number 88ABW-2014-5534.

Toward Temporally-aware MT: Can Information Extraction Help Preserve Temporal Interpretation?

Taylor Cassidy

Army Research Laboratory, Adelphi, MD 20783, USA

taylor.cassidy.civ@mail.mil

Jamal Laoudi

ARTI, Fairfax, VA 22030

jamal.laoudi.ctr@mail.mil

Clare Voss

Army Research Laboratory, Adelphi, MD 20783, USA

clare.r.voss.civ@mail.mil

Abstract

Users of MT systems often need to glean information about the world from foreign language texts for specific tasks, such as documenting how events, as mentioned in those texts, fit on a time line. Current systems have not been systematically evaluated for their adequacy in preserving *temporal interpretation*, i.e., the set of temporal relations that a reader naturally takes to hold among the states, events, and time expressions mentioned in the text, as well as the intrinsic temporal properties of each, although some MT research has focused on exploiting linguistic mechanisms, such as verbal tense or aspectual markers to convey temporal information. We describe ongoing work to develop a method for (i) building parallel TimeBanks with annotated temporal interpretation on parallel texts, (ii) leveraging these resources to train and evaluate the emerging class of *temporal interpretation extraction systems* on new languages, and (iii) developing *time-aware MT systems* that aim to preserve the temporal interpretation of source language text in their target language outputs. We present our approach and results from our exploratory analyses into the preservation of temporal interpretation in Arabic-English MT, and propose shared tasks to bring together research in information extraction and machine translation, geared toward building time-aware MT.

Users of MT systems often need to be able to glean information about the world from foreign language texts for specific tasks, such as documenting their understanding of how events, as mentioned in those texts, fit on a time line. While task-based metrics have evaluated the extent to which MT preserved who, when, and where information (Voss and Tate, 2006) or information required to pass language proficiency tests (Jones et al., 2005; Matsuzaki et al., 2015), current systems have not been systematically evaluated for their adequacy in preserving *temporal interpretation*, i.e., the set of temporal relations that a reader naturally takes to hold among the states, events, and time expressions mentioned in the text, as well as the intrinsic temporal properties of each.

Some MT research has focused on exploiting linguistic mechanisms, such as verbal tense or aspectual markers, when available in text to convey specific forms of temporal information. MT systems incorporating this research address the challenge of preserving temporal content narrowly, e.g., selecting the correct target language tense for each source language verb, or selecting the correct sense to translate temporal discourse connectives. However different languages rely on a much wider range of explicit temporally-significant linguistic mechanisms to convey underlying temporal content, including tense, aspect, function words, discourse con-

nectives, syntactic relations, idiomatic expressions.¹ These mechanisms are manifested asymmetrically across languages; as a result, reference translations may use different mechanisms compared with those in the source.² Thus proper selection and use of these mechanisms is non-trivial. The MT challenge of preserving temporal interpretation from source language to the target language output goes beyond current approaches and necessarily subsumes working with many forms of temporal information.³

A growing body of computational research now studies temporal interpretation in text, having initially emerged to support systems performing tasks such as information extraction and knowledge base construction, and thus has generally taken place outside the MT research community. This research has become multilingual and a variety of corpora in many languages, including parallel corpora (Forăscu and Tufiş, 2012), have been annotated using the TimeML annotation scheme (Pustejovsky et al., 2003a). In addition, temporal interpretation algorithms implemented within extraction systems have successfully made use of a variety of features drawn from the linguistic temporal mechanisms listed above (UzZaman et al., 2013; Bethard et al., 2016), though no one feature type stands out as dominant. Despite the fact that such annotation frameworks and automatic extraction algorithms for temporal interpretation exist, and translation of certain temporal linguistic mechanisms has been improved, little has been written about explicitly preserving temporal interpretation in MT. Simply put, MT engines are not built to be fully temporally-aware.

This paper describes ongoing work to develop a method for (i) building parallel TimeBanks with annotated temporal interpretation on parallel texts, (ii) leveraging these resources to train and evaluate the emerging class of *temporal interpretation extraction systems* on new languages, and (iii) developing *time-aware MT systems* that aim to preserve the temporal interpretation of source language text in their target language outputs. We present our approach and results from our exploratory analyses into the preservation of temporal interpretation in Arabic-English MT, and conclude by proposing shared tasks to bring together research in information extraction and machine translation, geared toward building time-aware MT.

2 Temporal Interpretation and Its Preservation Across Languages

2.1 Definition of Temporal Interpretation

Similar to Katz and Arosio (2001)’s “radically simplified semantic formalism,” we start with the notion that the temporal interpretation of a text is *the set of temporal relations that a reader naturally takes to hold among the states, events, and time expressions mentioned in the text, as well as the intrinsic temporal properties of each*. Intuitively, the temporal interpretation provides sufficient information to answer questions about when events occur and states obtain to the level of precision and certainty intended by the text’s author. Temporal interpretation is independent of a text’s accuracy with

¹E.g., “when women go to the farm” can mean during the hours 8-9am in Babungo Schaub (1985).

²Indeed, this asymmetric usage may not be optional. For example, Chinese has no clear analogue to English grammatical verb tense.

³We distinguish *information*, as processed raw text data, from *interpretation*, as a process by which information is grounded or annotated for inclusion, for example, in a knowledge base or a time line. (The processing in either case may be manual or automated.) As explained further in section 2, for us, temporal information is a narrow term that refers to explicit surface mentions with temporal content, while temporal interpretation is a broader, more inclusive term for a process, or the result of a process, that starts with temporal information, but also allows for inference to derive implicit temporal entities and relations using world knowledge, such as date-time arithmetic.

respect to what transpires in the actual world, and for temporal relations left vague in a text, their interpretation may vary among readers due to the unique perspectives and prior knowledge that they each bring to understanding the text.

Our working definition of the temporal interpretation of a text is inspired by the TimeML annotation scheme, as used by Pustejovsky et al. (2003b). Here we focus only narrowly on core aspects of temporal interpretation to give the reader a sense for what interpretation means procedurally, first in terms of what information is annotated in text and second, how the annotated information then is connected to specific concepts of time.⁴ We use the label SL to stand for a source language text and $I(SL)$ for the temporal interpretation over that text.⁵ We define that interpretation as a five-tuple, $\langle SL_e, SL_t, \tau_{SL}, \sigma_{SL}, r_{SL} \rangle$. First, two sets designate what information is annotated: the set SL_e consists of the events and states (henceforth shortened collectively to “events”) mentioned in SL and the set SL_t contains the time expressions mentioned in SL . The elements of their union are referred to collectively as *temporal entities*. Second, the remaining items in the five-tuple are functions that designate how these entities are connected to concepts of time. The function τ_{SL} maps mentioned time expressions, SL_t , into actual times, as captured in T , the set of all possible time values (e.g., as identified in ISO-8601). The function σ_{SL} maps mentioned events, SL_e , into a space of semantic property values, such as *class*, *polarity*, *modality*, *tense*, etc.⁶ The function r_{SL} maps pairs of temporal entities into temporal interval relations, elements in set S (e.g., as in Allen’s interval relations).

Several shared task workshops in the past decade have tackled automating the temporal interpretation of text. To evaluate the algorithms in computational systems built for this task, the TempEval and Clinical TempEval tracks of previous SemEval workshops used, as their ground truth, manually annotated corpora such as TimeBank (Pustejovsky et al., 2003b) and THYME (Styler IV et al., 2014). Both corpora use a version of TimeML annotation schema, with guidelines for identifying and annotating event and time expression words or phrases (defined above as SL_e and SL_t). Each *temporal entity* corresponding to an event or time expression is conceptualized as having a *temporal extent* that is either an interval or a set of intervals.⁷ Events in TimeML are also labeled with semantic properties such as *class*, *polarity*, *modality*, *tense*, etc. At the current time, inter-annotator agreement (IAA) rates however on the manual task of annotating texts for temporal interpretation have varied considerably depending on the particular setting and they tend to be lower than for similar text annotation tasks. For example, TimeML relations generally achieve Kappa scores between .4-.8, while PropBank annotations for argument roles achieve higher .91-.96 Kappa (Palmer et al., 2005).

Nonetheless even with such varied IAA results, researchers have built computational systems to perform temporal interpretation extraction⁸ with supervised machine learn-

⁴We later extend this to less obvious sources of temporal information, such as definiteness.

⁵Similarly TL and $I(TL)$ stand for a target language text and its temporal interpretation.

⁶That space is $\times_i^N S_i$, the cross product of N sets of event semantic property values, S_i .

⁷Other details of TimeML’s interval annotation include the following. End points may be precise dates and times, or as less specific values from a pre-defined set, e.g. “morning”. Intervals may or may not be anchored to an actual time line. Time expressions are assigned their extents directly via a Timex3 tag. When an event and its location in time are determined, that pair of temporal entities is tagged with a TLINK characterizes their relationship (e.g. Before, Overlap). TLINKs are also used to relate events in time relative to other events.

⁸We refer information extraction (IE) systems that focus explicitly on incorporating temporal interpretation into its annotations as *temporal interpretation systems* or *temporal extraction systems*.

ing algorithms to identify events, time expressions, and their semantic properties. The systems work with features for word and character n-grams, POS tags, information from lexical ontologies, and distributional semantic vectors. For labeling higher-order temporal entity pairs with interval relations, systems make use of additional types of linguistic knowledge such as syntactic context from the dependency path between temporal entities. For the specific subtask of time expression normalization, the highest-performing systems have generally been rule-based (Strötgen and Gertz, 2010; Chang and Manning, 2012). Researchers have had less success in modeling implicit world knowledge in temporal extraction systems. For example, Mirza and Tonelli (2014) observed that adding a feature encoding typical event duration to an SVM classifier decreased system accuracy. Unsurprisingly, annotation guidelines differ on whether such inferences are allowed. The TempEval corpora annotation guidelines encourage the use of world knowledge (Verhagen et al., 2009), while the still-evolving Richer Event Description (RED) guidelines prohibit its use.⁹

To date, manual temporal interpretation annotation is carried out monolingually by native speakers of the language of the text being annotated, who are trained with language-specific guidelines. And so temporal interpretation systems are also constructed independently in separate languages based on those resources. Schematically, we illustrate this situation with separate rows in figure 1a¹⁰.

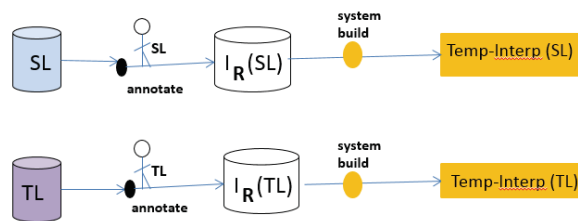


Figure 1a: Annotation of Temporal Interpretation in Texts for System Construction: Monolingual workflows here are independent for source & target language (SL & TL).

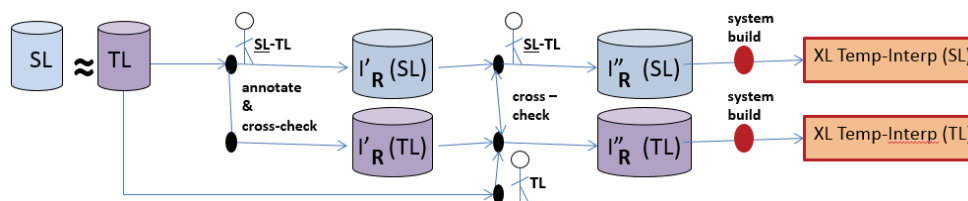


Figure 1b: Annotation of Temporal Interpretation in Parallel SL and TL Texts for System Construction: Cross-lingual workflows here are interdependent by establishing semantic correspondence of temporal annotations in SL & TL prior to system builds.

⁹For example, given the text “We diagnosed her cancer last week”, TempEval guidelines would permit annotating that the cancer preceded the diagnosis, whereas RED guidelines would not. (see <https://github.com/timjogorman/RicherEventDescription>.)

¹⁰In our figures, cylinders represent datasets, circles represent human or automated processes, and colored rectangles represent software systems.

2.2 Preservation of Temporal Interpretation

Given that established methods now exist for annotating temporal interpretation in text and that computational systems can extract temporal information, we now ask, how might these monolingual methods, corpora, and systems be leveraged to work cross-lingually to *preserve* temporal content from the source language in machine translation to the target language?

When evaluating the translation of a text from one language to another, it is natural to ask whether the meaning of the text is fully preserved by the translation. Here we focus on preservation of the text's temporal interpretation only. Consider two texts: an SL and its translation TL . As show in figure 1a, let $I_R(SL)$ and $I_R(TL)$ be their temporal interpretations as derived independently by a native speaker of each language.¹¹ We say that the interpretations are *semantically equivalent* when all their identified core components are *semantically equivalent*, i.e., their identified temporal entities correspond cross-lingually (SL_e with TL_e and SL_t with TL_t) and the values of temporal grounding functions (τ_{SL}, σ_{SL} , and r_{SL}) over SL identified entities or entity pairs in their domains, correspond cross-lingually with function values (τ_{TL}, σ_{TL} , and r_{TL}) over TL identified entities or entity pairs in their domains. When $I(SL)$ and $I(TL)$ are semantically equivalent, we say that the translation TL *strongly preserves* the interpretation $I(SL)$.

This formal notation is of course a conceptual-level abstraction, removed from the realities of actual cross-language divergences in how and where temporal information is expressed. Information that is explicitly lexical in a sentence in one language may be grammaticalized in its translation in another language, and left implicit in another language. These divergences complicate the detective work of identifying the temporal content that is preserved in actual translation by humans and machine translation. Intuitively, we would like to say that a translation of SL to TL preserves temporal content to the extent that native speakers of each language independently arrive at the same temporal interpretation. However, even parallel texts may yield distinct temporal interpretations, and so we have begun experimenting with a process for converging those interpretations as part of the annotating process over parallel texts.¹²

Operationally, we consider the most likely scenario for such preservation in corpus construction, when annotators seek to interpret temporal information in parallel texts. We assume both a bilingual SL-TL annotator who is native in the SL, shown schematically in figure 1b as stick person labeled SL-TL with SL underlined, and a monolingual TL speaker, also shown as a stick person with label TL only. The bilingual annotator can read given parallel SL and reference TL texts, and develop annotations by cross-checking texts and their temporal interpretations in tandem. We designate their initial interpretations I'_R , where the single quote indicates the first pass and the subscript R indicates reference interpretation.¹³ The bilingual aims in their annotation for semantic correspondence of $I'_R(SL) = I'_R(TL)$. For quality control, the native TL speaker can read TL and $I'_R(TL)$, and then cross-check the TL annotations for semantic correspondence to the annotations only in $I'_R(SL)$, by conducting a systematic review of temporal entities and their relations in both texts with the bilingual annotator. Since first-pass interpretations and the reference translation itself may be changed in the review, we adopt the second-pass results, $I''_R(SL)$ and $I''_R(TL)$, as reference interpretations for

¹¹The subscript R indicates these are human reference interpretations.

¹²Our initial efforts in developing this process are described in section 3.

¹³The quote and subscript will later serve to distinguish this from, respectively, subsequent second pass with double quotes I''_R and automated interpretations with subscripts to identify the system.

system builds and evaluations.

We also extend the notion of preserving temporal information for the purpose of evaluating machine translation output, " TL ".¹⁴ When an MT user who knows they are reading MT output and uses that knowledge in drawing inferences, then creates a temporal interpretation of " TL ", we denote their interpretation $\hat{I}("TL")$. Furthermore, we say the MT engine has *weakly preserved* the source interpretation when the MT output has translation errors in conveying that original temporal content, but the user can overcome that information with their background knowledge, as shown when subsequent evaluation of the MT user's $\hat{I}("TL")$ shows it to be semantically equivalent to the reference $I'(TL)$.

2.3 Time-Aware Machine Translation: Current State of the Art

Our long-term research goal is to develop time-aware machine translation systems that preserve temporal interpretation from the source text in the target language output. The most relevant research efforts have aimed to correctly translate specific types of linguistic mechanisms that speakers use to convey temporal interpretation. These efforts therefore indirectly aim to preserve temporal interpretation. Access to tense (Klavans and Chodorow, 1992), lexical aspect and temporal connectives (Dorr and Gaasterland, 2002), have been shown to help lexical choice, critical for all MT systems. For statistical systems, automatic prediction of target language tense based on source language verbs has been a popular task, especially for Chinese-to-English translation due to the lack of overt tense on Chinese verbs (Olsen et al., 2001; Ye et al., 2006; Baran, 2013; Ge et al., 2015; Loaiciga et al., 2014).

There have been few efforts to integrate linguistic temporal mechanisms from source text directly into a statistical MT system, and evaluate its impact on MT performance. Meyer et al. (2013) used a factored translation model that included a binary narrativity tag on each English source verbs in the simple past tense as a feature to improve choice of French output tense. Loaiciga et al. (2014) extended this model by using a supervised machine learning classifier to further tag each English source text verb with one of nine possible French verb tenses. It's worth noting that one of the features used by English verb tagger was derived from event tense, aspect, and class for event pairs as labeled by a temporal interpretation extractor. They were able to achieved 10% improvement on tense translation. Gong et al. (2012) re-score translation hypotheses during decoding time using (1) English tense labels automatically assigned to Chinese source-language verbs, and (2) a tense n-gram language model that models the probability of a given sequence of tenses in English text. Meyer et al. (2015) use a factored language model to leverage discourse connective marker types. They train a supervised classifier to predict Penn Discourse Treebank style tags on discourse connectives, whose values include time-relevant values such as temporal, temporal-durative, temporal-punctual, temporal-contrast, and temporal-causal. This work uses similar features to Loaiciga et al. (2014).

While it is reasonable to expect improving translation of verbal tense or discourse connectives in particular will increase the likelihood that temporal ordering will be preserved, we are eager to develop a broader approach to improving temporal interpretation preservation in MT. We will return to these questions pertinent to this broader approach after the exploratory analysis section: (i) What resources are needed to assess the extent to which current MT systems preserve temporal information? (ii) How might temporal interpretation extraction systems be integrated into the workflow for building

¹⁴The quotation marks distinguish MT output as " TL " from reference translation text labeled TL , to remind readers that the accuracy and fluency of MT output is rarely equivalent to its reference TL .

MT engines, and how will this impact the quality of MT output?

3 Exploratory Analysis

This section documents our ongoing effort to determine how well current MT systems preserve temporal interpretation. Our procedure is depicted in Figure 2. The results of the procedure are a reference translation that strongly preserves the temporal interpretation of the source, MT output for each source sentence, and an evaluation of whether MT output weakly preserves the source’s temporal interpretation.

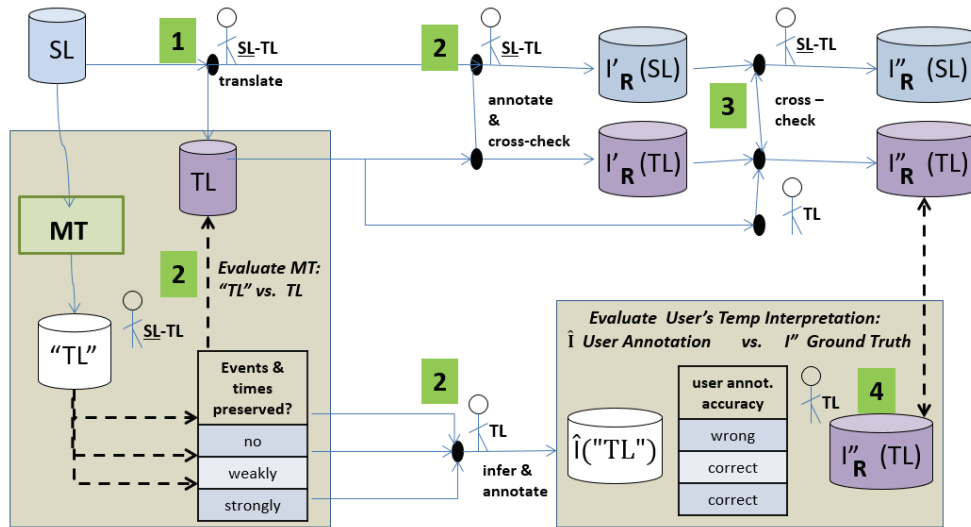


Figure 2: Procedure for Exploratory Analysis.

This procedure requires two participants, a bilingual source language native speaker fluent in the target language (P_{SL-TL}), and a native speaker of the target language (P_{TL}).

In step 1 P_{SL-TL} translates SL text to obtain TL aiming to preserve temporal interpretation, while MT translates SL to obtain "TL". These translations are independent of each other and so can be done during this same step. In the step 2, P_{SL-TL} derives a temporal interpretation for both SL and TL, $I'_R(SL)$ and $I'_R(TL)$. In addition, P_{SL-TL} identifies the events and time expressions in "TL" corresponding to those identified in the Arabic source, noting where this temporal information is strongly, weakly, or not at all preserved. Meanwhile, P_{TL} independently constructs $\hat{I}("TL")$ without any access to the reference translation.¹⁵ In step 3, P_{SL-TL} and P_{TL} work together on two tasks. P_{TL} provides target language expertise to help P_{SL-TL} alleviate inconsistencies between $I'_R(TL)$ and $I'_R(SL)$, yielding revised interpretations $I''_R(TL)$ and $I''_R(SL)$ and possibly revisions to TL text. P_{SL-TL} works with P_{TL} , who can also consult the reference translation TL during this step of cross-checking. In step 4, after the cross-checking is complete, P_{TL} is then able to assess the extent to which $\hat{I}("TL")$ is equivalent to $I''_R(TL)$, assessing whether MT users can overcome MT output errors, and so allow

¹⁵A variety of factors bear on one’s ability to determine $\hat{I}("TL")$, and the methodological choices made in doing so, such as: (i) knowledge about the MT algorithm (e.g. statistical vs. rule-based), (ii) experience reading MT output, both in general and from the MT engine in question, (iii) knowledge of SL and access to the source text, (iv) experience manually translating the source to TL.

for the possibility that some preservation, what we are calling weak preservation, was indeed achieved by the MT engine.

Our analysis thus far consisted of a shallow pass through the procedure described above. No annotation guidelines were strictly enforced, but we roughly adhered to TimeML guidelines for identifying events and time expressions. Temporal relations were not explicitly annotated; however, both participants were previously trained to perform the temporal relation annotation task, and holistically developed informal temporal interpretations. Rather than compute quantitative results, we identified representative cases illustrating phenomena relevant to preservation of temporal interpretation. Selected results are reported in section 4.

The participants were authors of this paper. P_{TL} is a native English speaker, and P_{SL-TL} is a native Arabic speaker and fluent in English, with machine-aided translation experience. We used 10 documents from DARPA's GALE dataset and several manually translated recent sports news documents.

4 Phenomena Pertaining to Temporal Interpretation Preservation

We encountered a variety of phenomena that lead to a failure to preserve temporal information. However, instances where MT output fail even to weakly preserve the Arabic temporal interpretation are of particular concern given that we are driven primarily by practical usage of MT. In this section we discuss instances where one of the following was the case: (i) English MT output failed to weakly preserve Arabic temporal interpretation (thus, strong preservation failed as well); or, (ii) English MT output failed to strongly preserve Arabic temporal interpretation, but did weakly preserve it, in spite of the fact that elements of temporal significance were incorrectly translated.

Incorrect translation of events: Perhaps the most obvious source of error is when events and time expressions themselves are not translated correctly. In some cases, an Arabic event word that should have been translated was transliterated, or vice versa. In other cases the incorrect English word in MT output can lead the user to unintended interpretation. Table 1 provides two representative examples where temporal interpretation is not preserved

Poor time expression interpretation: Temporal expressions often serve as *temporal containers* with respect to which many events can be related. Rather than annotate every single pair of events, annotators can effectively specify a great deal of a text's temporal interpretation by relating each event to a few key time expressions. Thus, assigning the wrong temporal extent can greatly distort temporal interpretation. We found instances where time expressions rely on the semantics of certain verbs for their correct interpretation. Consider the temporal interpretation of "year" in the following example:

- وقال "لدينا فريق رائع، جعلنا نستمتع في كل مباراة. من المؤكد أننا بصدد عاما صعبا للغاية"، ولكننا مستعدون"
- Ref: He said: "we have an amazing team that allowed us to enjoy every game. Surely **we are embarking on a very difficult year**, but we are ready".
- MT: He said, "We have a great team, made us enjoy every game. Sure, we **are in a very difficult year**, but we are ready."

In the reference translation, "year" is a time period that begins in the future or has just begun speech time. In the MT, by virtue of losing the verb "embarking", we

Arabic	MT Output	Reference	Explanation
المباحثات تناولت الوسائل الكفيلة بتعزيز التعاون	Alambagesat dealt with ... enhancing cooperation	The discussions were concerned with ways to enhance cooperation ...	The Arabic word المباحثات was misspelled as المباحثات , which lead MT to transliterate to “Alambagesat” instead of translating to “discussion”.
تطرق رئيس نادي برشلونة الإسباني، جوزيب ماريا بارتوميو، إلى الحالة الاقتصادية القوية...اليوم	Turning president of FC Barcelona, Josep Maria Bartomeu, the strong economic situation ... today	The president of the Spanish club Barcelona, Josepi Maria Bartomero, addressed the current strong financial status ... today ...	“addressed” mistranslated as “Turning” could lead to interpretation that the presidency is just beginning today

Table 1: Examples of Incorrect Event Translations with Explanations.

lose the interpretation that the year is just beginning or about to begin is not preserved (however, “we are ready” does suggest this interpretation).

We encountered instances where noun phrase definiteness affects the interpretation of time expressions. Consider the following example:

- وقال في هذا الصدد "نريد غلق بعض الملفات دون تعجل. سندافع عن اللاعبين كلما اقتضت الحاجة"، مشيدا بالعام الذي "لا ينسى" على المستوى الرياضي بعد الأربعة ألقاب التي حققها الفريق".
- Ref: He stated about this topic that “we want to close these cases in a timely fashion. We will defend the players as needed,” praising **this “unforgettable” year** at the competitive level after the four trophies the team has won.
- MT: He explained, ”We enjoyed it a lot and managed to destroy the attacking triangle of all the numbers, we’ve lived unforgettable **years** and we have made four titles reflect a lot of personal competitive team.”

Here, the definite singular demonstrative determiner “this” in the reference translation indicates a single year, where the time of speech is at or near the end of that year. The source text uses a definite singular determiner. The indefinite plural “years” in the MT output, however, indicates multiple years in the past, not necessarily contiguous, leaving whether the current year is part of that set of years unspecified. Thus, the source text temporal interpretation is not preserved.

Missing or Mistranslated Function Words in MT Output: A single missing word can dramatically alter the meaning of a sentence, including its temporal interpretation. In the following case the missing word is the conjunction “and”, though we observed missing and mistranslated prepositions as well.

- وتأتي زيارة¹ ساركوزي الى المغرب بعد اربعة ايام² من زيارة رئيس مكتب التحقيقات الفدرالي الاميركي (اف بي أي) روبرت اس مولر الى المغرب وتسليم³ الولايات المتحدة ثلاثة معتقلين مغربيين من معتقل قاعدة غوانتانامو الى المغرب

- Ref: Sarkozy's **visit**_{e1} to Morocco comes **four days** after the **visit**_{e2} to Morocco of FBI Chief Robert S. Mueller, **and** the **handing over**_{e3} of three Moroccan prisoners from the Guantanamo detention camp to Morocco by the United States.
- MT: Sarkozy's **visit**_{e1} to Morocco, **four days** after the **visit**_{e2} of the US Federal Bureau of Investigation (FBI) Robert S. Mueller to Morocco, The United States **delivered**_{e3} three Moroccan detainees from Guantanamo Bay base to Morocco.

Both reference translation and MT output show that e1 occurred after e2 (with four days separating the two events). The presence of “and” in the reference indicates that e3 also occurred four days after e2 and an equivalent word plays the same role in the source text. However, the relative ordering of e2 and e3 is not so clear in the MT output. It seems likely that a word or phrase is missing between the comma and “The United States” in the MT output, but its impossible to tell if it should be “and”, or “just before”, or something else. The missing main verb “comes” adds additional complexity. The source temporal interpretation is therefore not preserved.

Phrases are Incorrectly Ordered: We found instances where incorrectly ordered phrases significantly impacts temporal interpretation.

- MT: Sarkozy's **visit**_{e1} to Morocco after **four days**_{t1} of **visiting**_{e2} Chairman of US Federal Bureau of Investigation (FBI) Robert Mueller Vegas to Morocco and **extradition**_{e3} to The United States three Moroccan prisoners from Guantanamo to Morocco.

The relative order of “after” and “four days” is switched, resulting in the possible interpretation that “four days” describes the length of the “visit” (e2) as opposed to the time elapsed between e1 and e2. Here, a single change in word order leads the reader to favor a syntactic analysis where “four days of visiting” is a linguistically motivated phrase. While this interpretation leaves the existence of the second “to Morocco” unexplained, recovering the fact that “Robert Mueller” is the agent of event e2 and that “to Morocco” describes the location of that visit is quite difficult; thus, temporal interpretation is not preserved.

Tense and Aspect: Tense and aspect are generally considered important cue to temporal interpretation. Most machine learning and rule based temporal information extraction systems use tense and aspect as features. The literature on tense-aware MT presents cases where poor tense translation results in a failure to preserve temporal interpretation (e.g. Meyer (2014)), which we also observed. In contrast, we present here cases in which temporal interpretation is weakly preserved in spite of incorrect translation of tense and aspect.

- وأشار إلى أن الوزيرين "سيقترحان على نظيرهما الإسباني (خوسيه انطونيو الونسو) عقد اجتماع ثلاثي في القريب العاجل حول مكافحة تهريب المخدرات

- Ref: It also noted that both ministers “**will suggest** to their Spanish counterpart Jose Antonio Alonso **to hold** a tripartite meeting in **the near future** regarding combating drug trafficking.”
- MT: He noted that the two ministers “**will suggest** on Spanish counterpart (Jose Antonio Alonso) tripartite meeting **was held** in the near future on the fight against drug trafficking.”

In the reference translation we see that the meeting will take place after the suggesting, and during a period of time denoted by “the near future”. That Arabic similarly uses a past tense and infinite verb. It is fairly clear, however, that the past tense “was held” in the MT output is a tense error, as it seems likely that the meeting takes place during “the near future”. To the extent that one is familiar with Arabic to English translation, the MT system in question, and the training data, an incorrect tense translation (i.e. that “was held” should be “to hold”) is far more likely than a time expression denoting a past time being translated as “in the near future”. Another possibility is that a more accurate translation would put “in the near future” near the end of the sentence, so that it attaches to “drug trafficking” or “fight”. In this case we would have less evidence that “was held” has the wrong tense. It is not uncommon for an adverbial phrase to appear out of proper order. We also must face the problem that if tense translation errors are common, “will suggest” could have the wrong tense as well.

In spite of the various tense error possibilities, it seemed obvious that “was held” had the wrong tense, and so the source interpretation is weakly preserved. This is likely because in the context of the article, it does not make sense for the ministers to suggest that a meeting was held in the past. In another context, a similar construction might be more acceptable, as in:

- In his closing arguments, the defense attorney will suggest that a meeting was held for the purpose of framing his client in the near future.

Similarly, in the example below the lexical selection and semantics of the verb “to accuse” allow us to infer that an accusation of possession of weapons does not precede the possession itself, in spite of a tense translation error, resulting our categorizing this as a case of weak preservation by the MT. Note that replacing “to accuse” with “to convince” would not permit that inference, and we would not label this a case of weak preservation by the MT.

• والأدهى من ذلك وأمرٌ سعيٌ بعضهم إلى تلفيق أكاذيب لاتهم هذا البلد أو ذاك بامتلاك أسلحة دمار شامل

- Ref: Even worse and more bitter is the attempt by some to fabricate lies, **accusing** this or that country of **possessing** weapons of mass destruction ...
- MT: Worse still, some of them and is seeking to fabricate lies **to accuse** this or that country **to possess** weapons of mass destruction ...

5 Conclusion: Proposal for New Resources & Shared Tasks

Given the exploratory analyses with the wide range of challenges in preserving temporal interpretation in MT of Arabic into English, we can now return to our earlier questions and conclude with a proposed research way forward:

(i) What resources are needed to assess the extent to which current MT systems preserve temporal information?

Parallel TimeBanks. There already exist corpora with temporal interpretation annotation in a variety of languages,¹⁶ but we are only aware of one parallel effort, the Romanian TimeBank, built by translating English TimeBank and manually projecting the annotations to the Romanian translations. The creation of parallel TimeBanks, i.e., parallel corpora annotated for temporal interpretation, as in

¹⁶There exist TimeML annotated corpora in English, Spanish, French, Italian, Korean, Chinese, Indonesian, Brazilian Portuguese, Farsi, Estonian, Romanian.

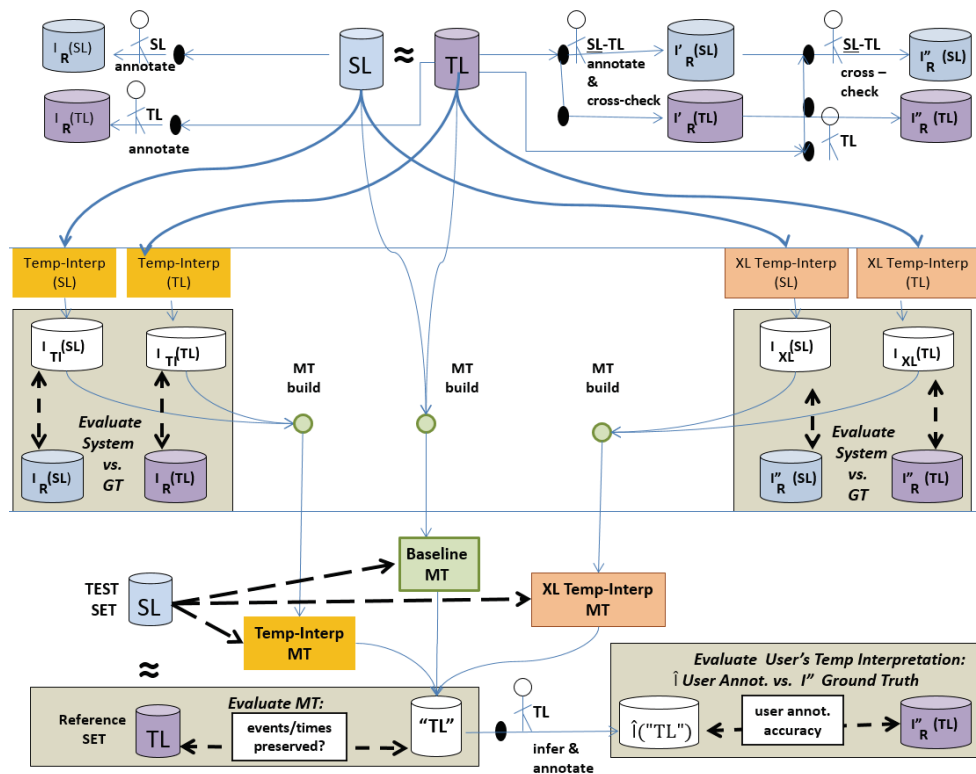


Figure 3: Workflow for Developing Parallel TimeBanks and Time-Aware Machine Translation.

figure 1a and the top of figure 3, will make it possible to support further development of temporal interpretation systems and new efforts toward time-aware MT, both shared tasks proposed below. We propose building a parallel English-Arabic TimeBank.

(ii) How might temporal interpretation extraction systems be integrated into the building of MT engines, and how will this impact the quality of MT output?

Shared Task: Temporal Interpretation Extraction Leveraging Parallel TimeBanks Shared tasks evaluating temporal interpretation algorithms have been conducted primarily in English. Each potential source language brings its own challenges due to the different ways in which temporal information is conveyed. We anticipate that the availability of parallel TimeBanks will encourage research on transfer learning for temporal interpretation extraction (see middle row of figure 3).

Shared Task: Time-Aware MT Leveraging Temporal Interpretation Extraction The desired results of the above proposed efforts are (1) a parallel English-Arabic TimeBank, and (2) publicly available tools for temporal interpretation extraction for Arabic text (those for English already exist). Given these resources, we propose a shared task to build MT systems that preserve the temporal interpretation of the source text. This task would address research questions pertaining to semantics-based MT as well as MT evaluation, both intrinsic and extrinsic (as in left and right evaluation boxes respectively in bottom row of figure 3).

References

- Baran, E. (2013). *Chinese Verb Tense? Using English Parallel Data to Map Tense onto Chinese and Subsequent Tense Classification*. PhD thesis.
- Bethard, S., Savova, G., Chen, W.-T., Derczynski, L., Pustejovsky, J., and Verhagen, M. (2016). Semeval-2016 task 12: Clinical tempeval. *Proceedings of SemEval*, pages 1052–1062.
- Chang, A. X. and Manning, C. D. (2012). Sutime: A library for recognizing and normalizing time expressions. In *LREC*, pages 3735–3740.
- Dorr, B. J. and Gaasterland, T. (2002). Constraints on the Generation of Tense , Aspect , and Connecting Words from Temporal Expressions. *Knowledge Creation Diffusion Utilization*, 1:1–47.
- Forăscu, C. and Tufiş, D. (2012). Romanian timebank: An annotated parallel corpus for temporal information.
- Ge, T., Ji, H., Chang, B., and Sui, Z. (2015). One Tense per Scene : Predicting Tense in Chinese Conversations. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 668–673.
- Gong, Z., Zhang, M., Tan, C. L., and Zhou, G. (2012). Classifier-based tense model for smt. In *COLING (Posters)*, pages 411–420. Citeseer.
- Jones, D., Shen, W., Granoien, N., Herzog, M., and Weinstein, C. (2005). Measuring Translation Quality by Testing English Speakers with a New Defense Language Proficiency Test for Arabic. *International Conference on Intelligence Analysis*, page 6.
- Katz, G. and Arosio, F. (2001). The annotation of temporal information in natural language sentences. *Proceedings of the workshop on Temporal and spatial information processing-Volume 13*, 13(1):15.
- Klavans, J. L. and Chodorow, M. (1992). Degrees of Stativity: The Lexical Representation of Verb Aspect. In *Proceedings of {COLING} 1992*, pages 1127–1131.
- Loaiciga, S., Meyer, T., and Popescu-Belis, A. (2014). English-french verb phrase alignment in europarl for tense translation modeling. In *LREC*, pages 674–681.
- Matsuzaki, T., Fujita, A., Todo, N., and Arai, N. H. (2015). Evaluating Machine Translation Systems with Second Language Proficiency Tests. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 145–149.
- Meyer, T. (2014). Discourse-level features for statistical machine translation.
- Meyer, T., Grisot, C., and Popescu-Belis, A. (2013). Detecting narrativity to improve english to french translation of simple past verbs. In *Proceedings of the 1st DiscoMT Workshop at ACL 2013 (51th Annual Meeting of the Association for Computational Linguistics)*, page 8.
- Meyer, T., Hajlaoui, N., and Popescu-Belis, A. (2015). Disambiguating discourse connectives for statistical machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(7):1184–1197.

- Mirza, P. and Tonelli, S. (2014). Classifying Temporal Relations with Simple Features. *Aclweb.Org*, (2006):308–317.
- Olsen, M., Traum, D., Van~Ess-Dykema, C., and Weinberg, A. (2001). Implicit Cues for Explicit Generation: Using Telicity as a Cue for Tense Structure in a Chinese to English MT System. In *Proceedings of MT Summit VIII, Santiago de Compostella, Spain*.
- Palmer, M., Gildea, D., and Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.
- Pustejovsky, J., Castano, J. M., Ingria, R., Sauri, R., Gaizauskas, R. J., Setzer, A., Katz, G., and Radev, D. R. (2003a). Timeml: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34.
- Pustejovsky, J., Hanks, P., Sauri, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L., et al. (2003b). The timebank corpus. In *Corpus linguistics*, volume 2003, page 40.
- Schaub, W. (1985). *Babungo*. Taylor & Francis.
- Strötgen, J. and Gertz, M. (2010). Heideltime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 321–324. Association for Computational Linguistics.
- Styler IV, W. F., Bethard, S., Finan, S., Palmer, M., Pradhan, S., de Groen, P. C., Erickson, B., Miller, T., Lin, C., Savova, G., et al. (2014). Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics*, 2:143–154.
- UzZaman, N., Llorens, H., Derczynski, L., Verhagen, M., Allen, J., and Pustejovsky, J. (2013). Semeval-2013 task 1: {T}empeval-3: Evaluating time expressions, events, and temporal relations. *Second joint conference on lexical and computational semantics (* SEM)*, 2(SemEval):1–9.
- Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Moszkowicz, J., and Pustejovsky, J. (2009). The tempeval challenge: identifying temporal relations in text. *Language Resources and Evaluation*, 43(2):161–179.
- Voss, C. R. and Tate, C. R. (2006). Task-based Evaluation of Machine Translation (MT) Engines : Measuring How Well People Extract Who , When , Where-Type Elements in MT Output. *Proceedings of the 11th Annual Conference of the European Associatio for Machine Translation (EAMT 2006)*, pages 203–212.
- Ye, Y., Fossum, V. L., and Abney, S. (2006). Latent Features in Automatic Tense Translation between Chinese and English. (July):48–55.

“Did You Mean...?” and Dictionary Repair: from Science to Engineering

Michael Maxwell, presented by Petra Bradley
(mmaxwell@casl.umd.edu, pbradley@casl.umd.edu)
University of Maryland, College Park MD 20742 USA

AMTA 2016

29 October 2016

Two problems with electronic (XML) dictionaries

- How can you look up a word? Especially when...
 - ▶ You can't spell gud
 - ▶ There's noise in the environment
 - ▶ You can't hear well
 - ▶ It's in a foreign language
- What happens when the dictionary has an error?
 - ▶ Missing information, information in the wrong field, wrongly structured fields
 - ▶ Errors range from typos to missing pages' worth of data
 - ▶ CASL has worked with dozens of XML dictionaries, mostly from government sources; no dictionary has been without errors
 - ▶ Errors are pervasive: ~10,000 errors in one Urdu dictionary (with ~50,000 entries)

- “Did You Mean...?” (DYM)
Smart fuzzy lookup. Three versions:
 - ▶ Basic DYM: Spell correction only
 - ▶ Morphologically Aware DYM (MADYM): Spell correction + morphological parsing
 - ▶ Cross-language DYMs: Basic or MADYM + lookup in multiple dialect/ language/ script dictionaries
 - Waziri Pashto: lookup in 1902 Waziri dictionary, plus backoff lookup in modern Pashto dictionary via sound changes
 - French-in-Arabic: lookup of Arabic script queries in Moroccan Arabic dictionary *and* in French dictionary (with Arabic morphological parsing)
- Dictionary repair tools
 - ▶ ADALT discovers anomalies: rare structures, data that doesn't “fit the mold”
 - ▶ VELMA is an editor specialized for dictionaries and similar documents

ADALT + VELMA and DYM are *tools*; neither replaces human labor

- J.C.R. Licklider, 1960 *Man-Computer Symbiosis*: “Computing machines can do readily, well, and rapidly many things that are difficult or impossible for man, and men can do readily and well, though not rapidly, many things that are difficult or impossible for computers. That suggests that a symbiotic cooperation, if successful in integrating the positive characteristics of men and computers, would be of great value.”
- ADALT and VELMA complement human abilities in error-finding
- DYM enables humans to better understand critical foreign language texts by doing smart fuzzy lookup

Two problems with electronic (XML) dictionaries

- How can you look up a word? Especially when...
 - ▶ You can't spell gud
 - ▶ There's noise in the environment
 - ▶ You can't hear well
 - ▶ It's in a foreign language
 - ▶ *"Did You Mean...?"*
- What happens when the dictionary has an error?
 - ▶ Missing information, information in the wrong field, wrongly structured fields
 - ▶ Errors range from typos to missing pages' worth of data
 - ▶ CASL has worked with dozens of XML dictionaries, mostly from government sources; no dictionary has been without errors
 - ▶ Errors are pervasive: ~10,000 errors in one Urdu dictionary (with ~50,000 entries)

How we got there: DYMs

- 2007: Dr. Anton Rytting (CASL) suggests need for spell correction for dictionary lookup by Arabic language analysts
- 2008: First “Did You Mean...?” for Georgetown Iraqi Arabic dictionary.
- 2009: Second DYM: Urdu, overwhelmingly positive response.
- Later: Additional Arabic dictionaries, gazetteers, 10,001 Arabic names; Pashto, Russian, Ukrainian, Somali, Persian (Farsi), Swahili...
- 2012-2013: Dr. Corey Miller adds Persian morphological parser to Persian DYM to produce first Morphologically Aware DYM (MADYM)
- Later: MADYMs for Arabic, Somali, Swahili, Korean; DYMs for Chinese (pinyin), Dhivehi, Punjabi, Portuguese (Open Street Maps for Rio de Janeiro)
- 2016:
 - ▶ DYM Toolkit for building (part basic of) DYMs for new languages
 - ▶ DYM Platforms for deploying basic DYMs and MADYMs

Electronic Dictionary Lookup

Ordinary dictionary lookup: type in a word, get back definition(s) *if* you spelled it correctly, and you chose the dictionary citation form

Wildcard lookup: For each letter you're uncertain about, type in a '*'
If you hear German [rat], type *ra**.

But this will give you unwanted words: *Rah, Rap, Rank, rar, rau, Rat, and Rad*—whereas only the last two are likely to be what you want.

Regular expression lookup: If you think it might end in a 'd' or a 't', type *ra[d|t]*. Works if you understand regular expressions, and if you know where the mistakes are likely to be, and if you remember what the possibilities are.

DYM: Builds in the knowledge of regular expressions, likely mistakes, possible substitutions (and deletions and insertions), and *how likely* a particular mistake is; and automagically applies this knowledge everywhere in the word where a mistake is possible. Type in *rat*, get back *Rat* and *Rad* (but not the others).

How does a DYM work? outside the box

Urdu Romanization. th = aspirated dental, t = unaspirated dental, Th = aspirated retroflex.

Did You Mean...?

[Version 2.0]



Enter a query in Latin script.

Urdu Alphabet Urdu Latin UseCase Urdu Latin

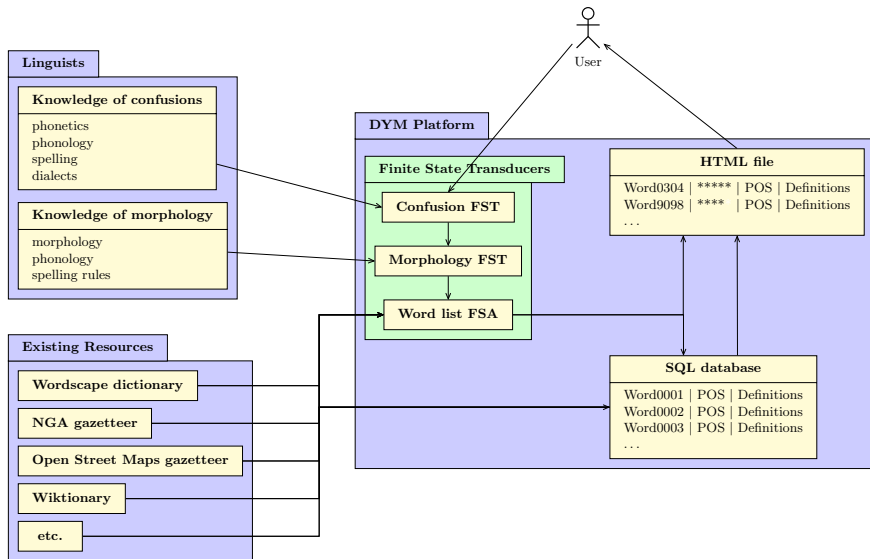
thair - Choose Lexicons - Count: 20 Max Defs: 5 Submit Help

Headword	Target	Urdu	Part of Speech	Grammar Note	Definition	Score	Source
tair	tair	طیر	N.	masculine plural	1. bird	★★★★☆	Kitabistan Urdu Dic
Thir	Thir	ٹھیر	N.	feminine	1. chill	★★★★☆	Kitabistan Urdu Dic
dair	dair	دیر	N.	masculine	1. monastery 2. temple	★★★★☆	Kitabistan Urdu Dic
taa ir	taa ir	طائر	N.	masculine	1. bird	★★★★☆	Kitabistan Urdu Dic
taahir	taahir	طاهر	ADJ.		1. pure; chaste	★★★★☆	Kitabistan Urdu Dic

Proceedings of AMTA 2016, vol. 2: MT Users' Track

Austin, Oct 28 - Nov 1, 2016 | p. 392

How does a DYM work? inside the box



But that's so complicated!

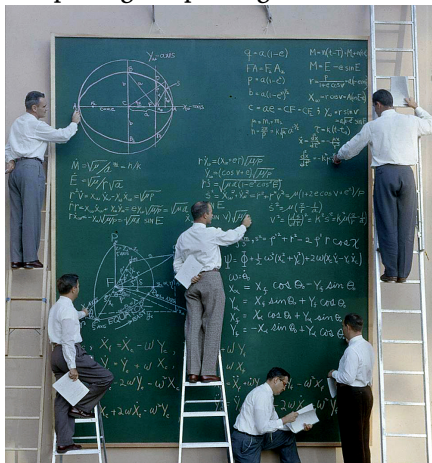
CASL tools to build DYMs:



- DYM Toolkit to incorporate the linguists' knowledge of confusions
 - ▶ DYM Toolkit enables end users (e.g. language analysts) to build confusion matrices for their confusions.
 - ▶ The toolkit can also be used to build confusion matrices for dialects, non-standard spelling systems, native speaker errors.
 - In conjunction with a transliterator, this can be used for Romanized scripts (Arabeze, Penglish/ Farslish,...)
- Dictionaries, gazetteers often exist (but stay tuned for part 2 of this talk)
- DYM Platform takes care of the software infrastructure

But that's so complicated!

Morphological parsing...



...is still rocket science (and optional).

- Platforms 1 and 2, with DYMs, submitted as deliverable (includes documentation and about ten DYMs); USG has free use.
 - ▶ Also available with consulting/ training.
 - ▶ DYM Toolkit included.
- What do I need to build a new DYM?
 - ▶ Dictionary or other lexical resource (e.g. gazetteer)
If it needs cleanup...stay tuned!
 - ▶ Confusion matrix
 - Available for non-native listener confusions for ten languages
 - ...or language analysts can build their own with the provided Toolkit
 - ▶ ...or talk to us about building a new one.

Planned work

- Cross-language DYMs (like Waziri Pashto, French-in-Arabic)
 - CASL has built such DYMs, but current Platforms do not support them
...future work.
- Stand-alone DYMs (without network connection/ server)
 - This could also be used to rapidly field stand-alone dictionaries, even without a 'real' confusion matrix.
- Coming soon to your cellphone?

Two problems with electronic (XML) dictionaries

- How can you look up a word? Especially when...
 - ▶ You can't spell gud
 - ▶ There's noise in the environment
 - ▶ You can't hear well
 - ▶ It's in a foreign language
- What happens when the dictionary has an error?
 - ▶ Missing information, information in the wrong field, wrongly structured fields
 - ▶ Errors range from typos to missing pages' worth of data
 - ▶ CASL has worked with dozens of XML dictionaries, mostly from government sources; no dictionary has been without errors
 - ▶ Errors are pervasive: ~10,000 errors in one Urdu dictionary (with ~50,000 entries)
 - ▶ *ADALT and VELMA*

VELMA: dictionary editor

- VELMA = Visual Environment for Lexicography and MACHine learning



New version = Zelda



- VELMA is designed for editing of existing dictionaries (not building new ones)
 - ▶ Allows easy manipulation of existing XML nodes and text...
 - ▶ ...meaning Ordinary Working Lexicographers (not computer scientists)* can modify dictionaries.
 - ▶ Integrated with ADALT.
 - ▶ Uses ADALT output to suggest a workflow for linguists/lexicographers.
- *Ok, you need to learn about XML...

VELMA in pictures

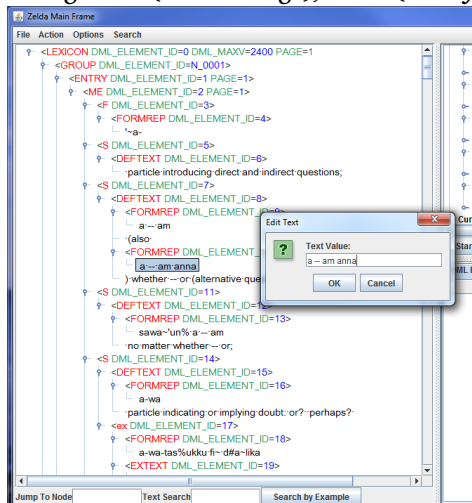
Dictionary file on left, actions performed upper right pane; search results (or ADALT output) lower right pane (here: search for 'camel').

The screenshot displays the VELMA software interface, titled "Zelda Main Frame". The interface is divided into three main panes:

- Left Pane:** A list of dictionary entries. Each entry is a line of text starting with a symbol (like a blue circle or triangle) and followed by a tag and ID, such as "<ENTRY DML_ELEMENT_ID=36884 PAGE=124>". The entry "<EX DML_ELEMENT_ID=36981>" is expanded to show its content: "<ex DML_ELEMENT_ID=36982>" followed by "<FORMREP DML_ELEMENT_ID=36983>", "<EXTTEXT DML_ELEMENT_ID=36984>", and "<EXTTEXT DML_ELEMENT_ID=36985>". The text "he has neither a bleating (sheep) nor a braying (camel)." is highlighted in a grey box.
- Upper Right Pane:** A tree view showing the actions performed on the dictionary file. The path is "/home/dzajic/pandym/arabicWehr/dml/manual.dml". The tree includes "Initial DML Content", "# START ACTION", "# END ACTION", "# START ACTION", "# END ACTION", "# START ACTION", "# END ACTION", "# START ACTION", "# END ACTION", "# START ACTION", "# END ACTION", and "# START ACTION". The actions include "moveElement (290, after, 262)", "removeElement (288)", "removeTail (291)", and "removeElement (292)".
- Lower Right Pane:** A search results pane with two tabs: "Search Result" and "DML Runtime Error". The "Search Result" tab is active, showing a list of search results for the term "camel". The results include: "[TEXT 428] camels", "[TEXT 6842] the moaning bray of a camel", "[TEXT 14178] camel", "[TEXT 18926] saddle, packsaddle (for donkeys and ca...", "[TAIL 19698] the camel kneel down", "[TEXT 23276] girth (of a camel)", "[TEXT 24139] camel", "[TEXT 25193] young camel", "[TEXT 36985] he has neither a bleating (sheep) nor a br...", "[TEXT 41600] wooden post on which camels rub themse...", "[TEXT 43308] the front part of a camel's neck", "[TEXT 43969] slaughter camel", "[TEXT 49019] camel", "[TEXT 49020] camel", "[TEXT 57105] camel saddle".

VELMA in pictures

Supports typical editor operations: search and replace, drag-and-drop, change text (or XML tags), undo (of *any* action, not just most recent)



So what is this ADALT stuff?

How do you find errors in the electronic equivalent of hundreds of pages of a dictionary?

- Errors come in two types:
 - ▶ Frequent: Some common structure is mis-represented throughout
 - Easy for humans to notice at least some of them.
 - ...but finding all instances of an error type can be hard!
 - ▶ Rare: One-offs, typos
 - Needle-in-a-haystack
 - Hard for humans to find these...
 - ...especially when you don't know what kinds of needles there are!

Finding those needles in a haystack

You need...

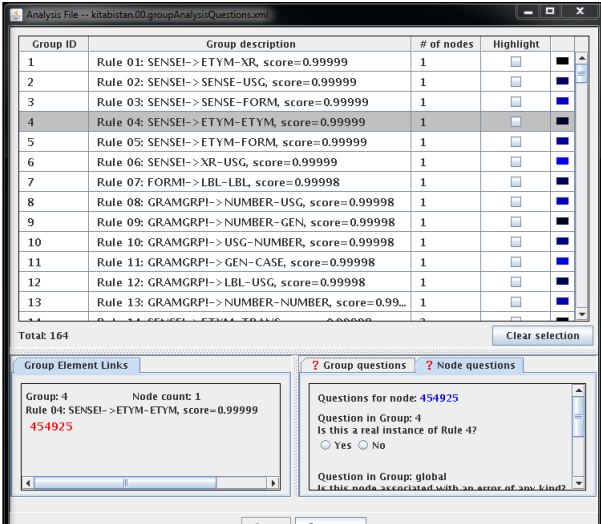


ADALT (Automatic Detection of Anomalies in Lexicographic Text)

- ADALT uses machine learning (AI) to find rare structures
- These are anomalies; some anomalies are errors, some are simply rare structures.
- Examples:
 - ▶ (Spanish, but based on actual anomaly in an Urdu dictionary):
raro *adj.* (Usage: rare)
Missing definition!
 - ▶ (English, but similar anomalies found in Urdu dictionary):
colo(u)r
 - Prevents lookup: user will never search for this spelling!
 - ▶ (Illustrative; actual errors would be in XML)
- ADALT builds models of structures found in a particular dictionary, and uses heuristics to find structures which are rare within those models.
- It therefore doesn't know (and doesn't need to be told) what is 'correct'.

Example of ADALT usage: the metal detector

The metal detector has found anomalies; user chooses one to work on (a single lexeme that has two etymologies):



The screenshot shows the ADALT software interface. At the top, a window titled "Analysis File -- kitabistan.00.groupAnalysisQuestions.xml" contains a table of rules. A red arrow points to the fourth row of the table, which is highlighted. Below the table, the "Group Element Links" section shows details for the selected rule. To the right, the "Group questions" section displays a question for the selected node.

Group ID	Group description	# of nodes	Highlight
1	Rule 01: SENSEI->ETYM-XR, score=0.99999	1	<input type="checkbox"/>
2	Rule 02: SENSEI->SENSE-USG, score=0.99999	1	<input type="checkbox"/>
3	Rule 03: SENSEI->SENSE-FORM, score=0.99999	1	<input type="checkbox"/>
4	Rule 04: SENSEI->ETYM-ETYM, score=0.99999	1	<input type="checkbox"/>
5	Rule 05: SENSEI->ETYM-FORM, score=0.99999	1	<input type="checkbox"/>
6	Rule 06: SENSEI->XR-USG, score=0.99999	1	<input type="checkbox"/>
7	Rule 07: FORMI->LBL-LBL, score=0.99998	1	<input type="checkbox"/>
8	Rule 08: GRAMGRPI->NUMBER-USG, score=0.99998	1	<input type="checkbox"/>
9	Rule 09: GRAMGRPI->NUMBER-GEN, score=0.99998	1	<input type="checkbox"/>
10	Rule 10: GRAMGRPI->USG-NUMBER, score=0.99998	1	<input type="checkbox"/>
11	Rule 11: GRAMGRPI->GEN-CASE, score=0.99998	1	<input type="checkbox"/>
12	Rule 12: GRAMGRPI->LBL-USG, score=0.99998	1	<input type="checkbox"/>
13	Rule 13: GRAMGRPI->NUMBER-NUMBER, score=0.99...	1	<input type="checkbox"/>

Total: 164

Group Element Links

Group: 4 Node count: 1
Rule 04: SENSEI->ETYM-ETYM, score=0.99999
454925

Group questions Node questions

Questions for node: 454925

Question in Group: 4
Is this a real instance of Rule 4?
 Yes No

Question in Group: global
Is this node associated with an error of any kind?

Example of ADALT usage: The needle is found!

VELMA jumps to that anomaly; the first ‘etymology’ contains just a language tag (Persian), the second lacks a language tag.

The screenshot shows the VELMA application window. The main pane displays XML data with various tags like <ENTRY>, <FORM>, <GRAMGRP>, <SENSE>, <ORTH>, <PRON>, <GRAMGRP>, <POS>, <GEN>, <TRANS>, <TR>, <ETYM>, <LANG>, and <MENTIONED>. The tag <SENSE DML_ELEMENT_ID=454925 PAGE=545> is highlighted in green. The right pane shows the 'Attributes of <SENSE>' table:

Name	Value
DML_ELEMENT_ID	454925
PAGE	545

Below the table are buttons for 'Add attribute', 'Delete attribute', and 'Edit a...'. The bottom right pane shows search results for '4:Instance of Rule 04;Rule 04: SENSEI-> ETYM-ETYM, score=0.99999;'. The bottom status bar indicates 'Proceedings of AMTA 2016, vol. 2: MT Users' Track' and 'Austin, Oct 28 - Nov 1, 2016'.

Example of ADALT usage: grab the needle

The user uses VELMA to remove the needle from the haystack (merge the two etymologies)

The screenshot shows the VELMA application window. The main pane displays XML data with a tree view on the left and a text view on the right. The tree view shows a hierarchy of elements including <FORM, <GRAMGRP, <SENSE, <ORTH, <PRON, <GRAMGRP, <POS, <GEN, <TRANS, <TR, <ETYM, <LANG, <MENTIONED, and <ENTRY. The text view shows the XML content for the selected element, including the word 'تپ لوزہ' and its transcription 'ta'p-e lar'zah'. The right pane shows the 'Attributes of' table, which is currently empty. Below it, the 'DML' tab is active, showing a log of actions performed on the document, including 'moveElement' and 'removeElement' operations.

Attributes of

Name	Value
------	-------

Add attribute Delete attribute Edit a

DML DML errors Notes Search result

```
*** SESSION START ***
# START ACTION Tue, Oct 18, 2016, 15:07:
moveElement (454937, after, 454935)
# END ACTION

# START ACTION Tue, Oct 18, 2016, 15:07:
removeElement (454936)
# END ACTION

# START ACTION Tue, Oct 18, 2016, 15:07:
removeElement (454936)
# END ACTION
```

...about those frequent errors

Latest version of VELMA (Zelda!) provides “query-by-example” (QBE) and “edit-by-example” (EBE)

- If the user notices multiple of instances of some structural error...
 - Dozens might be evident on inspection
 - ...and hundreds more might be lurking.
- ...then use QBE to find all instances of that structure (or similar structures).
- Once the user has verified that QBE finds the desired erroneous structures, correct one, then use EBE to correct others in the same way.

ADALT and VELMA software

- Mature: CASL has used this (or previous versions) on about a dozen dictionaries.
- VELMA is preferable to text editors/ Perl/ version control systems because...
 - ▶ Users are less likely to make mistakes
 - ▶ ...and if you do make a mistake, it can be undone six months from now without messing up everything done between now and then.
 - ▶ Provides audit trail of changes.
 - ▶ Users report working faster.
- Submitted as deliverable (includes documentation); USG has free use.
 - ▶ Also available with consulting/ training.
- What can I do with this?
 - ▶ Import/ clean up dictionaries
 - ▶ Import/ clean up gazetteer data
 - ▶ Potentially useful for other kinds of semi-structured text data

Summary

- DYMs and dictionary repair started as CASL research projects less than ten years ago.
- Many useful tools and resources (cleaned up dictionaries, DYMs) came out of the work along the way.
- Dictionary repair and building DYMs are now (mostly) engineering.
 - ▶ Fielding a new DYM can be done in a few days.
 - ▶ Repairing a dictionary can take a few months (rather than a year or more)
...and the output is of higher quality than CASL's early dictionaries.
- These tools help—not replace—human beings.
- Contact information:
 - ▶ Dr. Mike Maxwell (Technical Director for HLT):
mmaxwell@casl.umd.edu 301-356-2639
 - ▶ Dr. David Zajic (Research Scientist): dzajic@casl.umd.edu
301-356-8995
 - ▶ Dr. Mike Bunting (CASL Executive Director):
mbunting@casl.umd.edu 301-356-8894

- Arabic (Iraqi, MSA, Sudanese, Levantine...; dictionaries, gazetteers, 10,001 Arabic names); *French in Arabic (code switching)
- Urdu, Pashto, *Waziri Pashto, Punjabi
- Russian, Ukrainian
- Somali, Maay (related to Somali), Swahili, Chimwiini (related to Swahili)
- Persian (Farsi, Dari)
- Korean
- Mandarin Chinese (pinyin search)
- Dhivehi

*Not available in current Platforms



Principle-Based Preparation of Authentic Bilingual Text Resources

Michelle Vanni, Ph.D.

Army Research Laboratory



- 1. The context: Speech recognition for military**
- 2. The research questions: Where does the material fit?**
- 3. The problem: Material and task description**
- 4. The principles: Constraints on organization**
- 5. The examples: What you would do & why**



OPERATIONAL ACCENTED SPEECH ADAPTATION INITIATIVE

Vision and Objective

- Automated Speech Recognition (ASR) technology trained on authentically accented data for operations
- High quality ASR for military-relevant languages spoken in operational scenarios
- Algorithms adapting general purpose ASR technology to military operational needs



Problems Being Addressed

- Algorithms to adapt ASR to new types of variation
- Expeditionary Force: local populace & coalition partners
- Army Challenges
 - 1 Situation Awareness: Adversary intent & capabilities
 - 2 Security Force Assistance: HQ ASR for effectiveness

Impact

- Understand foreign media and captured document content
- HLT-equipped soldiers: Train & serve with coalition partners
- Focus: Variations of high military, low commercial value



- *Given the modest amounts of **bilingual in-domain speech data** available, which approaches to Automatic Speech Recognition (ASR) adaptation have the most impact on language modeling for Army-centric technologies?*
- *Can ASR software components and algorithms be trained to achieve **better performance** with African-Accented speakers?*
- *Is it possible to generalize—and to what extent—ASR adaptation algorithms designed to address individual speaker differences, over sets of non-native pronunciations present in communities?*

Strategy:

1. Improve techniques for ASR adaptation on **bi-modal—speech-text aligned--accented data**.
2. Algorithms for low resource languages, dialects and accented variations.
3. Assess for **general purpose tech** to process the speech of African accented **high-resource languages**

Technical Barriers:

1. Valuable on individual non-native variations, maximum likelihood linear regression (MLLR) & maximum a posteriori (MAP) adaptation likely improve high-resource language ASR on similar variations in accented speakers, **e.g., French and accented French: However experimentation for specific Army operations is required.**
2. Morphological analysis improves performance of translation for low resource languages. Methods require extensive training of bilingual humans and are not cost-effective.
3. French an official language in 21 of 54 African countries. Phonetic variation in African French clusters around national accents. Assessment of speech recognition accuracy for African-accented speech needed to support operations.

Approach:

1. Experiment with language modeling software offering Deep Neural Network technology on compiled **parallel aligned data sets** for low resource languages **of military interest.**
2. Test a new unsupervised morphological analyzer on Pashto data compiled in **domains of military interest.**
3. Use speech data collected in Cameroon and Gabon to test accuracy of a **French** speech recognizer with one type of African-accented speech. Adapt the French speech recognizer with a modest amount of Cameroon-accented French and compare accuracy using Word Error Rate.



- Lots of raw data in an operational format

- In this case, Power Point slides
 - Could be bilingual web data or digitized books

- Bi-text needed for multiple purposes
 - Speech recognition pronouncing dictionary
 - Machine translation domain adaptation
 - Glossaries and Translation Memory

- First Step: Change the format!
 - Find a suitable editing environment
 - In this case, MS Excel worked fine



A SAMPLE OF THE TEXT EDITING PRINCIPLES:

1. **Each row** represents a single semantic unit, word, phrase or clause and can be simple or complex.

2. **Punctuation:**
 - a. Periods only after full clauses, with or without grammatical subjects and where appropriate by convention.
 - b. Commas as appropriate
 - c. Usually delete colons, except when required on the basis of content.

3. **Capitalization:** As appears in raw text for application-specific pre-processing

.... **AND SO ON**

Proceedings of AMTA 2016, vol. 2: MT Users' Track

Austin, Oct 28 - Nov 1, 2016 | p.418



A SAMPLE OF THE TEXT EDITING PRINCIPLES:

4. **Insertion:** Without substitution of original material, syntactic support structures for creation of a corpus usable for training machine translation of genres other than the genre of provenance.

5. **Insertion:** Conventionally accepted orthographic forms, without substitution of forms presented in original data.

6. **Insertion:** Without substitution of original material, of semantically accurate and similarly structured translation, when given dynamic equivalent rendering is structurally divergent.

.... AND SO ON



U.S. ARMY
RDECOM

EXAMPLES

ARL

EEIFA

les éléments essentiels d'information des forces amies

ce que nous voulons cacher de la menace

la SECOP

COA

cause de l'action

zone géographique

où l'information pour répondre à un EIP peut être recueillie

zone géographique où l'information pour répondre à un EIP ou confirmer / refuser un COA de la menace peut être recueillie

EEFI

essential elements of friendly information

what we want to hide from the threat

OPSEC

COA

cause of action

geographic area

where information to answer a PIR can be collected

geographical area where information to answer a PIR or confirm/deny a threat COA can be collected



Questions?

Contact us at:

michelle.t.vanni.civ@mail.mil

Machine Translation of Canadian Court Decisions

L. Langlois, M. Simard, E. Macklovitch

AMTA 2016

Austin TX, October 2016



*This presentation is
dedicated to Lucie Langlois*

*Nous dédions cette présentation
à Lucie Langlois*

Courts Administration Service

- Created in 2003 to rationalize services offered to four Canadian tribunals:
 - Federal Court of Appeal, Federal Court, Court Martial Appeal Court, Tax Court of Canada
- Responsible for meeting courts' administrative needs & ensuring public access to all court records & decisions
- OLA: court decisions *must* be published in both official languages

Federal Court



Cour fédérale

Date: 20151127

Docket: T-575-15

Citation: 2015 FC 1323

Ottawa, Ontario, November 27, 2015

PRESENT: The Honourable Mr. Justice Locke

BETWEEN:

ALCON CANADA INC.,
ALCON LABORATORIES, INC.,
ALCON PHARMACEUTICALS LTD.,
and ALCON RESEARCH, LTD.

Plaintiffs/
Defendants by Counterclaim

and

ACTAVIS PHARMA COMPANY

Defendant/
Plaintiff by Counterclaim

ORDER AND REASONS

I. Overview

[1] The plaintiffs, Alcon Canada Inc., Alcon Laboratories, Inc., Alcon Pharmaceuticals Ltd., and Alcon Research, Ltd. (collectively referred to as Alcon), appeal from an Order of Prothonotary Martha Milczynski dated September 24, 2015, dismissing Alcon's motion to strike

CAS and Translation

- CAS responsible for ensuring the timely translation & publication of all court decisions
 - approx. 8 million words/year; mostly Eng > Fr
 - all outsourced; revised internally by jurilinguists
 - requirement for high quality; both linguistic versions have equal force before the law
 - simultaneous publication on Web
- Long translation delays; traditional workflow unable to cope

A pilot project in MT

- Launched at the initiative of L. Langlois, DG, Judicial Services
 - extensive experience in NLP and translation
 - in her view, solution could only come from MT
- LL contacts NRC re: MT pilot (early 2015)
 - contacts EM to act as independent consultant
 - imposing translation workload, but CAS has assets
 - NRC begins by analysing available corpora

NRC at CAS

- Main deliverable: provide CAS with the best possible Machine Translation
- Strategy: Build specialized MT engines for each of the four tribunals
- MT technology: NRC's Portage
 - Phrase-based MT technology
 - Continuous development since 2004
 - Participated in numerous shared tasks: WMT, NIST, etc.
 - Commercially available since 2010

Building Specialized Engines

General procedure for building specialized MT:

- Collect domain translations
- Process corpus
- Train engines
- Test and evaluate
- Repeat



Collecting CAS Data

- Historically, all translation was outsourced
→ no structured Translation Memory (TM)

However...

- All CAS court decisions are on the Web since the mid-1990s
→ all decisions of the last 20 years available in HTML format

Collecting CAS Data

		CMAC	FC	FCA	TCC	Total
Documents	Paired	142	25.3k	6.5k	8.8k	40.7k
	Orphan	1	1.9k	348	561	2813
TU's		28k	3.4M	888k	1.8M	6.6M
Words	EN	600k	89M	17M	35M	141.6M
	FR	600k	103M	19M	41M	163.6M

}	CMAC	=	Court Martial Appeals Court	}
	FC	=	Federal Court	
	FCA	=	Federal Court of Appeal	
	TCC	=	Tax Court of Canada	

CAS Data Analysis: Linguistic Complexity

Corpus	Court	Type-Token Ratio (@100k words)	Growth Rate (@100k words)	BLEU
References	“Rich”	0.141	1/13	33.2
	“Medium”	0.109	1/18	42.9
	“Poor”	0.078	1/26	50.7
	Weather Reports	0.018	1/200	↑
CAS	CMAC	0.079	1/29	?
	FC	0.103	1/19	
	FCA	0.101	1/19	
	TCC	0.094	1/20	

CAS Data Analysis: Translation Memory

TM coverage (% source words)

Court-specific TMs	Court	70%+	85%+	Exact
	CMAC	7.2	5.9	3.9
	FC	13.6	11.6	9.0
	FCA	11.9	10.3	7.7
	TCC	12.5	9.9	6.8

Global TM	Court	70%+	85%+	Exact
	CMAC	8.4	6.8	4.5
	FC	13.9	11.9	9.3
	FCA	16.5	14.8	10.8
	TCC	13.1	10.7	7.2

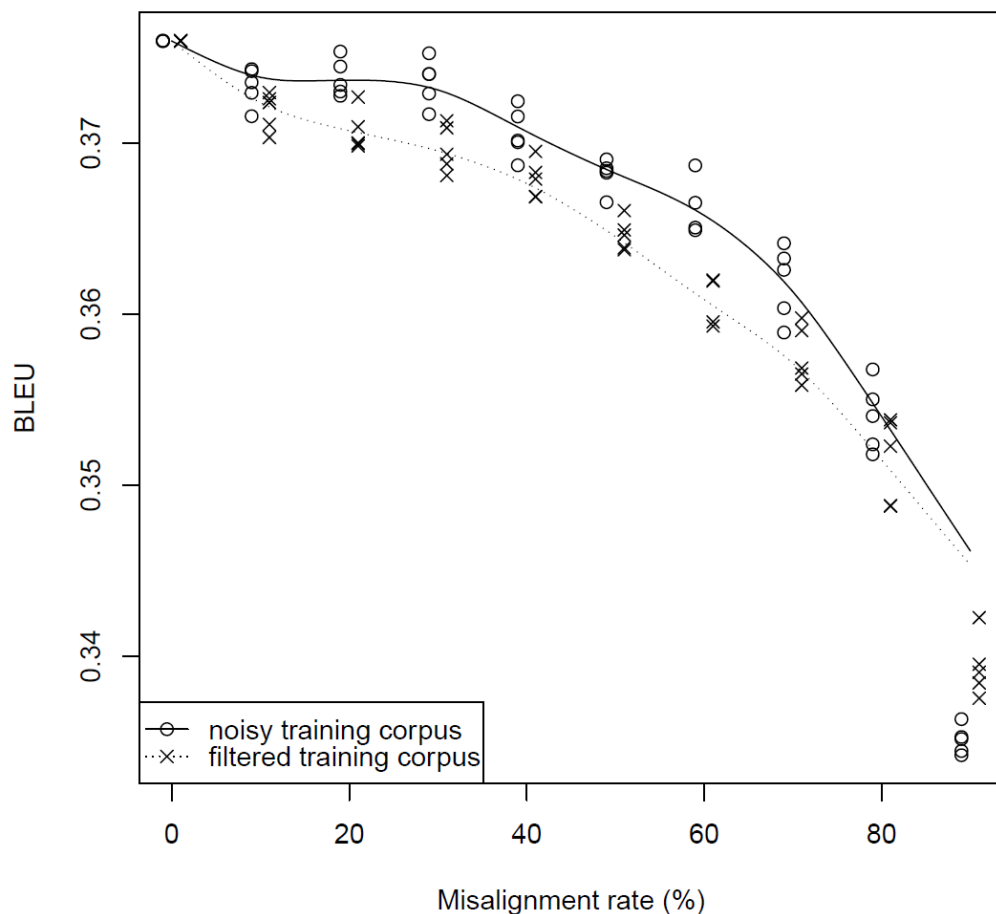
Processing CAS Data

A 3-step process:

1. Pair up documents
 2. Extract text, segment (into translation units), normalize
 3. Align segments
- Initially done using NRC tools
 - Recently: *AlignFactory* (Terminotix)

Sentence (mis)alignment

Impact of noise on BLEU



- SMT highly tolerant to “noise” in alignment

[Cyril Goutte, Marine Carpuat, George Foster (2012).

[The Impact of Sentence Alignment Errors on Phrase-Based Machine Translation Performance.](#)

AMTA 2012]

True only when noise is “uniform”!

Sentence (mis)alignment

Court	Basic Alignment		Improved Alignment	
	Accuracy (%)		Accuracy (%)	
CMAC	89.5		96.5	
FC	89.0		93.5	
FCA	88.5		97.0	
TCC	90.0		99.5	

- To measure alignment accuracy: sample 100 random pairs (A,B), assign labels:

Label	Description	Accuracy
Good	A is a translation of B	1
Partial	Part of A is a translation of part of B	½
Bad	A not a translation of B	0
Unusable	Something is weird	0

Sentence (mis)alignment

Court	Basic Alignment		Improved Alignment	
	Accuracy (%)	MT (BLEU)	Accuracy (%)	MT (BLEU)
CMAC	89.5	40.4	96.5	41.2
FC	89.0	46.5	93.5	49.3
FCA	88.5	42.7	97.0	47.1
TCC	90.0	44.0	99.5	47.1

- Obviously, alignment errors are not “uniform”
 - systematic bias is hurting quality of MT
 - Better alignments mean Portage has more “meaningful” data to learn from

Data Filtering

- Untranslated quotation in text
→ same language in both versions

[5] Dans sa décision, la Commission rappelle d'abord les circonstances ayant entouré le double meurtre commis par le demandeur et ce qui a pu pousser celui-ci à les commettre, circonstances qu'elle décrit de la manière suivante :

The victims were 15 and 17 year-old adolescents. The youngest victim was the brother of your ex-girlfriend and the other victim was one of his friends. On February 28, 1989, you entered the residence of the youngest victim and hid in the basement with a loaded rifle. When the two boys arrived after school, you shot both of them in the head. Each victim was shot twice. They were murdered in cold blood, with planning and deliberation.

[...]

According to your file, those violent crimes were committed in

Data Filtering

- Bilingual quotation in both versions of text

[3] The Tax Court found that the lump sum payment was properly included in Mr. Butler's income under paragraph 56(1)(v) of the *Income Tax Act*, R.S.C. 1985, c. 1 (5th Supp.) [the *ITA*]. This paragraph provides:

56. (1) Without restricting the generality of section 3, there shall be included in computing the income of a taxpayer for a taxation year,	56. (1) Sans préjudice de la portée générale de l'article 3, sont à inclure dans le calcul du revenu d'un contribuable pour une année d'imposition :
---	--

[...]

[...]

(v) compensation received under an employees' or workers' compensation law of Canada or a province in respect of an injury, a disability or death;	v) une indemnité reçue en vertu d'une loi sur les accidents du travail du Canada ou d'une province à l'égard d'une blessure, d'une invalidité ou d'un décès;
--	--

[4] The Tax Court held that the term "in respect of an injury" is to be broadly interpreted to mean all amounts paid in relation to a compensable injury, relying on the

Data Filtering

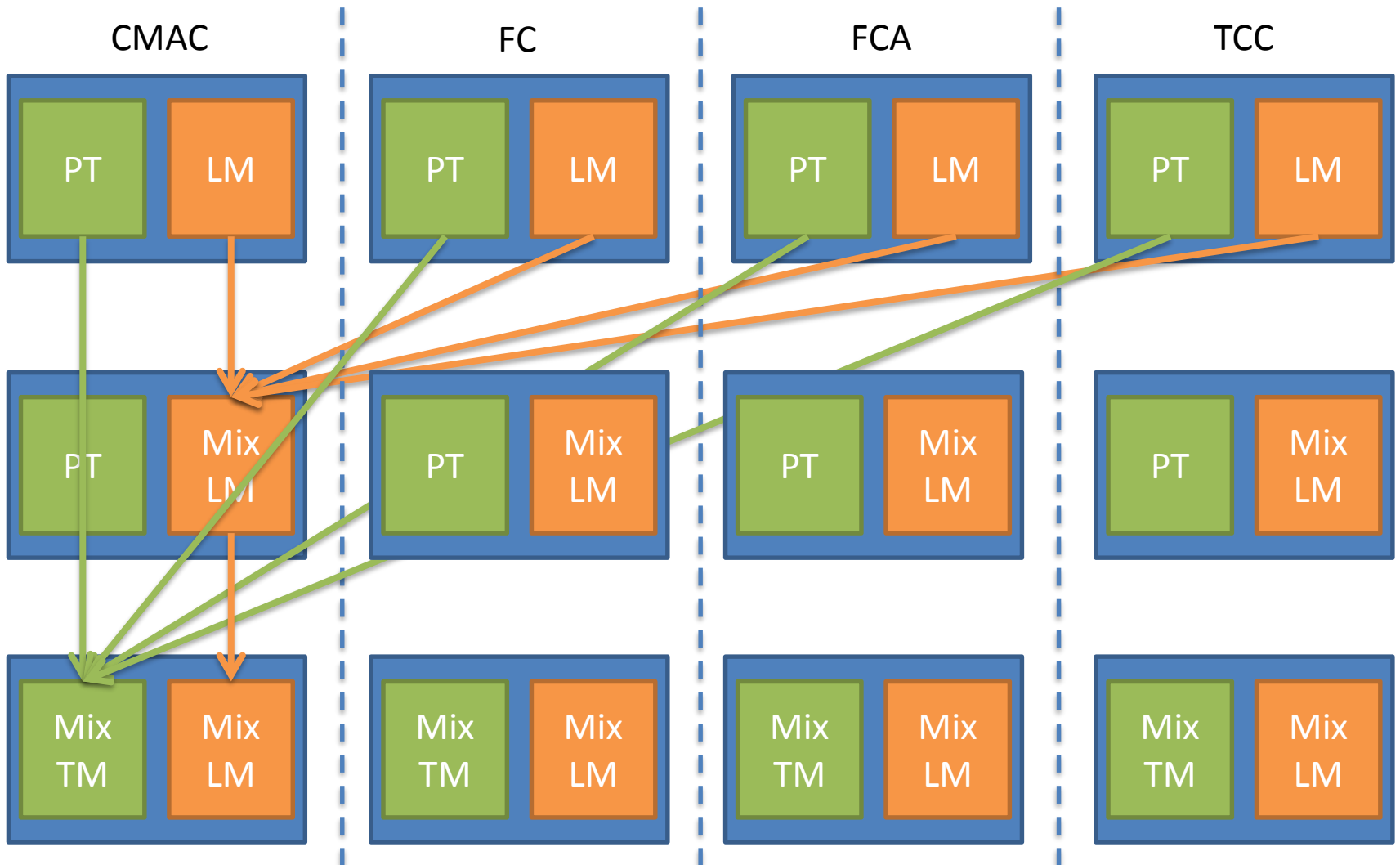
- We applied a simple filter, based on short lists of frequent French and English words
- Filters out 1-5% of training data
- BLEU gains between 0.3 and 1.5

Court	Unfiltered	Filtered
CMAC	41.5	41.8
FC	49.0	50.4
FCA	47.4	48.9
TCC	47.3	47.6

System Combinations

- Portage allows different combination strategies
- Best results are obtained with *mixture models*, that assign different weights to each component, to optimize performance on a certain type of text
 - mixLM: mixture target language model
 - mixTM: mixture translation model (“phrasetable”)

System Combinations



System Combinations

Court	Baseline	+ mixLM	+ mixLM + mixTM
CMAC	41.8	43.1	45.4
FC	50.4	50.0	50.3
FCA	48.9	49.5	51.7
TCC	47.6	47.6	47.8

- CMAC & FCA benefit the most from combinations
 - CMAC is small
 - FCA is very much related to matters in FC
- No clear benefit for FC and TCC
 - FC is much larger than other domains
 - TCC is probably distinct
- **Combinations never (significantly) hurt performance**

Final Systems

Court	EN→FR	FR→EN
CMAC	45.4	46.5
FC	50.3	52.6
FCA	51.7	54.4
TCC	47.8	52.0

- Gains relative to initial baseline systems range from **+3.8** (FC & TCC) to **+9.0** (FCA)

Integrating Portage in Matecat

- Needed a translation environment that allows MT to be integrated with translation memory
 - no TenT being used at CAS
- Matecat: a cloud-based CAT system
 - product of EU FP7 aimed at minimizing PE time
 - advantage for CAS: requires no local infrastructure or computer support; accessible everywhere
 - Matecat is free! Perfect for a pilot project
 - allows integration of different MT systems

Pilot Project at CAS

- To what extent can Portage help TRs increase productivity and decrease turnaround times?
- Two translation students hired for summer
 - pro: enthusiastic & open to technology
 - con: little experience in legal translation
- Translations carefully revised by professionals before publication
- Compare translation times with/without MT

Pilot Project Framework

- Focus on immigration decisions
- Statistics obtained from onsite coordinator
 - total no. texts/ words translated by each student
 - no. of texts with/without MT
 - productivity with/without MT
- Follow-up training provided to students
 - feedback obtained from two revisers
- Trial began on 11 May and ended on 25 August

The Results

Translator	<u>No MT</u>			<u>With MT</u>			diff. # words/hr	gain +MT vs. -MT
	total # texts	total # words	avg. # words/hr.	total # texts	total # words	avg. # words/hr.		
ADB	14	19,998	238	85	77685	373	135	57%
AL	19	20,538	291	109	86,918	390	99	34%

- Results compare very favorably with legal translators currently handling CAS decisions

Trial Results (cont'd)

- Caveats:
 - TRs didn't have access to a complete TM; only the one created as they translated. Some of gain attributed to MT would normally come from TM
 - We should have recorded revision times to ensure +TM texts didn't require more revision
- Still, no doubt that student TRs benefited substantially from Portage input
 - revisers report surprising errors in non-MT

Room for improvement

- Matecat had its problems:
 - handling intricate formatting; not always parallel in English & French
 - reintegrating results of spelling & grammar checking
 - lack of flexibility in revision mode
- Portage had its problems
 - handling named entities, i.e. knowing when and when not to translate these NPs
 - surprising number of errors of grammatical agreement, particularly in E > F direction

Discussion

- Recall: these students had no prior experience in legal translation
 - yet with the help of MT, in a few short months...
 - aided by Portage's acquisition of terms & phrases that are common in court decisions
- What makes these court decisions such a good application for MT
- Future plans

Thank you for your attention!

Any questions?



CENTER FOR ADVANCED
STUDY OF LANGUAGE

JOHNS HOPKINS
UNIVERSITY



human language technology
center of excellence

Putting the "human" back in HLT:

The importance of human evaluation in assessing the quality and potential uses of translation technology

Erica Michael & Petra Bradley

University of Maryland CASL

Paul McNamee & Matt Post

Johns Hopkins University HLT COE



Outline

- **Part 1:** Human comprehension of machine translation (MT) output
- **Part 2:** Use of MT for translation and comprehension of Chinese texts
- **Part 3:** Utility of translation memory (TM) in an operational context





CENTER FOR ADVANCED
STUDY OF LANGUAGE

JOHNS HOPKINS
UNIVERSITY



human language technology
center of excellence

Human comprehension of MT output: Findings from 2016 SCALE

SCALE: Summer Camp for Applied Language Exploration

SCALE 2016a: Knowledge Rich Statistical Machine Translation

Approaches to MT

Rule-based	Human experts create rules
Example-based	Sentences are matched against previous translations
Statistical	Phrase translations are learned from many examples
Deep Learning	Neural nets are trained from many examples



Features of each approach

Rule-based

- Rules are composed by language experts
- Performs a deep source language analysis
- Easy to update, adapt to new domains
- Very fast

Statistical

- Learns automatically from example translations
- Doesn't require language-specific knowledge
- Leverages Big Data



Workshop questions

- How do the **Rule-based** and **Statistical** paradigms compare in terms of translation quality?
- Can we improve translation quality by combining them?



Rule-based

Human constructed
Knowledge-rich

Statistical

Learned automatically
Generic
Language-agnostic
Commercially dominant

Hybrid

Best of both worlds ?

Workshop themes

- Explore **Hybrid MT** *in a number of languages*
- *Augmenting* SMT using linguistic information from Rule-based MT
- Evaluate *both* intrinsic MT quality and document comprehensibility
- Make *engineering changes* to support fast updates, easy adoption

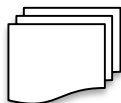


Statistical translation

Translation Model (learned from parallel texts)



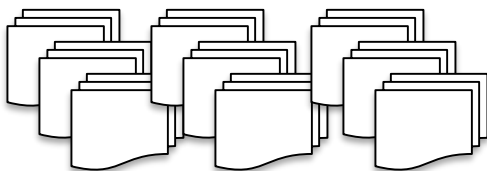
'Это было ошеломляющее зрелище' –
сказал Керри после посещения парка.



'It was a stunning sight' - Kerry said after
visiting the park.

$p(\text{канада} \mid \text{canada}) = 0.7$
 $p(\text{канада} \mid \text{canadian}) = 0.1$
 $p(\text{канада} \mid \text{montreal}) = 0.1$
 $p(\text{украине} \mid \text{ukraine}) = 0.6$

Language Model (learned from text)



$p(\text{"the eyes of texas"}) = \text{high}$
 $p(\text{"eyes texas the of"}) = \text{very low}$
 $p(\text{"like to eat crabs"}) = \text{high}$
 $p(\text{"like to eat forks"}) = \text{low}$



Used Open Source Apache Joshua during workshop:
joshua.incubator.apache.org



Hybrid approaches - I

- Dictionary Extraction
 - Add translation pairs from RBT lexicon to augment the SMT translation model
 - Pros: direct; can be done once; reduces OOVs
 - Cons: requires lots of morphological analysis to accurately convert lexicon base forms to the surface forms needed in SMT



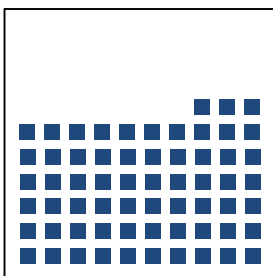
Hybrid approaches - II

- “Black Box”
 - Apply RBT to a set of foreign sentences. Use generated translations as supplementary training data (“imperfect bitext”)
 - Pros: simple; can take advantage of additional RBT processing (e.g., pre- or post-corrections, transliteration of unknown words)
 - Cons: possibly adding errorful training data

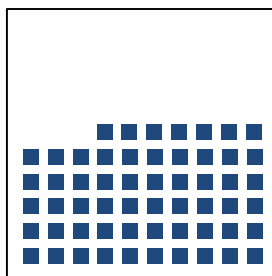


Training data

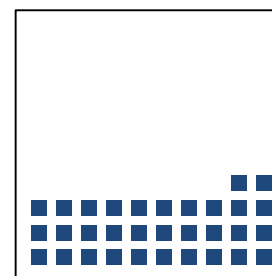
Higher Resource



Russian (63)

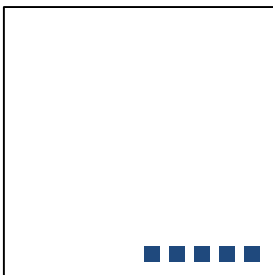


Arabic (57)

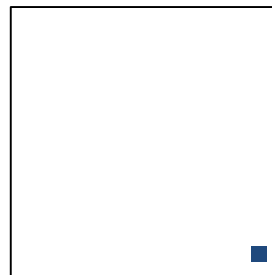


Portuguese (32)

Lower Resource



Farsi (5)



Swahili (0.2)

■ = 1 million sentences
with English translations



Two evaluations

- Automatic Metrics
 - Given reference translations, compute scores that characterize the fidelity and fluency of MT system output
- Human Evaluation
 - More directly measures quality and the ability to use MT to support analyst workflows



Motivation for human evaluation

- **Research questions**

- How well can human users comprehend machine-translated text?
- Are differences in BLEU scores reflected in human comprehension scores?

- **Goals for implementing the study**

- Materials: as authentic as possible
- Task: similar to some of the ways in which analysts might use MT output



Design overview

- **5 languages**
 - Arabic, Farsi, Portuguese, Russian, Swahili
 - Range of morphological complexity and resources
- **4 translation types**
 - **Human**, **Rule-based**, **Statistical**, **Hybrid**
- **16 passages per language**
 - For most languages, 8 News + 8 Conversation
- **2-6 questions per passage**
 - Topic, Main Idea, Fact, Inference
 - Questions based on human translation



Materials

- **Arabic & Farsi**

- Created by Doug Jones (MITLL)
- All news items
- Similar to Voice of America and BBC items
- Originally in English, translated to Arabic and Farsi
- Average number of words per passage = 340.8
- 2 or 3 questions per passage



Materials

Portuguese Russian Swahili

News			
Sources	Global Voices, BBC	Voice of America	BBC
Original language	English	English	Mostly English
Avg. # of words	258.4	308.8	138.4
Conversation			
Sources	Global Voices	Global Voices, course materials	JSPS Global COE, swahiliweb.net
Original language	Portuguese, English, other	Russian	Swahili
Avg. # of words	221.4	224.1	232.7



Types of questions

Similar to triage tasks

Topic	Main Idea
<p data-bbox="112 604 571 743">The topic of this passage is:</p> <ul data-bbox="112 792 413 1078" style="list-style-type: none"><li data-bbox="112 792 413 846">a. Polygamy<li data-bbox="112 868 390 922">b. Marriage<li data-bbox="112 943 355 998">c. Divorce<li data-bbox="112 1019 374 1073">d. Children	<p data-bbox="610 604 1765 735">Which of the following is the best title for this article?</p> <ul data-bbox="610 792 1804 1078" style="list-style-type: none"><li data-bbox="610 792 1804 846">a. Obama promotes voting on American Idol finale<li data-bbox="610 868 1634 922">b. Obama performs on American Idol finale<li data-bbox="610 943 1750 998">c. Obama asks for votes on American Idol finale<li data-bbox="610 1019 1410 1073">d. American Idol comes to an end



Types of questions

More relevant for finding essential elements of information

Fact	Inference
<p data-bbox="112 596 817 839">According to the IMF, what would increase indebtedness for emerging market economics?</p> <ul data-bbox="112 896 801 1310" style="list-style-type: none">a. High budget deficitsb. Increasing 50% of their gross domestic productc. Strong revenues with slow growthd. World government debt	<p data-bbox="852 596 1711 718">What does Person A imply about Kennedy and Krushchev?</p> <ul data-bbox="852 896 1802 1310" style="list-style-type: none">a. They were prudent and averted a war.b. They met because of the Caribbean Crisis.c. They caused the Caribbean Crisis.d. They could have worked together to avoid catastrophe.



Pilot testing

- **Guessability**

- Tested questions without passages to ensure correct responses could not be guessed
 - Criterion: $\leq .70$

- **Difficulty**

- Tested questions with human-translated passages to ensure questions were not too difficult
 - Criterion: $\geq .50$



MT examples

Portuguese

À medida que as pessoas envelhecem, os relógios biológicos começam a voltar a despertar mais cedo,

Human Translation

As people age, the biological clock starts waking up earlier again,

Rule-based MT

while the people age, the biological clocks begin to return to awaken earlier,

Statistical MT

As people age, the biological clocks are beginning to return to awaken sooner,

Hybrid MT

While people age, the biological clocks are beginning to return to wake up earlier



MT examples

Swahili

A: Labla Marekani katika jimbo gani? Marekani ni kubwa.

Human Translation

A: To be more precise, which state in America? America is vast.

Rule-based MT

A: LABLA America in/at what region? America is big.

Statistical MT

A: “Maybe America in what state? The United States is the greatest.

Hybrid MT

A: Maybe America in what region? The United States is big.



MT examples (with sample question)

According to the story, in what way are planes and trains alike?

- a. They both rarely have accidents.
- b. They both offer tea.
- c. They both have good window seats.
- d. They both have duty free.

Russian

В: Но на поезде почти никогда не бывает аварий...

А: Самолёты тоже падают очень редко.



MT examples (with sample question)

According to the story, in what way are planes and trains alike?

- a. They both rarely have accidents.
- b. They both offer tea.
- c. They both have good window seats.
- d. They both have duty free.

Rule-based MT

B: but on the train almost never is wrecks...

A: aircraft also fall very far-between

Russian

B: Но на поезде почти никогда не бывает аварий...

A: Самолёты тоже падают очень редко.



MT examples (with sample question)

According to the story, in what way are planes and trains alike?

- a. They both rarely have accidents.
- b. They both offer tea.
- c. They both have good window seats.
- d. They both have duty free.

Rule-based MT

B: but on the train almost never is wrecks...

A: aircraft also fall very far-between

Stat MT / Hybrid MT (identical)

B: but the train is almost never crashes...

A: the planes also fall very rarely.

Russian

B: Но на поезде почти никогда не бывает аварий...

A: Самолёты тоже падают очень редко.



MT examples (with sample question)

According to the story, in what way are planes and trains alike?

- a. They both rarely have accidents.
- b. They both offer tea.
- c. They both have good window seats.
- d. They both have duty free.

Rule-based MT

B: but on the train almost never is wrecks...

A: aircraft also fall very far-between

Stat MT / Hybrid MT (identical)

B: but the train is almost never crashes...

A: the planes also fall very rarely.

Russian

B: Но на поезде почти никогда не бывает аварий...

A: Самолёты тоже падают очень редко.

Human Translation

B: But by train there's almost never an accident...

A: Planes fall very rarely too.



MT examples (with sample question)

What portion of Mrs. Obama's speech is highlighted in Lines 1-2?

- Her belief that obese children can become seriously ill.
- A program to raise funds for terminally ill Americans.
- The need to help millions of children without health care.
- The NAACP's campaign to improve US medical care.

Arabic

والتي أطلقتها في فبراير الماضي إلى تسليط ضوءاً مستمراً على مرض سمنة "هيا نتحرك" وتهدف حملة السيدة أوباما الأطفال في الولايات المتحدة و أيضاً على ملايين الشباب المعرضين لخطر الإصابة بأمراض خطيرة ذات صلة بالسمنة.



MT examples (with sample question)

What portion of Mrs. Obama's speech is highlighted in Lines 1-2?

- Her belief that obese children can become seriously ill.
- A program to raise funds for terminally ill Americans.
- The need to help millions of children without health care.
- The NAACP's campaign to improve US medical care.

Arabic

والتي أطلقتها في فبراير الماضي إلى تسليط ضوءاً مستمراً على مرض سمنة "هيا نتحرك" وتهدف حملة السيدة أوباما الأطفال في الولايات المتحدة و أيضاً على ملايين الشباب المعرضين لخطر الإصابة بأمراض خطيرة ذات صلة بالسمنة.

Rule-based MT

2 And aim(s) campaign/download_it
Mrs/lady Obama ``come on, we move``
and that launched her last February to
focusing light continuing on disease
fat(ness)/Samnah the children in the
United States and also on the millions of
guys/al-Shabab the exhibitions to the
injury danger with illnesses dangerous
related with the fatness.



MT examples (with sample question)

What portion of Mrs. Obama's speech is highlighted in Lines 1-2?

- Her belief that obese children can become seriously ill.
- A program to raise funds for terminally ill Americans.
- The need to help millions of children without health care.
- The NAACP's campaign to improve US medical care.

Arabic

والتي أطلقتها في فبراير الماضي إلى تسليط ضوءاً مستمراً على مرض سمنة "هيا نتحرك" وتهدف حملة السيدة أوباما الأطفال في الولايات المتحدة و أيضاً على ملايين الشباب المعرضين لخطر الإصابة بأمراض خطيرة ذات صلة بالسمنة.

Rule-based MT

2 And aim(s) campaign/download_it
Mrs/lady Obama `come on, we move`
and that launched her last February to
focusing light continuing on disease
fat(ness)/Samnah the children in the
United States and also on the millions of
guys/al-Shabab the exhibitions to the
injury danger with illnesses dangerous
related with the fatness.



MT examples (with sample question)

What portion of Mrs. Obama's speech is highlighted in Lines 1-2?

- Her belief that obese children can become seriously ill.
- A program to raise funds for terminally ill Americans.
- The need to help millions of children without health care.
- The NAACP's campaign to improve US medical care.

Arabic

والتي أطلقتها في فبراير الماضي إلى تسليط ضوءاً مستمراً على مرض سمنة "هيا نتحرك" وتهدف حملة السيدة أوباما الأطفال في الولايات المتحدة و أيضاً على ملايين الشباب المعرضين لخطر الإصابة بأمراض خطيرة ذات صلة بالسمنة.

Rule-based MT

2 And aim(s) campaign/download_it Mrs/lady Obama "come on, we move" and that launched her last February to focusing light continuing on disease fat(ness)/Samnah the children in the United States and also on the millions of guys/al-Shabab the exhibitions to the injury danger with illnesses dangerous related with the fatness.

Statistical MT

2 The Obama campaign Ms. "come on move" last February, which unleashed by continuous casting ضوءاً سمنة satisfactory to children in the United States and also on the millions of youth المعرضين at risk of contracting serious diseases related to fat



MT examples (with sample question)

What portion of Mrs. Obama's speech is highlighted in Lines 1-2?

- a. Her belief that obese children can become seriously ill.
- b. A program to raise funds for terminally ill Americans.
- c. The need to help millions of children without health care.
- d. The NAACP's campaign to improve US medical care.

Arabic

والتي أطلقتها في فبراير الماضي إلى تسليط ضوءاً مستمراً على مرض سمنة "هيا نتحرك" وتهدف حملة السيدة أوباما الأطفال في الولايات المتحدة و أيضاً على ملايين الشباب المعرضين لخطر الإصابة بأمراض خطيرة ذات صلة بالسمنة.

Rule-based MT

2 And aim(s) campaign/download_it Mrs/lady Obama "come on, we move" and that launched her last February to focusing light continuing on disease fat(ness)/Samnah the children in the United States and also on the millions of guys/al-Shabab the exhibitions to the injury danger with illnesses dangerous related with the fatness.

Statistical MT

2 The Obama campaign Ms. "come on move" last February, which unleashed by continuous casting ضوءاً سمنة satisfactory to children in the United States and also on the millions of youth المعرضين at risk of contracting serious diseases related to fat



MT examples (with sample question)

What portion of Mrs. Obama's speech is highlighted in Lines 1-2?

- Her belief that obese children can become seriously ill.
- A program to raise funds for terminally ill Americans.
- The need to help millions of children without health care.
- The NAACP's campaign to improve US medical care.

Arabic

والتي أطلقتها في فبراير الماضي إلى تسليط ضوءاً مستمراً على مرض سمنة "هيا نتحرك" وتهدف حملة السيدة أوباما الأطفال في الولايات المتحدة و أيضاً على ملايين الشباب المعرضين لخطر الإصابة بأمراض خطيرة ذات صلة بالسمنة.

Hybrid MT

2 The Obama campaign Ms. "come on move" by last February which continued to shed light on disease fat children in the United States and also on the millions of youth the exhibitions at risk of contracting serious diseases related to fat.



MT examples (with sample question)

What portion of Mrs. Obama's speech is highlighted in Lines 1-2?

- Her belief that obese children can become seriously ill.
- A program to raise funds for terminally ill Americans.
- The need to help millions of children without health care.
- The NAACP's campaign to improve US medical care.

Arabic

والتي أطلقتها في فبراير الماضي إلى تسليط ضوءاً مستمراً على مرض سمنة "هيا نتحرك" وتهدف حملة السيدة أوباما الأطفال في الولايات المتحدة و أيضاً على ملايين الشباب المعرضين لخطر الإصابة بأمراض خطيرة ذات صلة بالسمنة.

Hybrid MT

2 The Obama campaign Ms. "come on move" by last February which continued to shed light on disease fat children in the United States and also on the millions of youth the exhibitions at risk of contracting serious diseases related to fat.

Human Translation

2 Mrs. Obama's "Let's Move" campaign, which she launched this past February, aims to shine a constant spotlight on childhood obesity in the United States, and the millions of young people at risk of developing related serious health conditions.



Participant overview

- Recruited via Amazon's Mechanical Turk
 - Required to be in the United States
- 77-78 participants per language
 - Each language tested separately
 - No information about what the source language was
 - Participants could do experiment in multiple languages
- All participants saw all translation types
 - Told that some were translated by humans and some by machines, but not which was which



Participant demographics

	Arabic	Farsi	Portuguese	Russian	Swahili
n =	77	78	78	78	78
M/F	40/37	38/40	34/44	43/34	46/32
Average age	36.8	38.9	38.9	40.6	36.8
Age range	20-69	22-67	20-68	20-69	19-69
At least some college	81%	91%	81%	79%	81%
Native language	English	English	English	English	English
Some knowledge of source language	n = 1 <i>proficiency rating = 3 (1-10 scale)</i>		n = 1 <i>proficiency rating = 3 (1-10 scale)</i>		

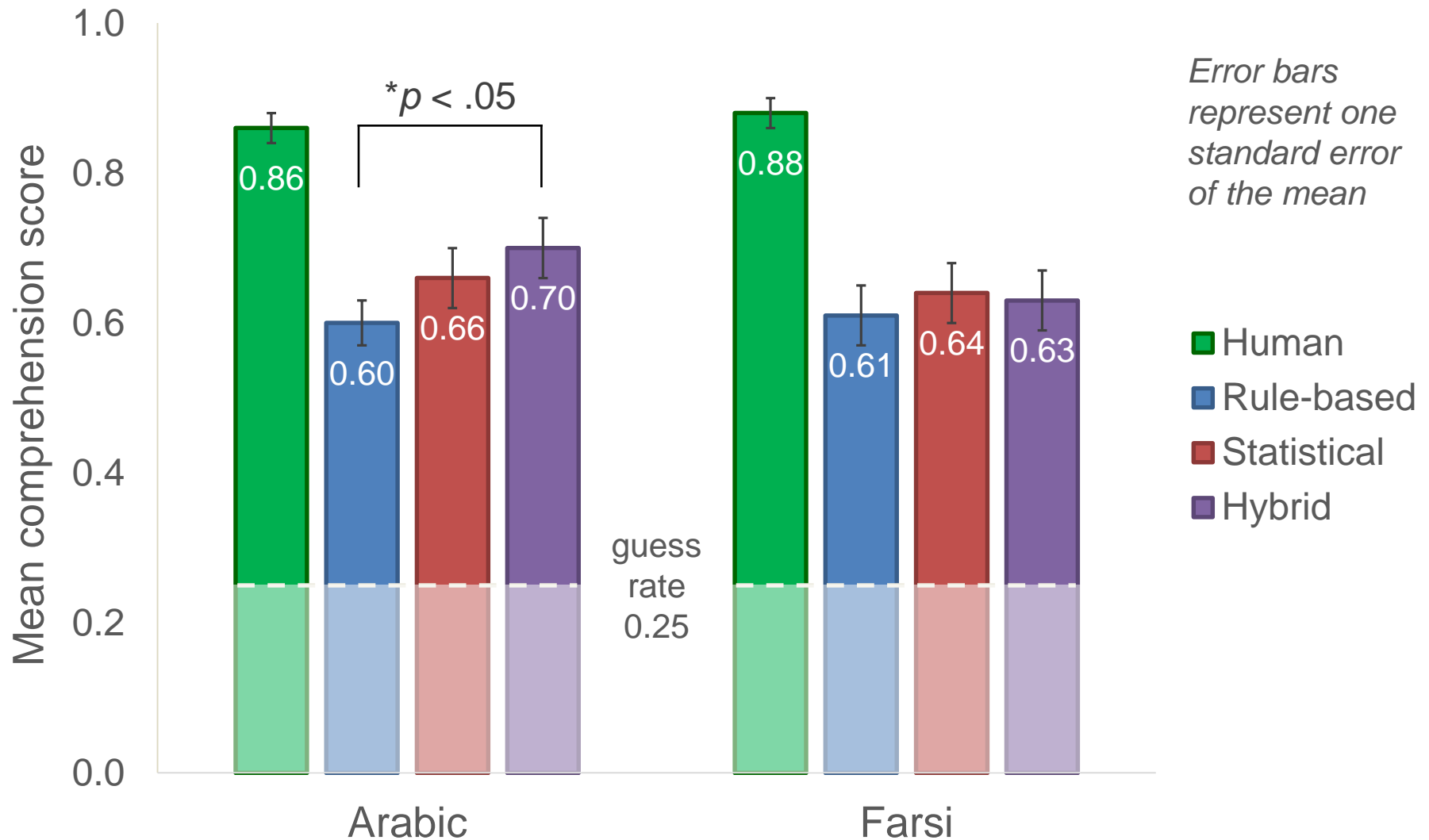


Participant comments

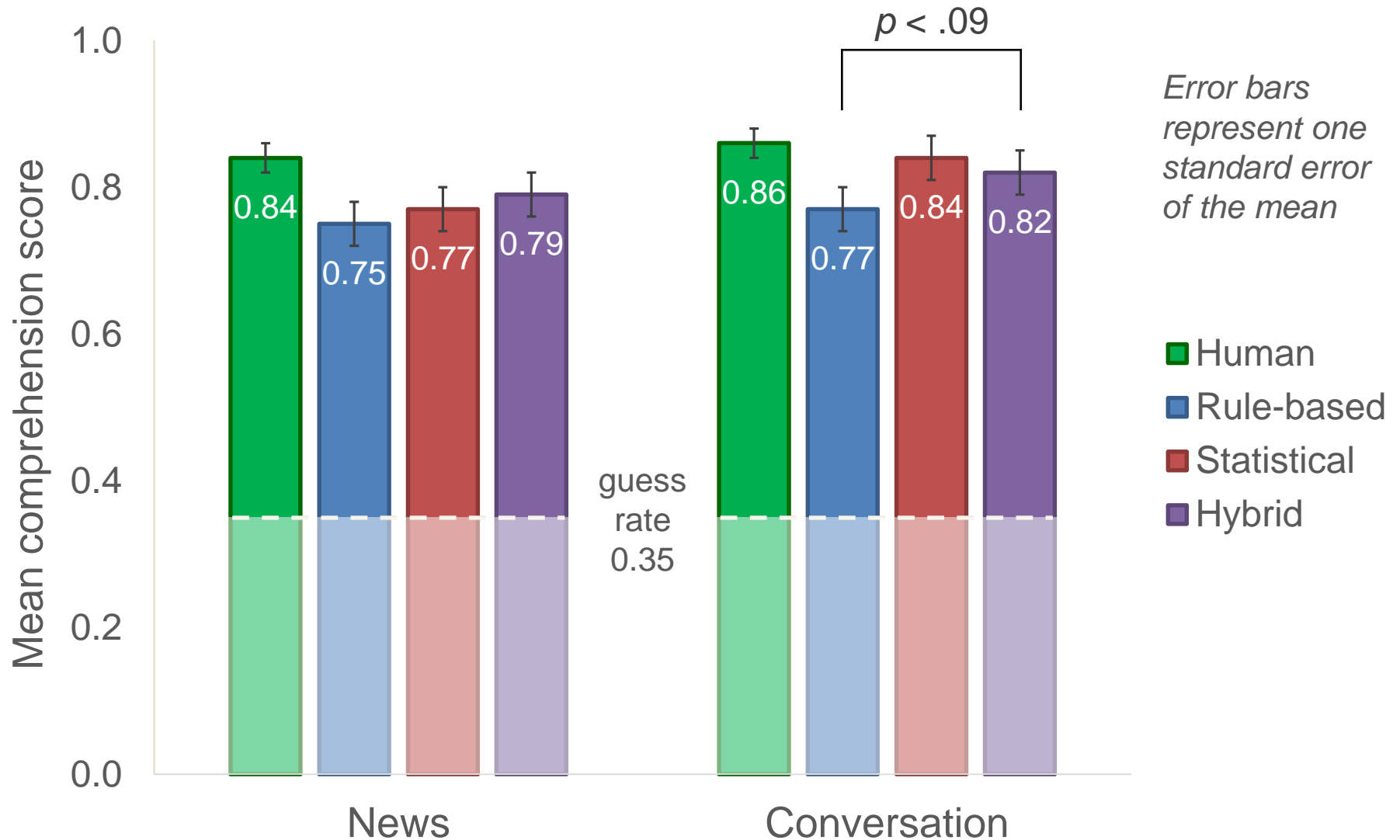
- I found the computer generated text very difficult to **get specific information** from
- This wasn't as easy as I thought it would be. Most of it was **gibberish** but I tried my best to understand the meaning.
- The reading was **really hard to follow** and often I had to read it 4-5 times just to make a guess.
- This **melted my brain.** / My **brain is fried.** / My **brain hurts!**
- Some questions asked for **reference to a specific person or phrase** and a couple of those **weren't actually found in the texts.**
- This task was really interesting, like I was **trying to decode something.** I often had to do this with my family, as my entire family came from Puerto Rico and spoke with very **broken English.**



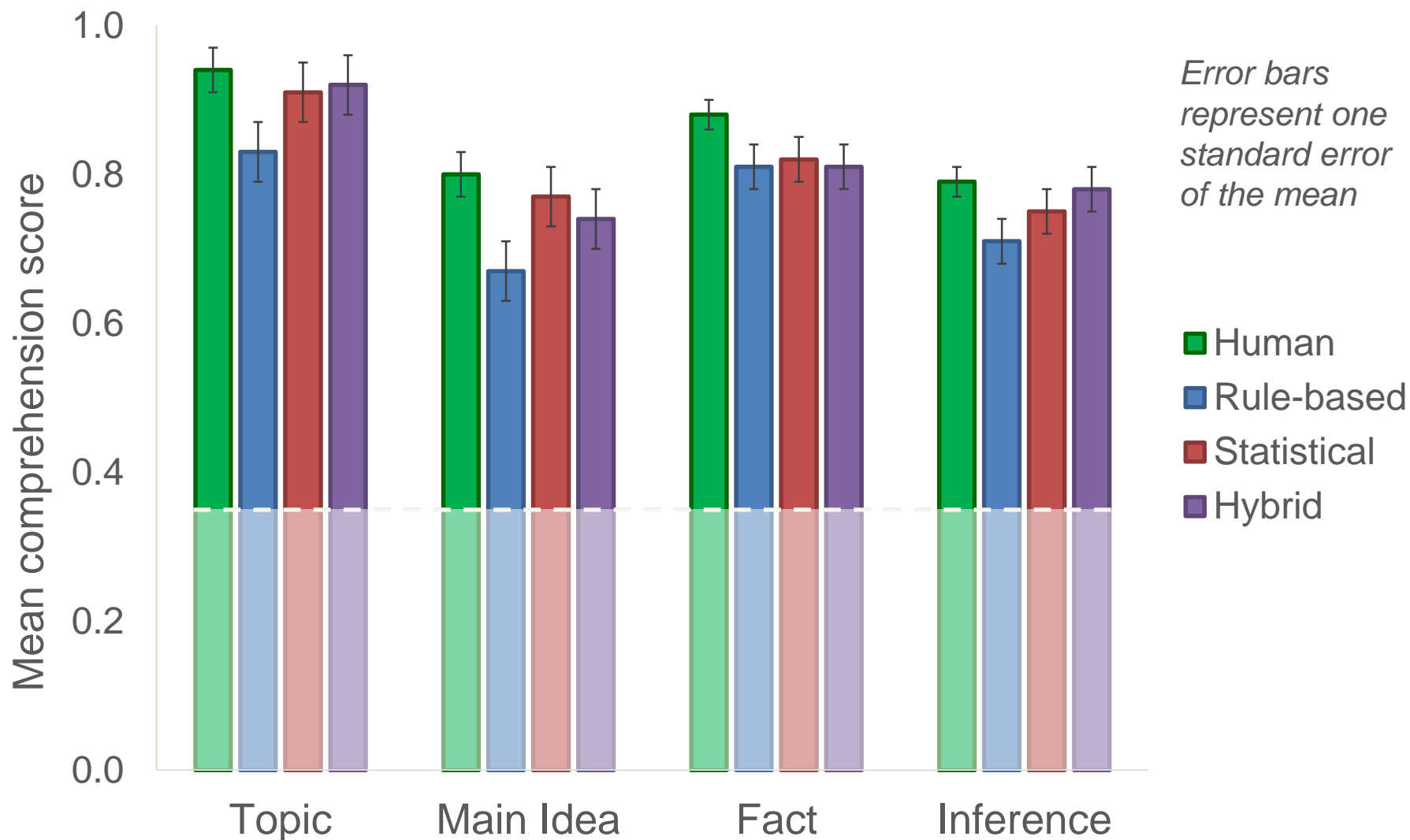
Arabic & Farsi comprehension



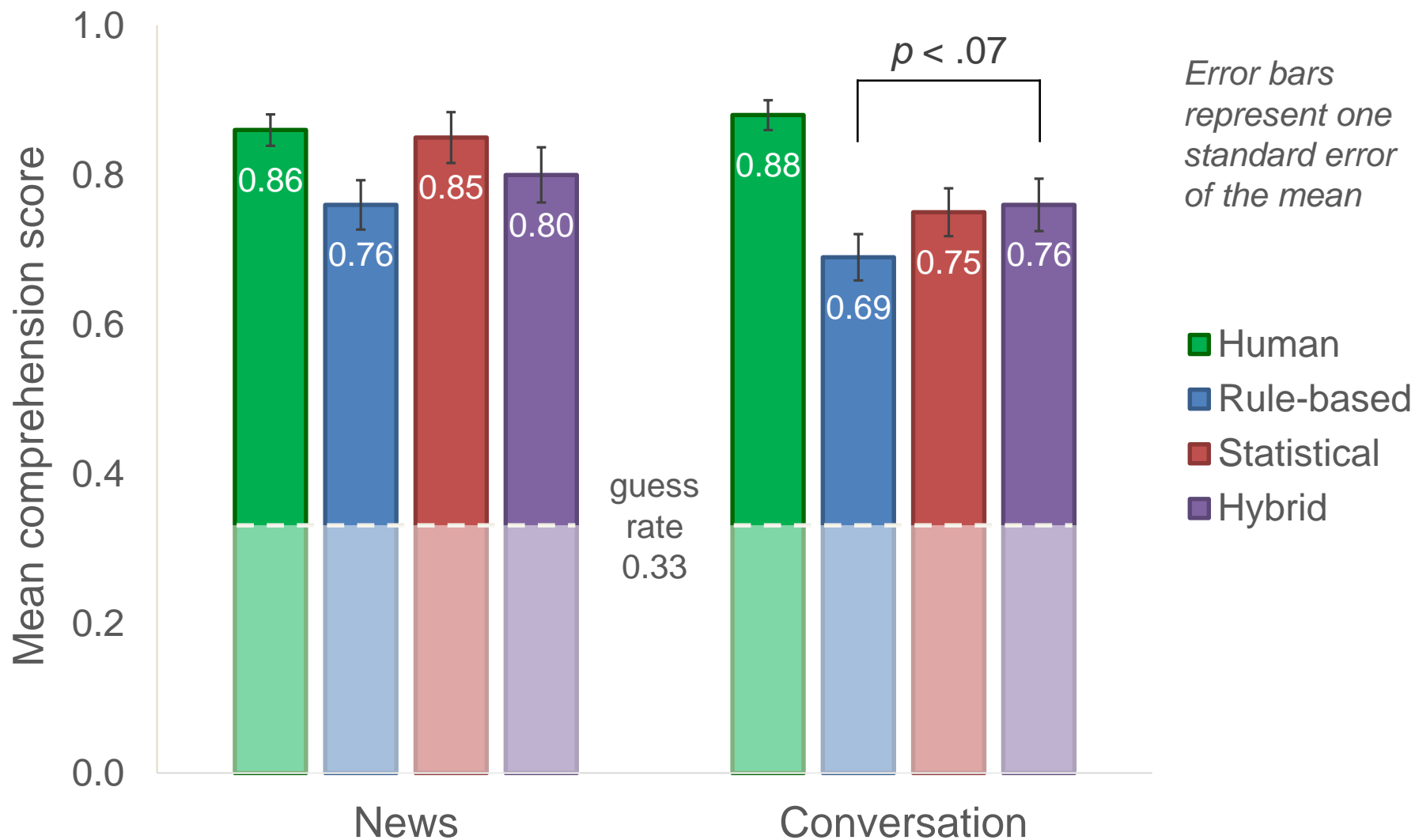
Portuguese comprehension



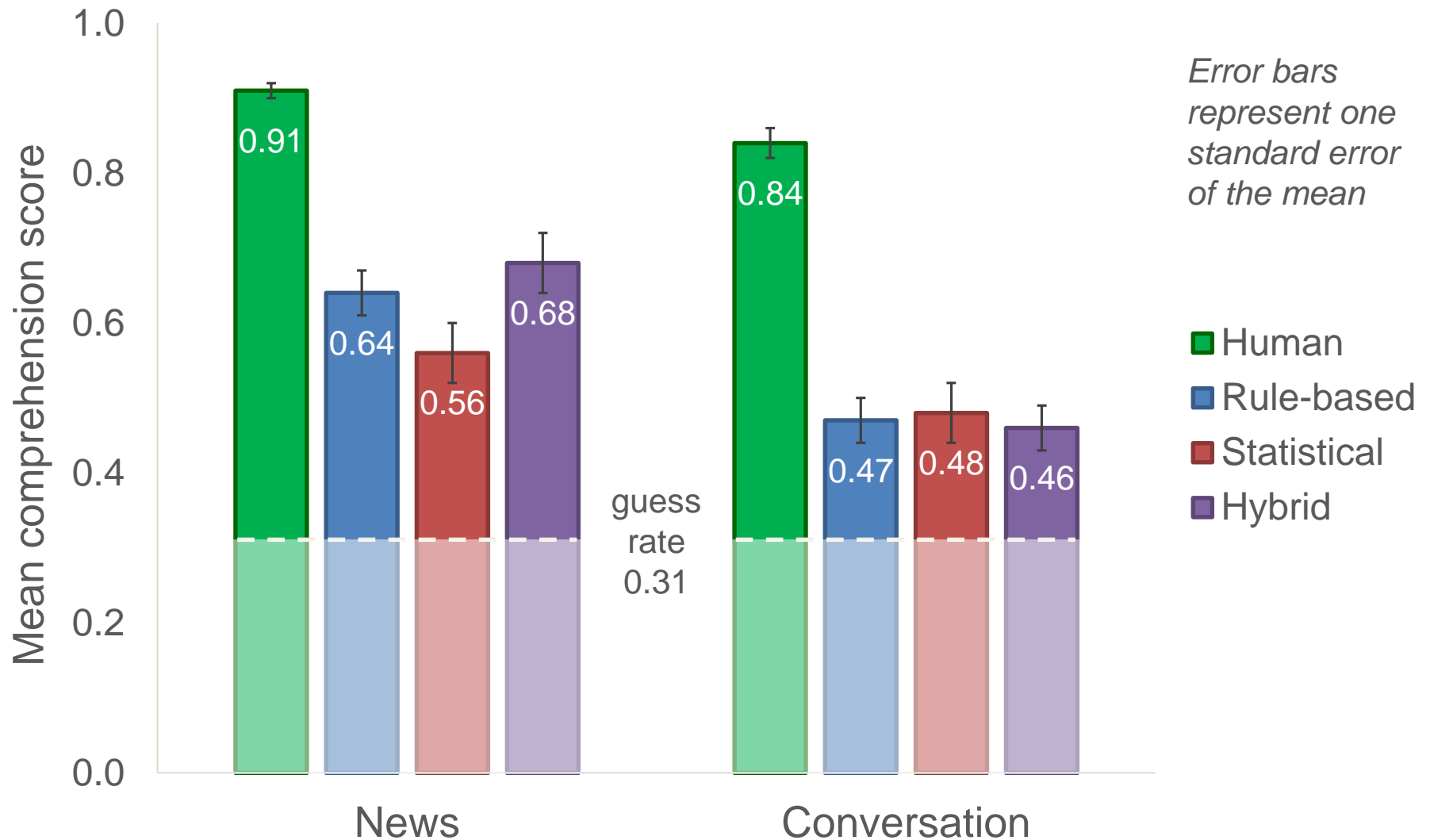
Portuguese comprehension



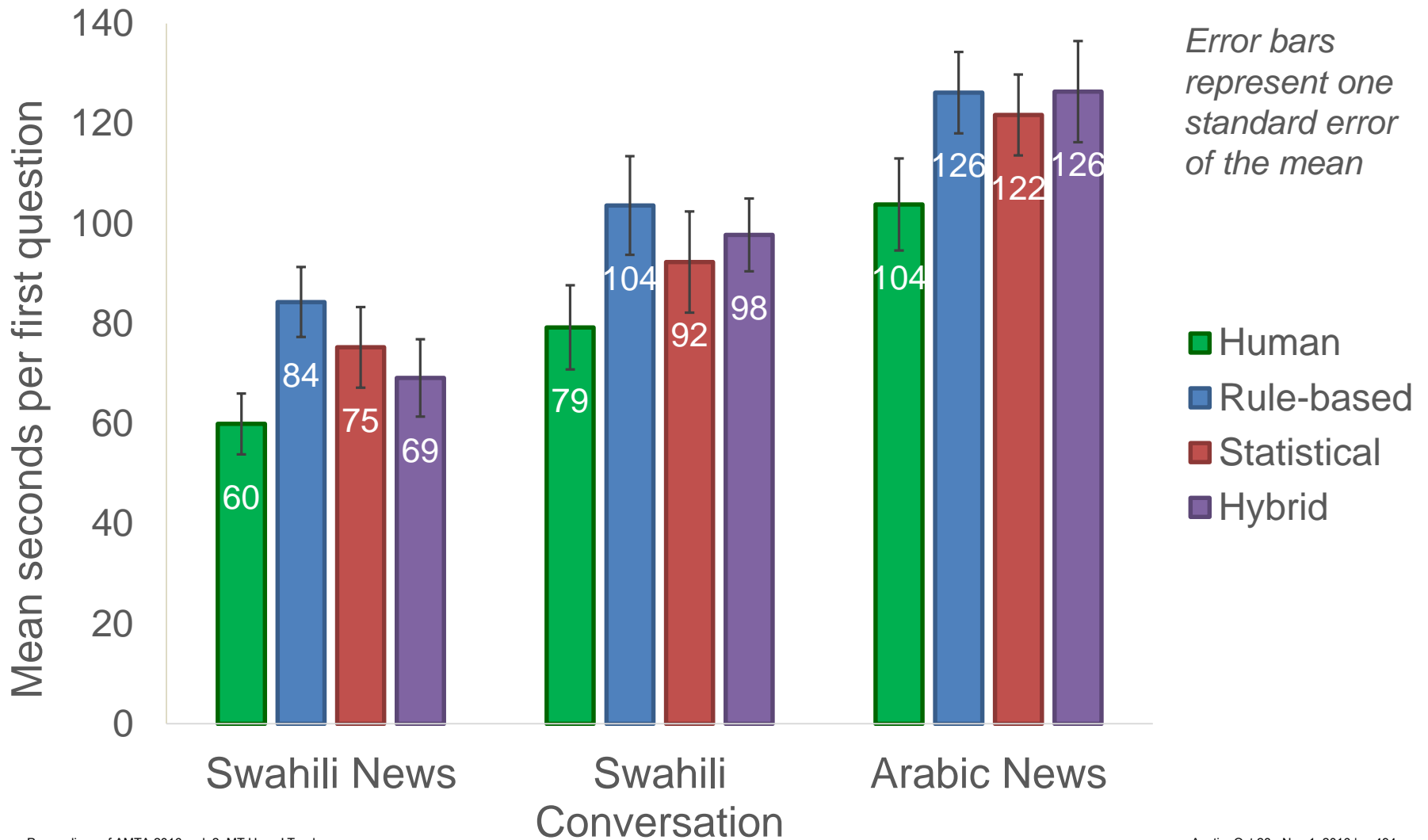
Russian comprehension



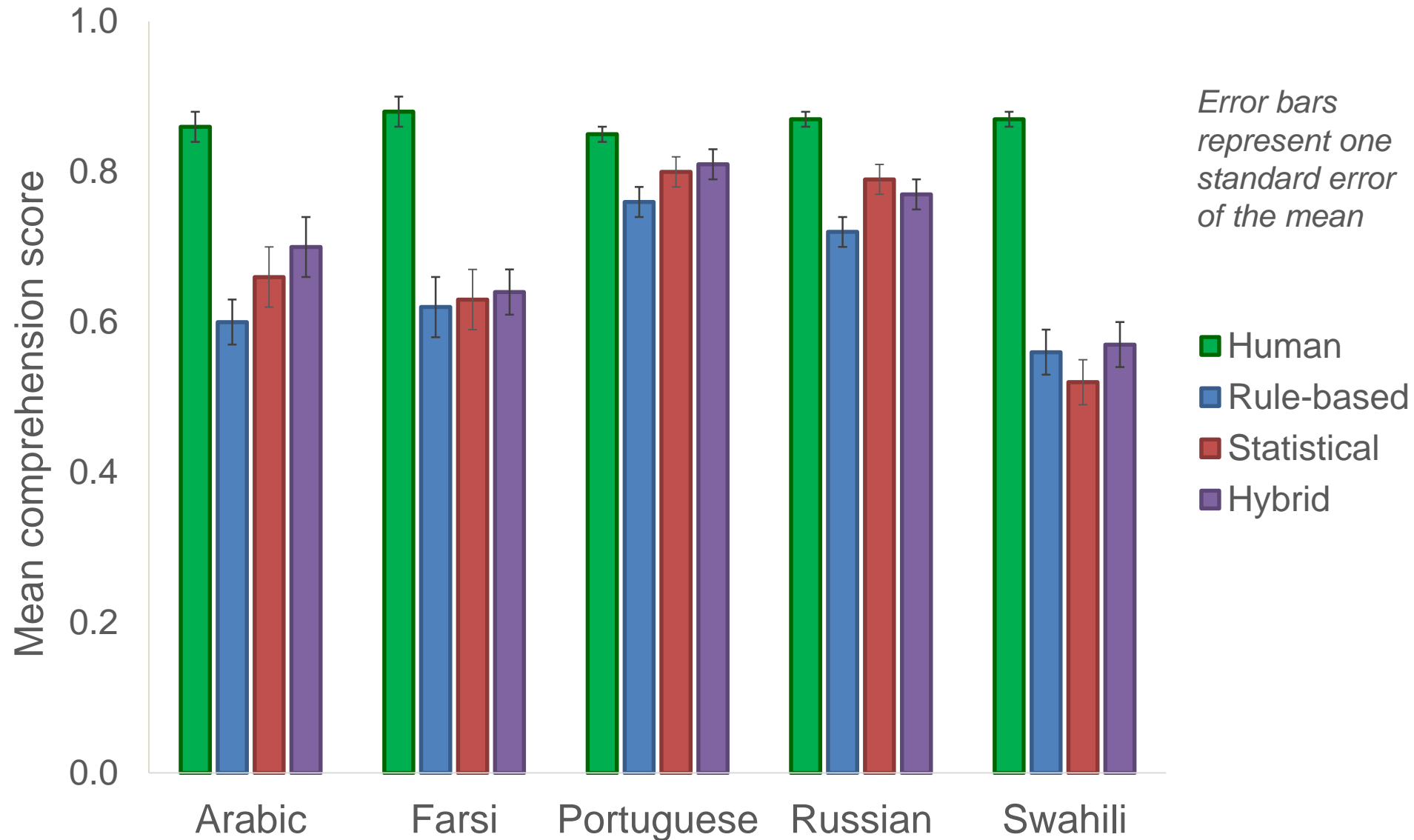
Swahili comprehension



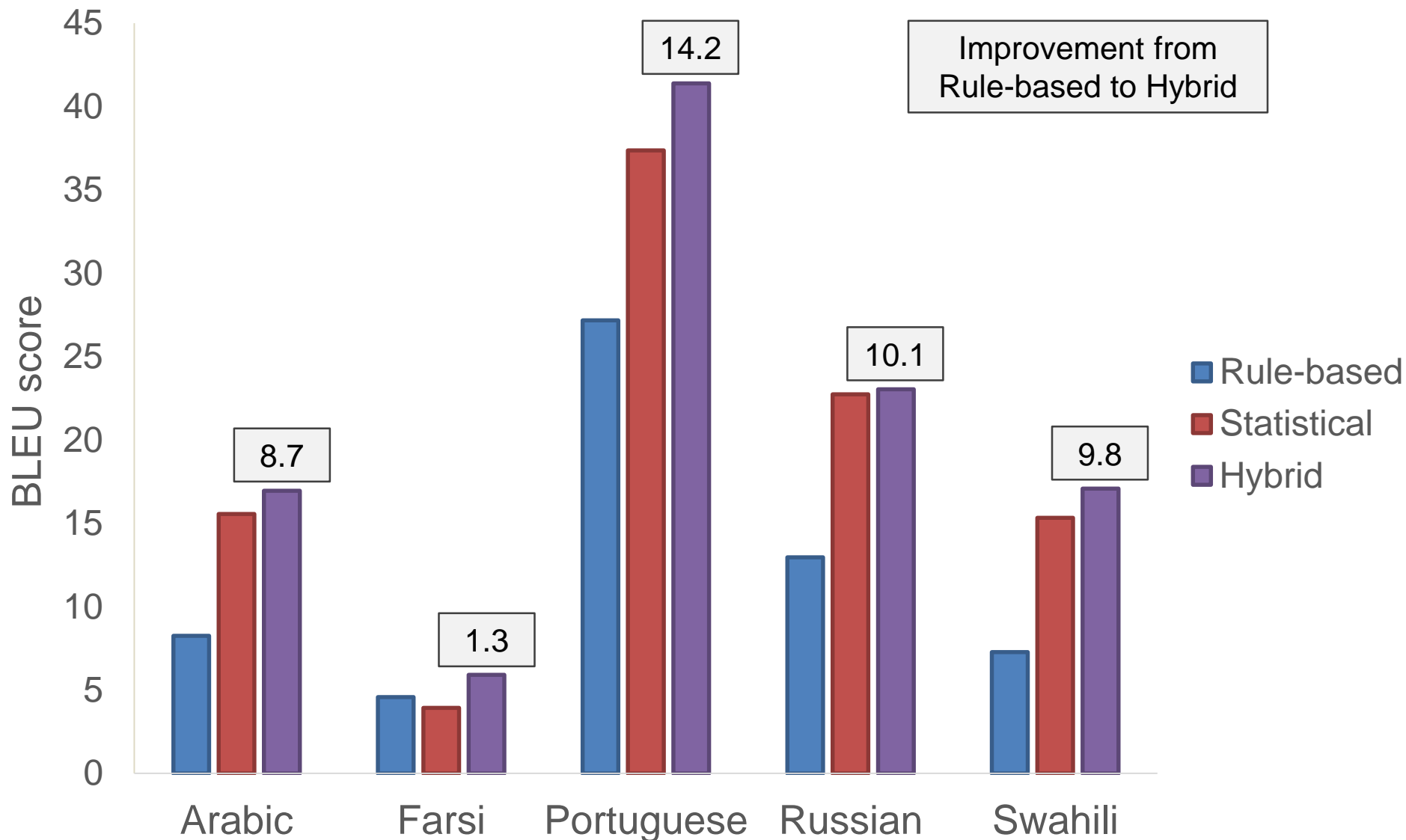
Time per first question (Swahili, Arabic)



Comprehension summary



BLEU scores



Best comprehension scores

	News	Conversation
Arabic	Hybrid	<i>Not tested</i>
Farsi	No sig differences	<i>Not tested</i>
Portuguese	No sig differences	Stat/Hybrid (marginal)
Russian	StatMT	Hybrid (marginal)
Swahili	Hybrid*	No sig differences

*Significantly higher than StatMT, but not significantly different from Rule-based



Automatic scoring metrics for MT

- TER – Translation Error Rate
- BLEU – BiLingual Evaluation Understudy
- METEOR – Metric for Evaluation of Translation with Explicit ORdering



TERPa

- Translation Error Rate Plus
 - Like other TER scores, measures the edits required to change MT output to match a reference translation
 - Edits: shifts, substitutions, insertions, deletions
 - Improves on TER to be more correlated with human judgments of translation quality compared to BLEU (Snover, Dorr, Schwartz, Micciulla, & Makhoul, 2006)



TERPa scoring

- Scoring
 - 0 = perfect match (can have synonyms)
 - 1.0 = default cap
 - 2+ = theoretical maximum
- Can use $1/\text{TERPa}$ for easier interpretation



BLEU

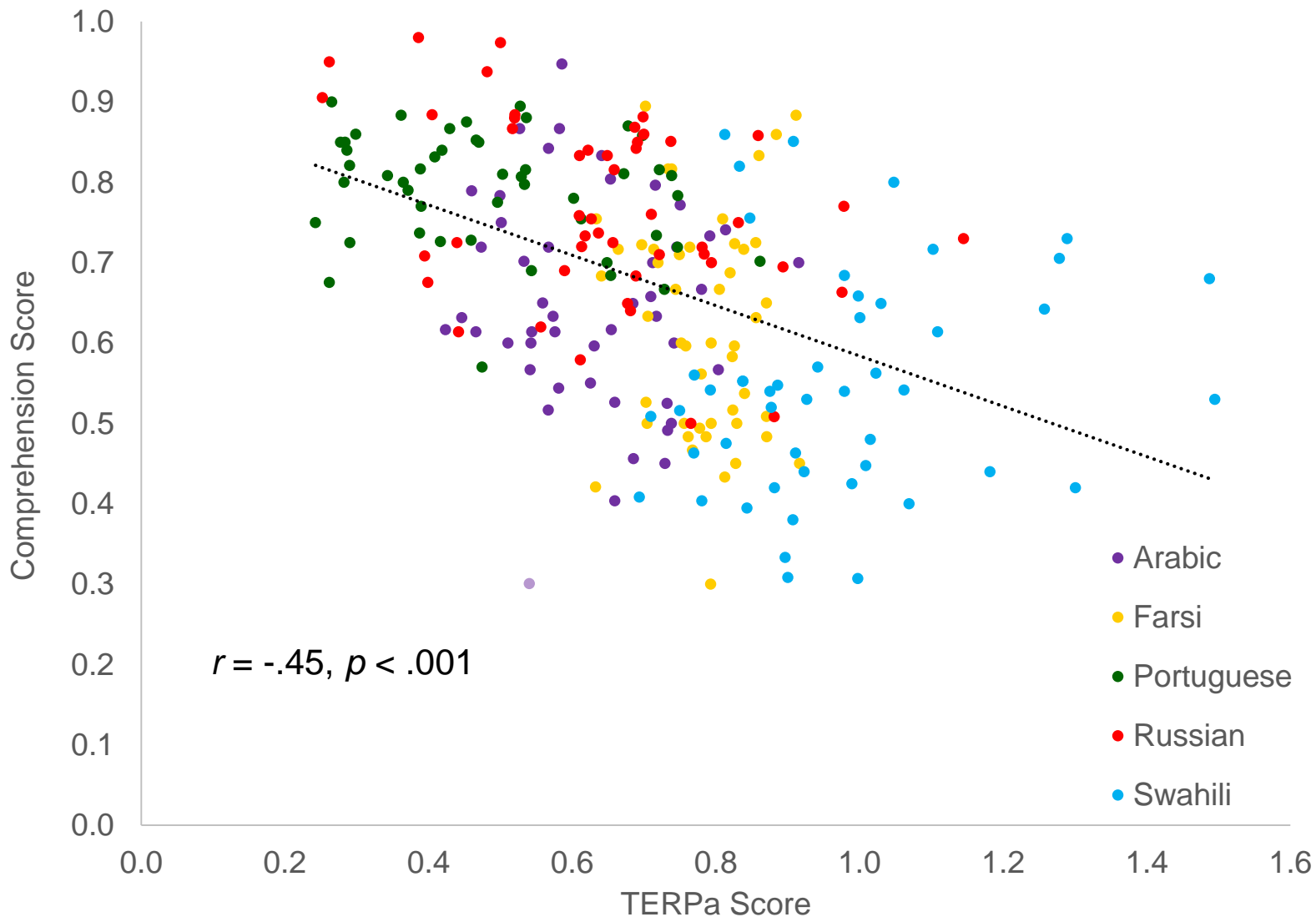
- Measures the number of n-grams from the MT that occur within the reference(s)
- Should be used with a large number of references and large number of sentences in order to correlate with human judgments

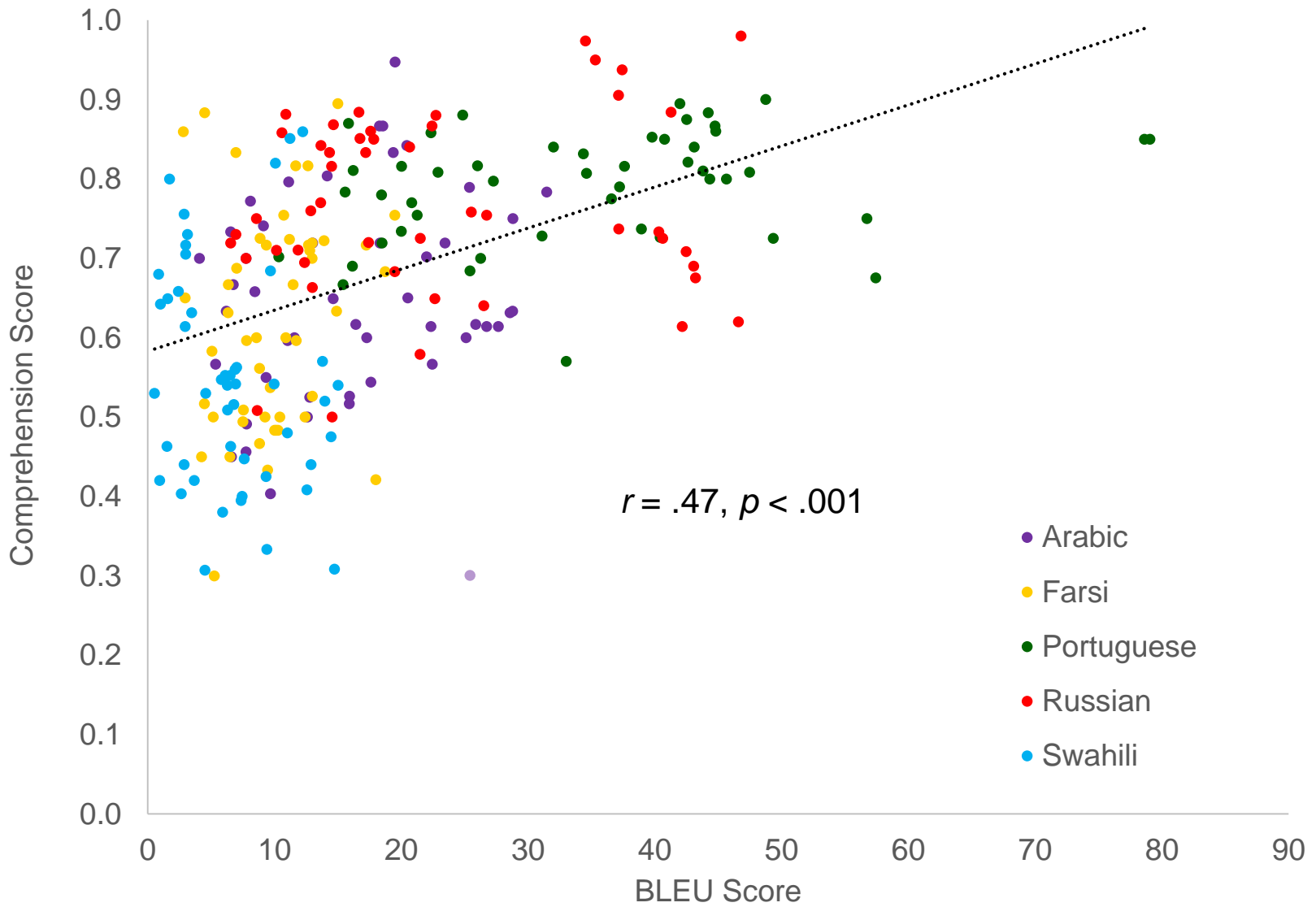


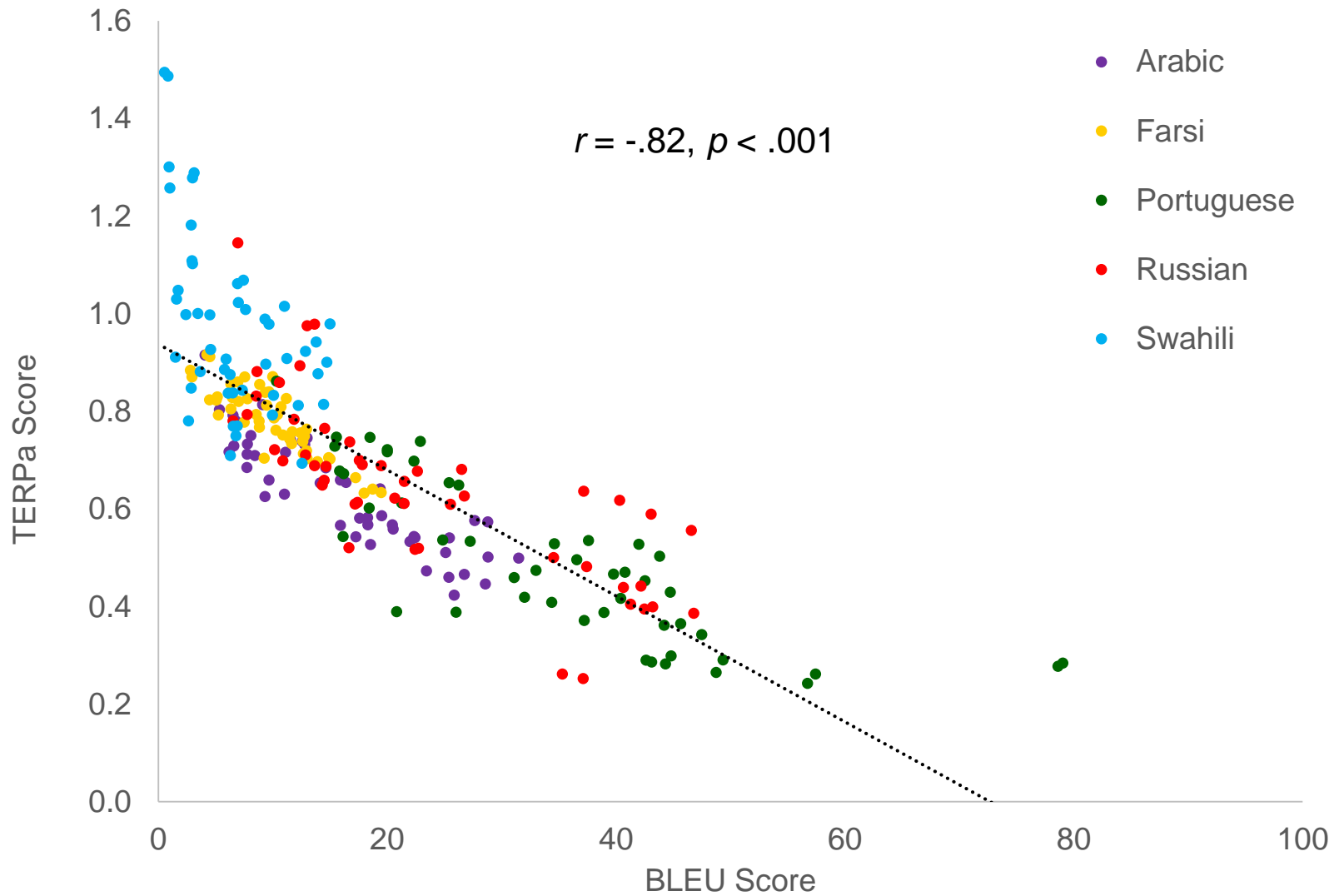
BLEU scoring

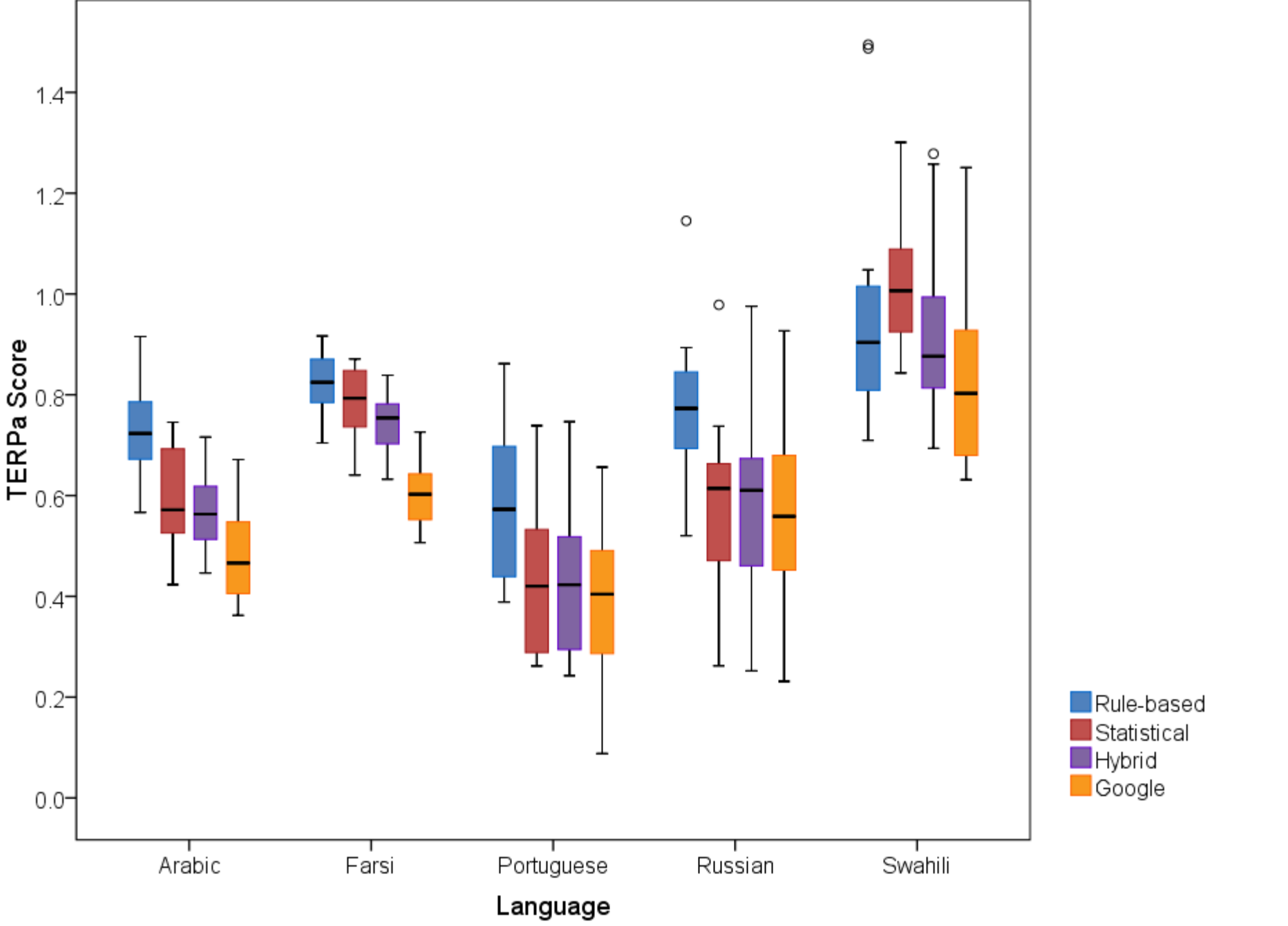
- 0 = poorest match
- 100 = perfect match
- Changes of 1-2 points considered “publishable”



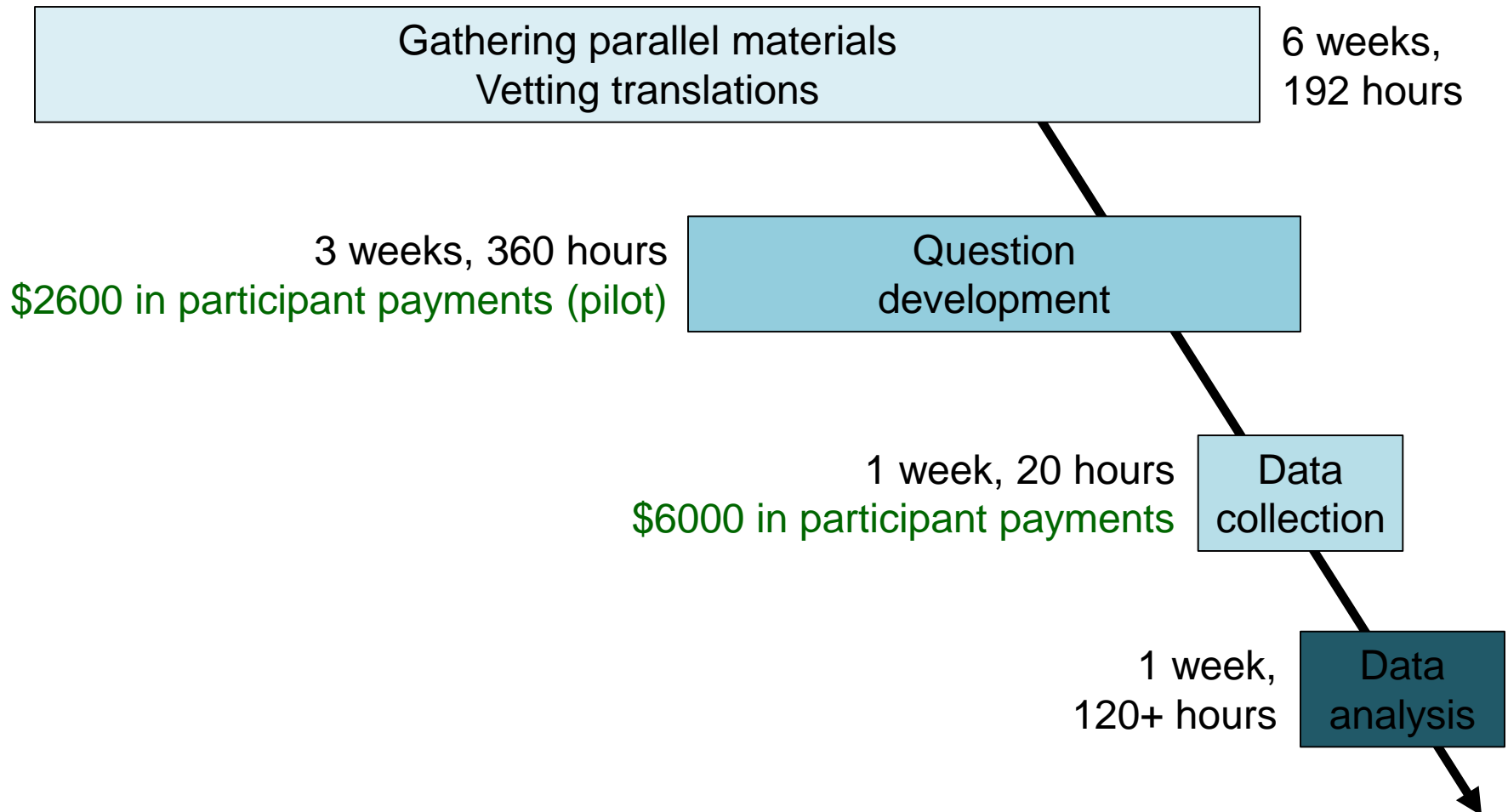








Level of effort required



Summary

- Comprehension of MT output was well above chance, meaning users were able to get some information from the MT even though they found it difficult to understand.
- Improvements in BLEU scores were associated with improvements in comprehension.
- In most conditions, comprehension scores for Hybrid MT output were better than for Rule-based MT.



Future directions

- Collect data with documents produced in the source language
 - Requires time for translation(s) and verification
- Design comprehension questions that focus on essential elements of information
 - EEIs = information critical to decision making
 - Vary based on organization/mission (e.g., FEMA EEIs might differ from law enforcement EEIs)



Future directions (cont'd)

- Compare comprehension scores to paired comparison scores
 - Human judgment of MT output usually done at the sentence level, often with paired comparison of overall quality (e.g., WMT16)
 - Process is less labor-intensive than developing comprehension questions





CENTER FOR ADVANCED
STUDY OF LANGUAGE

JOHNS HOPKINS
UNIVERSITY



human language technology
center of excellence

Questions on Part 1?



CENTER FOR ADVANCED
STUDY OF LANGUAGE

Use of MT for translation and comprehension of Chinese texts



Design

- Two task conditions
 - Full translation
 - Comprehension questions (multiple choice, short answer)
- Three MT conditions
 - From Scratch: No MT (access to online dictionaries only)
 - Post-Editing: Static MT output (from Google Translate)
 - Interactive: Opportunity to interact with MT system
- Post-task and post-experiment questionnaires



Why interactive condition?

Example of how MT output changes depending on how the input is segmented

Translation 1

在集成电路芯片设计过程中考虑不周全或恶意植入不受使用方控制的程序或电路

Not considered comprehensive or implant malicious programs or the circuit from the control of the use of the integrated circuit chip design process

Translation 2

在集成电路芯片设计过程中考虑

In the integrated circuit chip

设计过程中

The design process

考虑

Consider

不周全或恶意植入

Not comprehensive or malicious implants

不受使用方控制的程序或电路

Without the use of the control of the program or circuit



Topics

- Six scientific abstracts on highly technical topics from mainland China journals (125-175 characters)
 - **Text A:** Embedding information in satellite imagery
 - **Text B:** Detecting malware intrusions in Windows
 - **Text C:** Advances in helicopter collision avoidance radar
 - **Text D:** Antimissile command and control systems
 - **Text E:** Anti-interference capabilities of radar
 - **Text F:** Optimization models for firing effectiveness



Text characteristics

No correlation, i.e., the fluency of MT output does not necessarily give you a clue to other aspects of its quality

Text	Inverse TERp scores ^a	Fluency Ratings ^b	Difficulty Ratings ^c	Familiarity Ratings ^d
A	1.61	2.3 (0.50)	4.1 (1.01)	3.6 (0.67)
B	1.30	2.3 (0.82)	3.8 (0.95)	3.1 (0.93)
C	1.49	3.2 (0.75)	3.2 (0.98)	3.2 (0.78)
D	1.33	2.8 (0.75)	4.2 (0.74)	3.7 (0.66)
E	1.56	2.2 (0.83)	4.5 (0.76)	3.8 (0.52)
F	2.08	2.5 (0.93)	3.6 (1.00)	3.4 (0.76)

^a TERp (Translation Error Rate plus) = relative to gold standard; higher = more similar

^b 1 = not at all fluent; 5 extremely fluent (12 independent raters)

^c 1 = very easy; 5 = very difficult

^d 1 = very familiar; 4 = not at all familiar



Text characteristics

Significant correlation, i.e.,
more difficult texts were also
rated as less familiar

Text	Inverse TERp scores ^a	Fluency Ratings ^b	Difficulty Ratings ^c	Familiarity Ratings ^d
A	1.61	2.3 (0.50)	4.1 (1.01)	3.6 (0.67)
B	1.30	2.3 (0.82)	3.8 (0.95)	3.1 (0.93)
C	1.49	3.2 (0.75)	3.2 (0.98)	3.2 (0.78)
D	1.33	2.8 (0.75)	4.2 (0.74)	3.7 (0.66)
E	1.56	2.2 (0.83)	4.5 (0.76)	3.8 (0.52)
F	2.08	2.5 (0.93)	3.6 (1.00)	3.4 (0.76)

^a TERp (Translation Error Rate plus) = relative to gold standard; higher = more similar

^b 1 = not at all fluent; 5 extremely fluent (12 independent raters)

^c 1 = very easy; 5 = very difficult

^d 1 = very familiar; 4 = not at all familiar



Text characteristics

Proficiency self-ratings (not shown) negatively correlated with difficulty ratings, but not with familiarity ratings

Text	Inverse TERp scores ^a	Fluency Ratings ^b	Difficulty Ratings ^c	Familiarity Ratings ^d
A	1.61	2.3 (0.50)	4.1 (1.01)	3.6 (0.67)
B	1.30	2.3 (0.82)	3.8 (0.95)	3.1 (0.93)
C	1.49	3.2 (0.75)	3.2 (0.98)	3.2 (0.78)
D	1.33	2.8 (0.75)	4.2 (0.74)	3.7 (0.66)
E	1.56	2.2 (0.83)	4.5 (0.76)	3.8 (0.52)
F	2.08	2.5 (0.93)	3.6 (1.00)	3.4 (0.76)

^a TERp (Translation Error Rate plus) = relative to gold standard; higher = more similar

^b 1 = not at all fluent; 5 extremely fluent (12 independent raters)

^c 1 = very easy; 5 = very difficult

^d 1 = very familiar; 4 = not at all familiar



Sample comprehension questions

- Multiple choice: 3 questions standard across all passages
 1. Select the topic **domain** to which this article belongs.
 - a) Information Technology
 - b) Detection Technology
 - c) Nuclear Technology
 - d) Medical Technology
 - e) Command and Control
 - f) Weapons Technology
 2. Select three **keywords** that describe this passage.
 - a) Hidden Information
 - b) Detection Systems
 - c) Cyber Security
 - d) Safety Testing
 - e) Communication Systems
 - f) Perception & Cognition
 - g) Manufacturing
 - h) Robotics
 - i) Imagery/Imaging
 3. Select the best descriptive **title** for this passage.
 - a) Detecting Malware Intrusions
 - b) The Use of Behavioral Characteristics
 - c) RootKit for Windows
 - d) Reliable Windows Technology



Sample comprehension questions

- Multiple choice: 2 questions unique to each passage
 - On which of the following is the detection technique based?
 - a) Hook System Call
 - b) Hard System Call
 - c) Root System Call
 - d) Lock System Call
- True or false
 - According to the author, this detection method is completely reliable. (T/F)
- Short answer (expected response = a few words)
 - What kind of attack is the technique designed to detect?
- Open-ended (expected response = a sentence)
 - How does the detection technique work?



Demographics

- $n = 51$ (23 males, 28 females)
- Average age = 23.5 years (range = 18-58)
- 45 reported at least some college education
- Recruitment targeted advanced undergraduate learners of Chinese
 - All native English speakers
 - 4 also native Chinese (heritage speakers)



Chinese experience

- Average months of study = 43.2 (range = 6-168)
- Most had taken at least one course at the 300-level or higher
- Immersion experience: $n = 43$
 - Most in teens or 20's
 - Most for 1-5 months

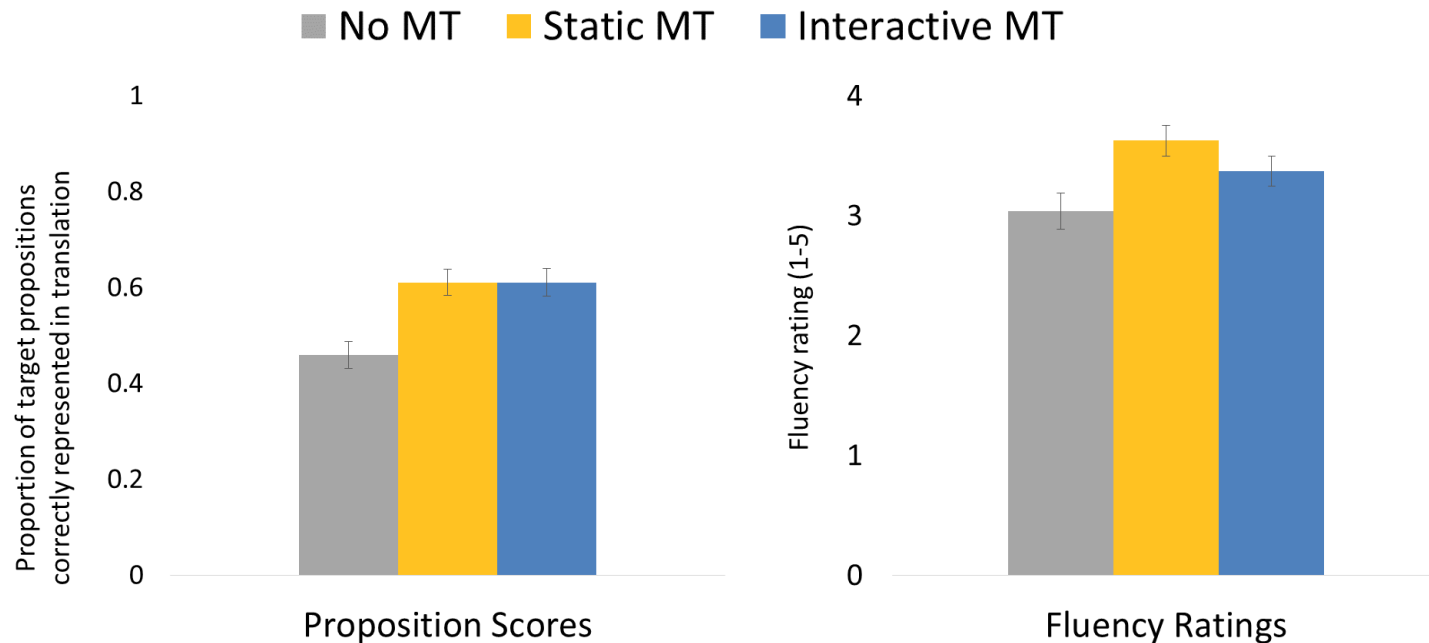


Translation outcome measures

- Proposition score (accuracy)
 - Scored by research team; counted presence or absence of pre-determined set of “propositions”
- Fluency ratings
 - Judged by 3 members of the research team
- Inverse TERp scores (higher = more similar)
 - Distance from gold standard
 - Distance from MT output
- Time on task



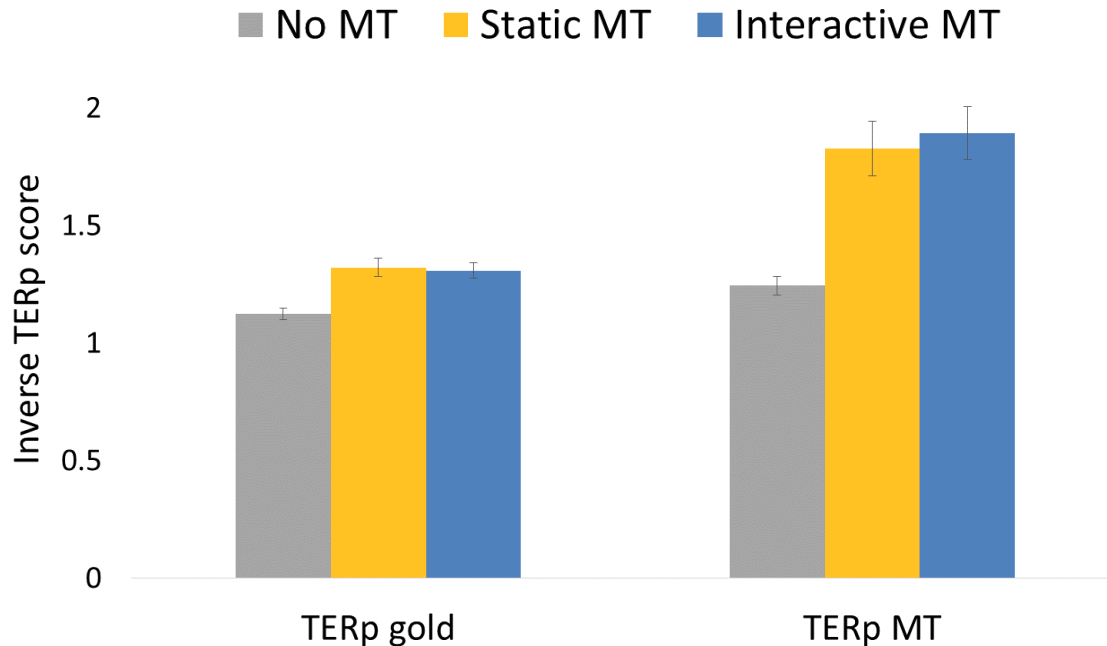
Translation quality



Proposition scores and fluency ratings were significantly lower in the No MT condition than in the two MT conditions, which did not differ significantly from each other.



TERp scores

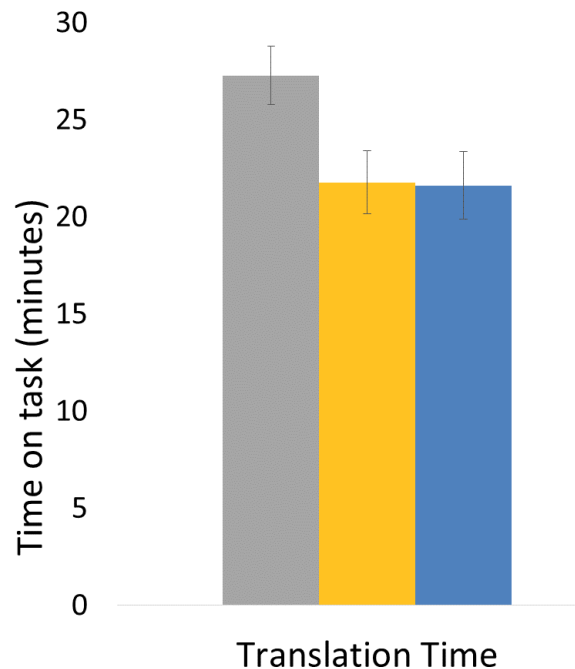


Translations produced in the No MT condition were significantly more distant from both the gold standard and the MT output than were translations produced in the two MT conditions, which did not differ significantly from each other.



Translation time on task

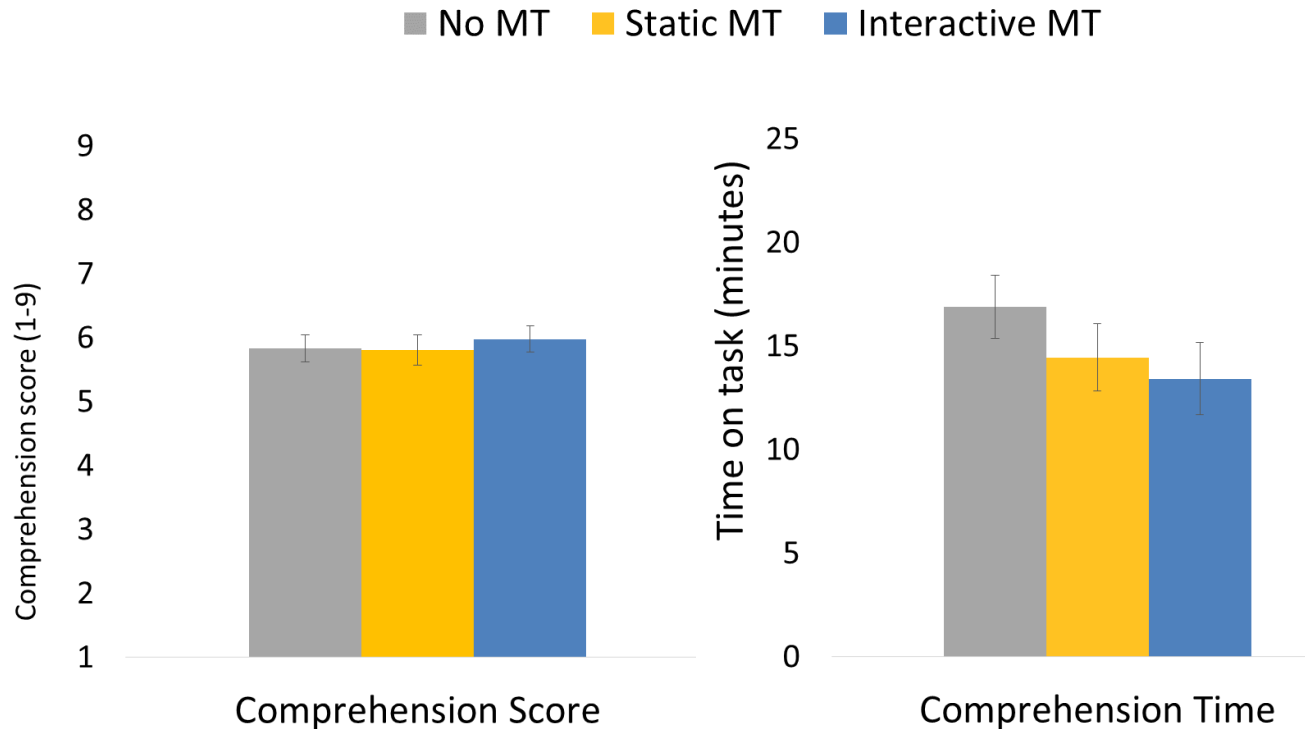
■ No MT ■ Static MT ■ Interactive MT



Participants took significantly longer to generate their translations in the No MT condition than in the two MT conditions, which did not differ significantly from each other.



Comprehension outcomes



- **Comprehension scores did not differ significantly across conditions.**
- Participants took significantly longer to answer the comprehension questions in the No MT condition than in the two MT conditions, which did not differ significantly from each other.



Comprehension vs. Translation

- How did you approach the task differently than you might have if you were writing a translation?
 - *Tried to get a "gist" of what was being said.*
 - *Focused on keywords / concepts rather than a cohesive translation.*
 - *Spent less time worrying about grammar.*
 - *I just trusted the provided MT output.*
 - *A lot of questions were related to very specific aspects of the text.*
 - *I had to really understand the meanings of technical terms, as they were needed to respond to the questions.*



Reasons for not interacting with MT

- Translation condition

- I decided to stick with using the online dictionary. I did check with the MT output to correlate with what I looked up.
- I have had better experiences using the online dictionary.
- I felt the MT output, dictionary, and my own knowledge and experience were sufficient.

- Comprehension condition

- I used the dictionary more.
- MT output sufficient.
- Because I felt that it would be the same.



Summary and conclusions

- Differential benefits of MT for translation and comprehension tasks
 - Notable that MT was sometimes helpful even when the quality of the output was relatively poor
 - For comprehension, MT may have helped participants get the gist more quickly, but helped less with details
- Interacting with MT did not help
 - Lack of use; perception that MT would not change
 - Translators may need training on how to use MT most effectively



Future directions

- Other languages and/or tasks
- Effects of proficiency
 - Lower: need more help
 - Higher: better able to discern when MT is misleading
 - Any benefit for native speakers?
- Other domains and genres
 - e.g., colloquial material without complete, grammatical sentences
- Combinations of tools





CENTER FOR ADVANCED
STUDY OF LANGUAGE

Questions on Part 2?



CENTER FOR ADVANCED
STUDY OF LANGUAGE

Utility of translation memory (TM) in an operational context

Motivation

- An USG organization with access to some CAT tools (MT, concordance search, on-line dictionaries)
- TM resources limited
 - Termbase sharing is manual
 - Labor-intensive to import termbases
- CAT resources have a history of performing differently in an USG context



Materials

- Provided by client
- Mixed genre
 - Transcribed/translated conversations
 - Translated written correspondence
 - Translated documents
 - Targeted summaries (discarded)
- Alignment
 - Often at the paragraph level



Approach

- Worked with 3 experienced analysts to align text
 - Source sentence as maximum length
 - Often could get to the level of an English phrase
- Formatted aligned translation units (TUs) as TMX



Choosing a comparison group

- “Ideal” TM scenario
 - Phone manuals
 - Within-text matches
 - Between text matches
- Scientific abstracts (very narrow domain)



Choosing a metric

- Is TM providing as many matches in the USG context?
 - N-gram matches
 - TU matches



Results

- The USG material had substantially fewer n-gram matches than any of the comparison groups
- No comparison for TU matches
- 30 matches within 1000 TUs



Unique applications

- Event-driven terminology
 - Пятое кольцо
 - “Fifth ring”



Unique applications

- Event-driven terminology

- Пятое кольцо
- “Fifth ring”



- 2014 Sochi Olympics
- Meaning: inept performance/inept performer or failure to complete a job



Unique applications

- Connecting the dots
 - Evolving terminology used to “talk around” topics or refer to individuals
 - Individuals using the same terminology may be connected in some way
 - Alerting analysts to areas of mission overlap and areas for potential collaboration



Unique applications

- Entity/event disambiguation
 - Less than exact matches may prompt analysts to investigate whether individuals, organizations, or events with similar names are in fact the same individual / organization / event
 - Access to prior reports allows quick access to data that may assist with disambiguation



Unique applications

- Entity/event disambiguation
 - International Council of Scientific Unions (1931-1998)
 - Changed to International Council for Science (1998-present)



The screenshot shows the Nature journal website. At the top, the word "nature" is written in a white serif font on a dark red background, with the tagline "International weekly journal of science" in a smaller white sans-serif font to its right. Below this, a breadcrumb trail reads "Journal home > Archive > News > Full Text". On the left side, there is a vertical menu with the heading "Journal content" and three items: "Journal home", "Advance online publication", and "Current issue", each preceded by a small red plus sign. The main content area on the right has the heading "News" and a sub-heading "Nature 403, 582 (10 February 2000) | doi:10.1038/35001201". The main article title is "International science council names first female president".



Translation Tools Platform performs functions such as:

- Segmentation
- Search for TM matches
- Render MT

Algorithm determines which suggestions should be presented to translator.

Source Text:

该种隐藏进程的检测方法十分可靠,可以检测出常规安全检测工具不能发现的系统恶意程序。

Translation Window:

This detection method is completely reliable,可以检测出常规 security detection 工具不能发现的系统恶意程序。

TM suggestion:

Accept Reject

Check for MT matches

MT suggestion:

Accept Reject

Check for TM matches

MT suggestions ON TM suggestions ON Spellcheck OFF



Obstacles to adoption

- At the level of the organization
 - Monetary/development investment
 - Need to invest in data
- At the level of the individual
 - Want seamless access/input
 - Resistant to additional burden
 - Want to be able to opt-in



Desired features

- Supported linking to:
 - Databases
 - Prior reporting
 - Prior authors
- Glossary creation
- Domain tagging
- Support for multiple languages



Summary and conclusions

- TM may provide fewer matches from memory compared to TM used in an “ideal” scenario
- May improve at scale
 - Will never approach the size of termbases of outside information
- Additional applications for TM may make the investment uniquely beneficial





CENTER FOR ADVANCED
STUDY OF LANGUAGE

Questions about Part 3?

Overall conclusions

- Important to consider the task/use case
- Even big changes in BLEU score might not be reflected in comprehension
- Both automatic evaluations and human comparisons are abstractions compared to human comprehension
- Any CAT tools:
 - may face adoption hurdles
 - may offer unanticipated benefits



Proto MT Evaluation for Humanitarian Assistance Disaster Response Scenarios

Doug Jones

Association for Machine Translation in the Americas

31 OCT 2016



This material is based upon work supported by the Defense Advanced Research Projects Agency under Air Force Contract No. FA8721-05-C-0002 and/or FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency or the U.S. Air Force. Approved for public release; distribution unlimited. Delivered to the U.S. Government with Unlimited Rights, as defined in DFARS Part 252.227-7013 or 7014 (Feb 2014). Notwithstanding any copyright notice, U.S. Government rights in this work are defined by DFARS 252.227-7013 or DFARS 252.227-7014 as detailed above. Use of this work other than as specifically authorized by the U.S. Government may violate any copyrights that exist in this work. Delivered to the U.S. Government with Unlimited Rights, as defined in DFARS Part 252.227-7013 or 7014 (Feb 2014). Notwithstanding any copyright notice, U.S. Government rights in this work are defined by DFARS 252.227-7013 or DFARS 252.227-7014 as detailed above. Use of this work other than as specifically authorized by the U.S. Government may violate any copyrights that exist in this work. © 2016 Massachusetts Institute of Technology.



Concept of Experiment

(Last Slide)



SPOILER ALERT

MT may be neither necessary nor sufficient for HADR mission relevance



HADR

Humanitarian Assistance Disaster Relief



Office of U.S. Foreign
Annual Report



Office of U.S. Foreign
Disaster Assistance

Annual Report for Fiscal Year 2010





Top 20 Languages of the World



Legend and Source

- Size of circles indicates number of speakers based on Ethnologue data. Lewis, M. Paul, Gary F. Simons, and Charles D. Fennig (eds.). 2015. Ethnologue: Languages of the World, Eighteenth edition. Dallas, Texas: SIL International.



Coverage of Seven Available MT Systems



Legend and Source

- Size of circles indicates number of speakers based on Ethnologue data. Lewis, M. Paul, Gary F. Simons, and Charles D. Fennig (eds.). 2015. Ethnologue: Languages of the World, Eighteenth edition. Dallas, Texas: SIL International.



~7500 Languages

123 456 789 1011 1213 1415 1617 1819 2021 2223 2425 2627 2829 3031 3233 3435 3637 3839 4041 4243 4445 4647 4849 5051 5253 5455 5657 5859 6061 6263 6465 6667 6869 7071 7273 7475 7677 7879 8081 8283 8485 8687 8889 9091 9293 9495 9697 9899 10000
 10001 10002 10003 10004 10005 10006 10007 10008 10009 10010 10011 10012 10013 10014 10015 10016 10017 10018 10019 10020 10021 10022 10023 10024 10025 10026 10027 10028 10029 10030 10031 10032 10033 10034 10035 10036 10037 10038 10039 10040 10041 10042 10043 10044 10045 10046 10047 10048 10049 10050 10051 10052 10053 10054 10055 10056 10057 10058 10059 10060 10061 10062 10063 10064 10065 10066 10067 10068 10069 10070 10071 10072 10073 10074 10075 10076 10077 10078 10079 10080 10081 10082 10083 10084 10085 10086 10087 10088 10089 10090 10091 10092 10093 10094 10095 10096 10097 10098 10099 10100
 ... [The rest of the page contains a dense grid of numbers and characters, likely representing a list of languages or a data set.] ...



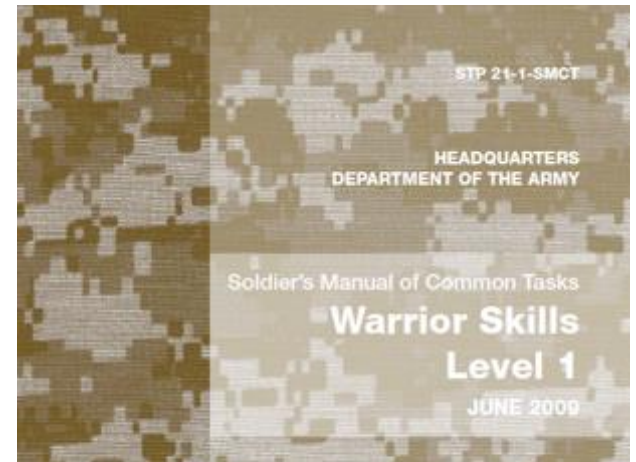
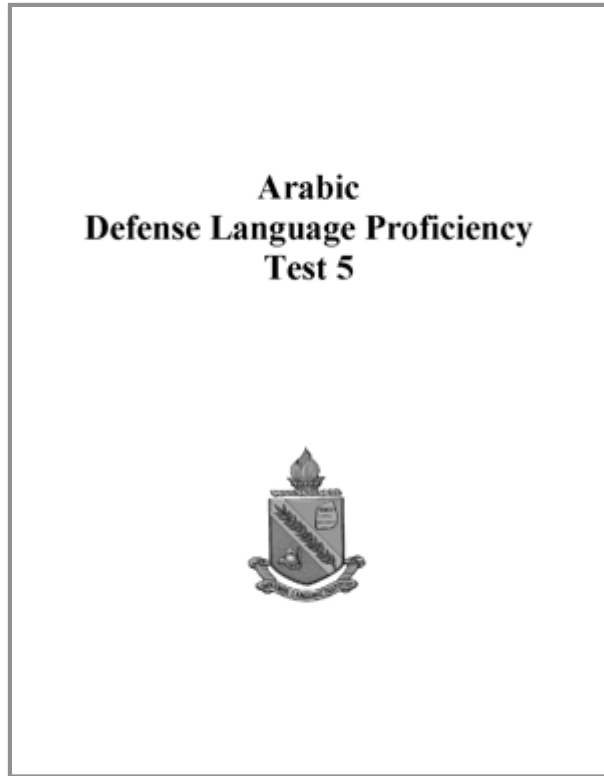
~7500 Languages – 1500 endangered ≈ 6500

104	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120	121	122	123	124	125	126	127	128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143	144	145	146	147	148	149	150	151	152	153	154	155	156	157	158	159	160	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175	176	177	178	179	180	181	182	183	184	185	186	187	188	189	190	191	192	193	194	195	196	197	198	199	200	201	202	203	204	205	206	207	208	209	210	211	212	213	214	215	216	217	218	219	220	221	222	223	224	225	226	227	228	229	230	231	232	233	234	235	236	237	238	239	240	241	242	243	244	245	246	247	248	249	250	251	252	253	254	255	256	257	258	259	260	261	262	263	264	265	266	267	268	269	270	271	272	273	274	275	276	277	278	279	280	281	282	283	284	285	286	287	288	289	290	291	292	293	294	295	296	297	298	299	300	301	302	303	304	305	306	307	308	309	310	311	312	313	314	315	316	317	318	319	320	321	322	323	324	325	326	327	328	329	330	331	332	333	334	335	336	337	338	339	340	341	342	343	344	345	346	347	348	349	350	351	352	353	354	355	356	357	358	359	360	361	362	363	364	365	366	367	368	369	370	371	372	373	374	375	376	377	378	379	380	381	382	383	384	385	386	387	388	389	390	391	392	393	394	395	396	397	398	399	400	401	402	403	404	405	406	407	408	409	410	411	412	413	414	415	416	417	418	419	420	421	422	423	424	425	426	427	428	429	430	431	432	433	434	435	436	437	438	439	440	441	442	443	444	445	446	447	448	449	450	451	452	453	454	455	456	457	458	459	460	461	462	463	464	465	466	467	468	469	470	471	472	473	474	475	476	477	478	479	480	481	482	483	484	485	486	487	488	489	490	491	492	493	494	495	496	497	498	499	500	501	502	503	504	505	506	507	508	509	510	511	512	513	514	515	516	517	518	519	520	521	522	523	524	525	526	527	528	529	530	531	532	533	534	535	536	537	538	539	540	541	542	543	544	545	546	547	548	549	550	551	552	553	554	555	556	557	558	559	560	561	562	563	564	565	566	567	568	569	570	571	572	573	574	575	576	577	578	579	580	581	582	583	584	585	586	587	588	589	590	591	592	593	594	595	596	597	598	599	600	601	602	603	604	605	606	607	608	609	610	611	612	613	614	615	616	617	618	619	620	621	622	623	624	625	626	627	628	629	630	631	632	633	634	635	636	637	638	639	640	641	642	643	644	645	646	647	648	649	650	651	652	653	654	655	656	657	658	659	660	661	662	663	664	665	666	667	668	669	670	671	672	673	674	675	676	677	678	679	680	681	682	683	684	685	686	687	688	689	690	691	692	693	694	695	696	697	698	699	700	701	702	703	704	705	706	707	708	709	710	711	712	713	714	715	716	717	718	719	720	721	722	723	724	725	726	727	728	729	730	731	732	733	734	735	736	737	738	739	740	741	742	743	744	745	746	747	748	749	750	751	752	753	754	755	756	757	758	759	760	761	762	763	764	765	766	767	768	769	770	771	772	773	774	775	776	777	778	779	780	781	782	783	784	785	786	787	788	789	790	791	792	793	794	795	796	797	798	799	800	801	802	803	804	805	806	807	808	809	810	811	812	813	814	815	816	817	818	819	820	821	822	823	824	825	826	827	828	829	830	831	832	833	834	835	836	837	838	839	840	841	842	843	844	845	846	847	848	849	850	851	852	853	854	855	856	857	858	859	860	861	862	863	864	865	866	867	868	869	870	871	872	873	874	875	876	877	878	879	880	881	882	883	884	885	886	887	888	889	890	891	892	893	894	895	896	897	898	899	900	901	902	903	904	905	906	907	908	909	910	911	912	913	914	915	916	917	918	919	920	921	922	923	924	925	926	927	928	929	930	931	932	933	934	935	936	937	938	939	940	941	942	943	944	945	946	947	948	949	950	951	952	953	954	955	956	957	958	959	960	961	962	963	964	965	966	967	968	969	970	971	972	973	974	975	976	977	978	979	980	981	982	983	984	985	986	987	988	989	990	991	992	993	994	995	996	997	998	999	1000
-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	------

≈ 1% MT



Use Standards and Doctrine



Performance Measures	GO	NO GO
1. Stopped the vehicle.	___	___
2. Informed the occupants of the reason for the search.	___	___
3. Identified the occupants by looking at their drivers' licenses or ID cards.	___	___
4. Directed the occupants to get out of the vehicle.	___	___
5. With the exception of the driver, directed the occupants to move to a place about 5 meters from the vehicle and out of the flow of traffic where they could be observed.	___	___
6. Directed the driver to open all doors and compartments, to include the ashtray, glove box and/or armrest, trunk, and hood.	___	___
7. Searched the vehicle in a sequenced manner.	___	___

STP 21-1-SMCT

18 June 2009

171-137-0001

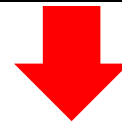
3-199



Flow of Information from Needs to Response

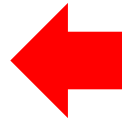
Individual Needs

WE HELP US
NEED FOOD



Aggregate Needs

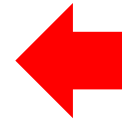
Deliver Assistance



General Mission

USAID/OFDA DoD Mission Tasking Matrix (MITaM)	
RESPONSE:	ETHAN Earthquake (EXERCISE ONLY)
	New Missions identified as of 1-Oct-11 at 1500 Local
Mission #	T-101
Priority	URGENT
WHO	WHO is Requesting US Military Assistance?
Name, Pos	Chivers, Dana
Org/Office	OFDA DART Civ-Mil Coordinator
email	dchivers@ofda.gov
phone	+1. 571.594-3937
WHAT	WHAT type of Service or Goods are Requested?
Describe as clearly as possible what you want the military to do.	Transport DART team on Aerial Recon of effected routes (Hwy1 & 2) to the south of the Capitol
WHEN	WHEN is it needed?
Date(s) & Time(s)	ASAP - request NLT 24 hours from now.
WHERE	WHERE is it needed? ...and HOW
If the request is for a static position:	
Site Name	N/A
Grid	N/A
POC on-site & contact info	N/A

Military Tasking





Flow of Information from Needs to Response

Individual Needs

WE HELP US
NEED FOOD

Deliver Assistance



Actionable Information



General Mission

USAID/OFDA DoD Mission Tasking Matrix (MITaM)	
RESPONSE:	ETHAN Earthquake (EXERCISE ONLY)
	New Missions identified as of 1-Oct-11 at 1500 Local
Mission #	T-101
Priority	URGENT
WHO	WHO is Requesting US Military Assistance?
Name, Pos	Chivers, Dana
Org/Office	OFDA DART Civ-Mil Coordinator
email	dchivers@ofda.gov
phone	+1.571.594-3937
WHAT	WHAT type of Service or Goods are Requested?
	Describe as clearly as possible what you want the military to do.
	Transport DART team on Aerial Recon of effected routes (Hwy1 & 2) to the south of the Capitol
WHEN	WHEN is it needed?
Dates(s) & Times(s)	ASAP - request NLT 24 hours from now.
WHERE	WHERE is it needed? ...and HOW
	If the request is for a static position:
Site Name	N/A
Grid	N/A
POC on-site & contact info	N/A

Military Tasking



Aggregate Needs





LORELEI HADR Topics

Low Resource Languages for Emergent Incidents

Emergency Indicators and Minimum Standards

Malnutrition Emergency Definitions

Global Acute Malnutrition (GAM) = Severe Acute Malnutrition (SAM) + Moderate Acute Malnutrition (MAM)

GAM for <5 age group

Z-Score <-2 MUAC <13.5 cm WFW/WFL <80%

SAM for <5 age group

Z-Score <-3 MUAC <11.0 cm WFW/WFL <70%

MAM for <5 age group

Z-Score >-3 and <-2 MUAC >11.0 and <13.5 cm WFW/WFL >70% and <80%

See section D of chapter III, Interpretation of Malnutrition Rates and Corresponding Actions, for malnutrition indicators.

Mortality Rate Emergency Indicators

Crude Mortality Rate (CMR): single most important indicator of serious stress in affected populations

CMR = deaths/10,000/day: emergency phase

- <1 = Under control
- >1 = Serious condition
- >2 = Out of control
- >4 = Major catastrophe

Mortality rate for <5 age group

- 1 = Normal in a developing country
- <2 = Emergency phase: under control
- >2 = Emergency phase: in serious trouble
- >3 = Emergency phase: out of control

Minimum Water Requirements

- Minimum maintenance = 15 lit
- Feeding centers = 30 lit
- Health centers and hospitals = 40-60 lit
- 1 tap stand/250 people not >100m from
- A large quantity of reasonably safe water of pure water

Minimum Food Requirements

Minimum maintenance = 2,100 Kcals/person/day

Minimum Shelter/Space

- Minimum shelter space = 3.5 m²/person
- Minimum total site area = 45 m²/person
- settled camps

Minimum Sanitation

- At least 1 toilet for every 20 persons
- Maximum of 1 minute walk from dwelling

Field Operations Guide

For Disaster Assessment and Response



USAID

Needs

- Water
- Food
- Shelter
- Medical
- Utilities, Energy, Sanitation
- Infrastructure Damage
- Search / Rescue
- Evacuation

Obstacles

- Civil Unrest
- Extreme Violence
- Regime Change





Mission Tasking Matrix (MITAM)

USAID/OFDA DoD Mission Tasking Matrix (MiTaM)	
RESPONSE: TITAN Earthquake (EXERCISE ONLY)	
New Missions identified as of 1-Oct-11	
at 1500 Local	
Mission #	T-101
Priority	URGENT
WHO WHO is Requesting US Military Assistance?	
Name, Pos	Chivers, Dana
Org/Office	OFDA DART Civ-Mil Coordinator
email	dchivers@ofda.gov
phone	+1. 571.594-3937
WHAT WHAT type of Service or Goods are Requested?	
Describe as clearly as possible what you want the military to do	Transport DART team on Aerial Recon of effected routes (Hwy1 & 2) to the south of the Capitol
WHEN WHEN is it needed?	
Date(s) & Time(s)	ASAP - request NLT 24 hours from now.
WHERE WHERE is it needed? ...and HOW	
If the request is for a static position:	
Site Name	N/A
Grid	N/A
POC on-site & contact info	N/A

what CARGO needs to be moved?		Total Weight () units	Total Volume () units
Total #/ pieces	What		
NONE	NONE		

total			
Hazmat?			
Special instructions			
PASSENGERS to be moved?		TOTAL:	5
Organization	Name & Position	Nationality	
DART	Dana Chivers, Civ-Mil Coordinator	USA	
DART	John Gambulputty, DART Tm Ldr	USA	
USAID/Titan	Bobby Houne, USAID MDRO	USA	
Min of Emerg	Goula Harnii, Vice Minister	Titan	
IOM	Francois Dubunet, IOM Field Officer	France	
Time on Objective & Special Instructions	Mr. Harnii will need to be dropped off at the village of Queriiia. All others get off with him for a 30 minute meeting vic the LZ, then need to get back on for transport back to base. Mr Harnii will stay at the village		
WHY	WHY is this requested of the military?		
Is the military your choice of last resort?	No civilian assets available at this time		

Form:MLU/007a rev 16AUG10. Previous editions obsolete



LORELEI Situation Frame Highlights

Low Resource Languages for Emergent Incidents

<Situation> ::= <Need> ! <Issue>

**<Need> ::= Water Supply ! Food Supply ! Shelter ! Medical Assistance !
Utilities, Energy, or Sanitation ! Infrastructure ! Search/Rescue
! Evacuation**

+ (Urgent ! Not Urgent) + (Current ! Past ! Future)

+ <Place>

**<Issue> ::= Civil Unrest or Widespread Crime ! Terrorism or other Extreme
Violence ! Regime Change**

<Place> ::= (<LOC> ! <GPE> ! Unknown)



LoReHLT16 Evaluations



NIST's new Low Resource human language technologies (LoReHLT) evaluation series aims to advance HLT that can provide rapid and effective response to emerging incidents where the language resources are very limited. Participation in LoreHLT is open to anyone who finds the task(s) of interest. LoReHLT 2016 offers three tasks:

- Machine translation (MT)
- Situation Frame (SF)
- Named entity recognition (NER)

Highlights

- Surprise language evaluation
- Two training conditions: constrained (required) and unconstrained
- Three evaluation checkpoints to gauge performance based on time and training resources given



Data from Haiti 2010 Earthquake

- **Mission 4636**
 - SMS Service during 2010 Earthquake for people to send messages requesting help
 - <http://www.mission4636.org/>
- **Ushahidi Collaboration Platform**
 - Volunteers translated and annotated requests and shared with responders
 - <http://www.ushahidi.com/>
- **Data publicly available**
 - [pydata book on github.com](#)
 - [Haiti Crisis Map on datahub.io](#)





Sample Mission 4636 Message*

Nou nan zòn tigwav nou ta renmen pou nou èd men fok se pajan mwen wè mesye minista ap fè moun yo nou bezwen anpil tant ak medikaman pou grip la fyè ak no Time: 2010-01-28 23:31:13

LORELEI seeks to enable English speakers to respond in HADR scenarios without knowing the local languages.



Message #2517* from Haiti 2010 Earthquake

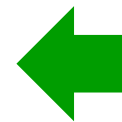
Nou nan zòn **tigwav** nou ta renmen pou nou èd men fok se pajan mwen wè mesye minista ap fè moun yo **nou bezwen anpil tant ak medikaman pou grip** la fyè ak no **Time: 2010-01-28 23:31:13**



Situation Frame

ID HT2010EQ01-U2517
Type Need Shelter, Medical
Entities UNK
LOC Petit Goave
Time 2010-01-28 23:31:13
SEC NONE
Subtopic NONE

From Gazetteer:



Name Petit Goave
Equiv tigwav
LAT 18.37743
LON -72.93157

*Ushahidi Data: <https://github.com/pydata/pydata-book/blob/master/ch08/Haiti.csv>



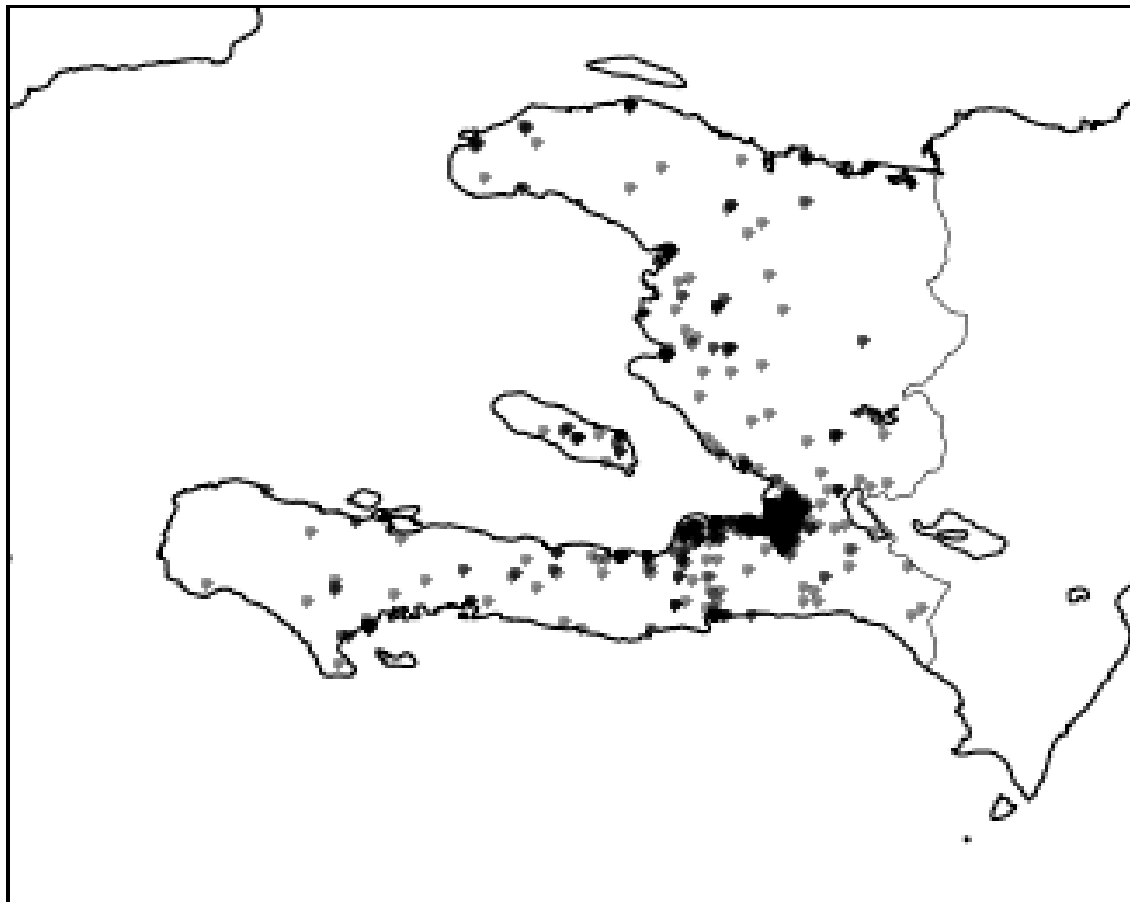
“Signal” in Low Resource Language “Noise”

- Trends and “Dots on a Map”
- Use confidence levels to help aggregate information



Dots on a Map

2a: Food Shortage



- <https://github.com/pydata/pydata-book/blob/master/ch08/Haiti.csv>
- <http://datahub.io/dataset/ushahidi/resource/81d058a8-173a-49d9-8ce9-4edf5e7cafc9>



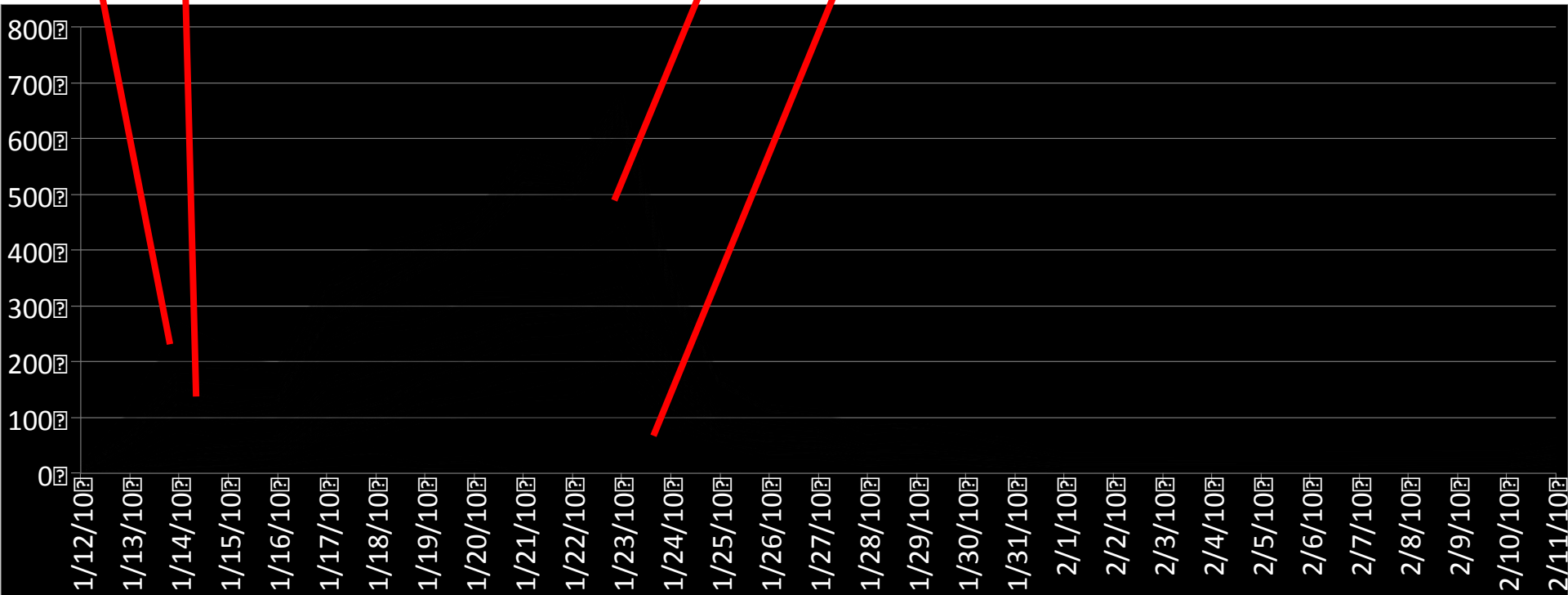
Trends in Message Categories

Persons News

Water Shortage

Collapsed structure

Food Shortage



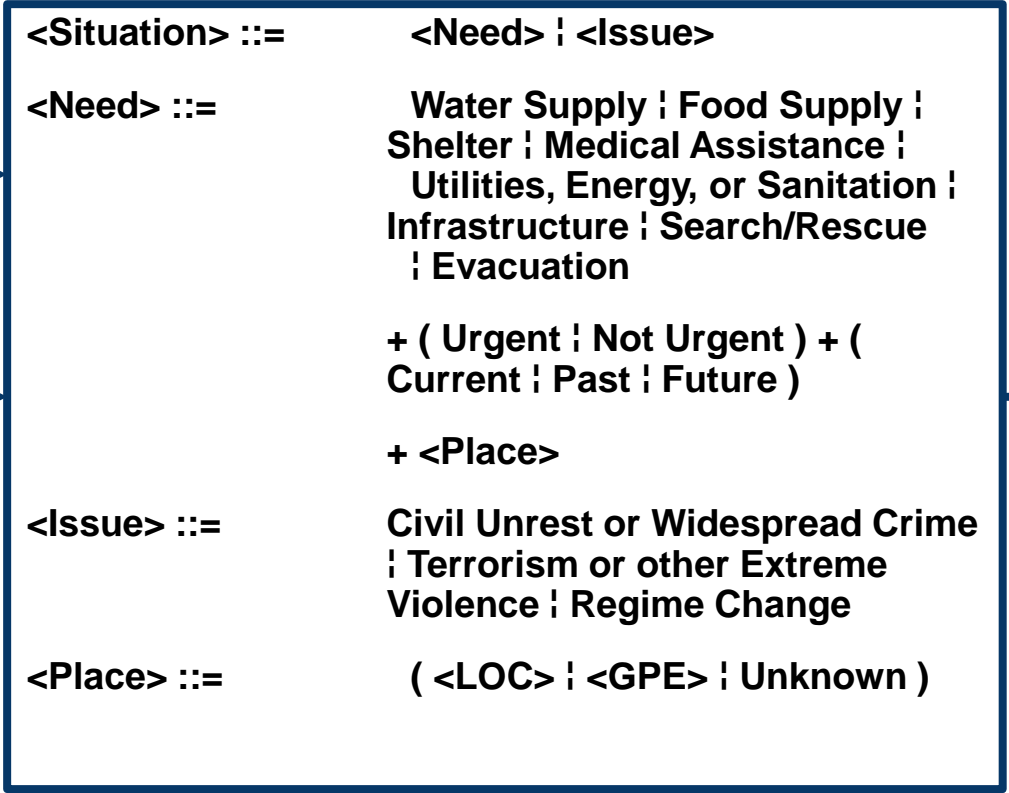


MT may be neither necessary nor sufficient for HADR mission relevance

Foreign Language

Foreign Language

English



High volume, uncertain inputs acceptable if aggregate output is useful



Concept of Evaluation

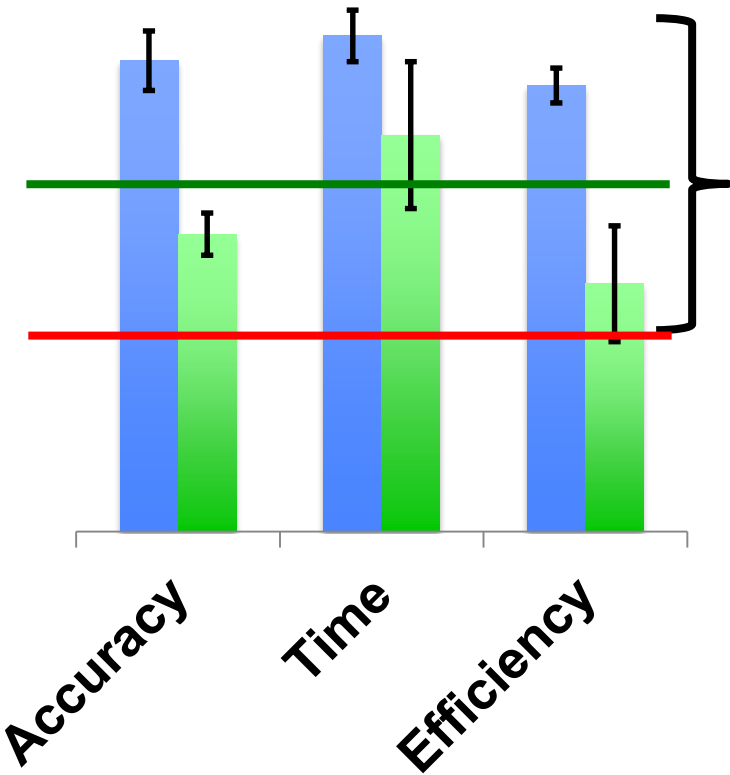
- **Inject LORELEI analytics into HADR exercise from low-resource language messages (social media, i.e.: Twitter)**
- **Measure outcomes: i.e.:**
 - Accuracy of Decision
 - Time to Right Decision
 - Reduction of Overlapping Effort (Efficiency)
- **Vary conditions for Situation Frames (SF) from Low Resource Languages: i.e.:**
 - Reference SF
 - System SF



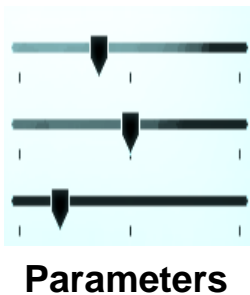
Concept of Experiment

Situation Frames

■ Reference
■ System



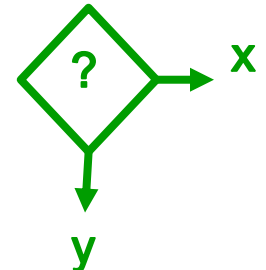
Technology Opportunity Zone



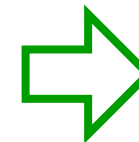
HADR Participant Experiments



Situation Frame



Decision





THANKS

Doug Jones DAJ@LL.MIT.EDU



DragonFly

Wearable Devices to Enable Communication via ASL

Presented at the
Association for Machine Translation in the Americas (AMTA) 2016 Conference
Austin, Texas
October 31, 2016

Overview

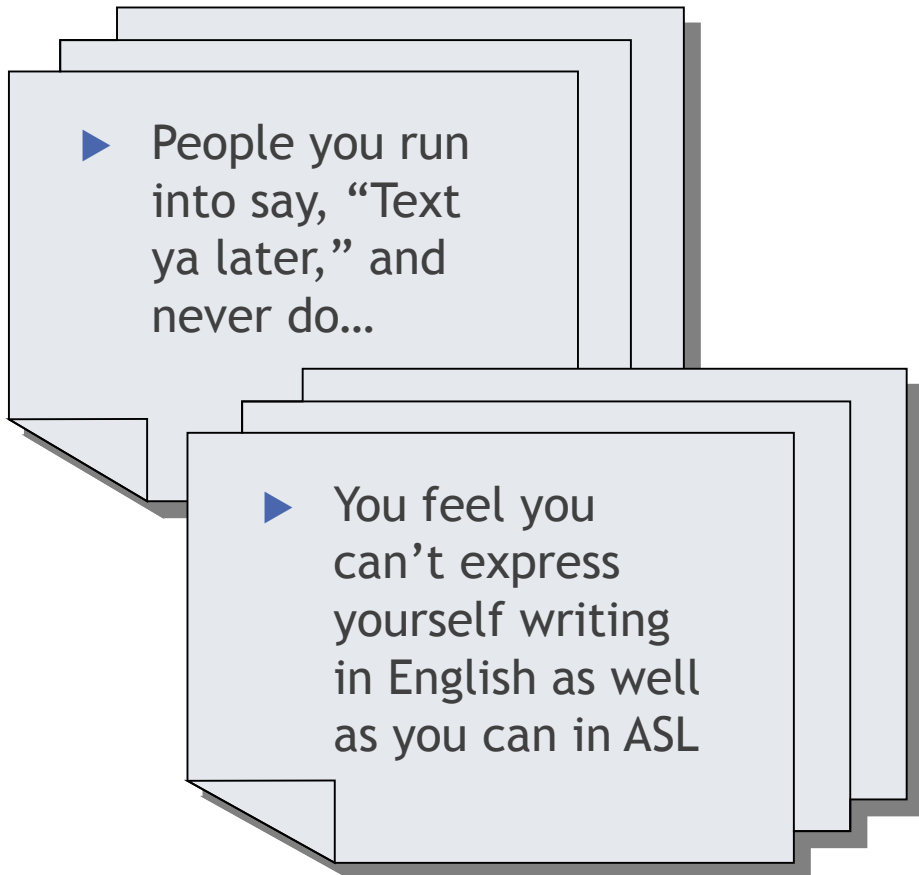
- ▶ Statement of User Need
- ▶ Our Technical Approach
- ▶ Project Status & Next Steps
- ▶ Wrap-Up

Statement of User Need

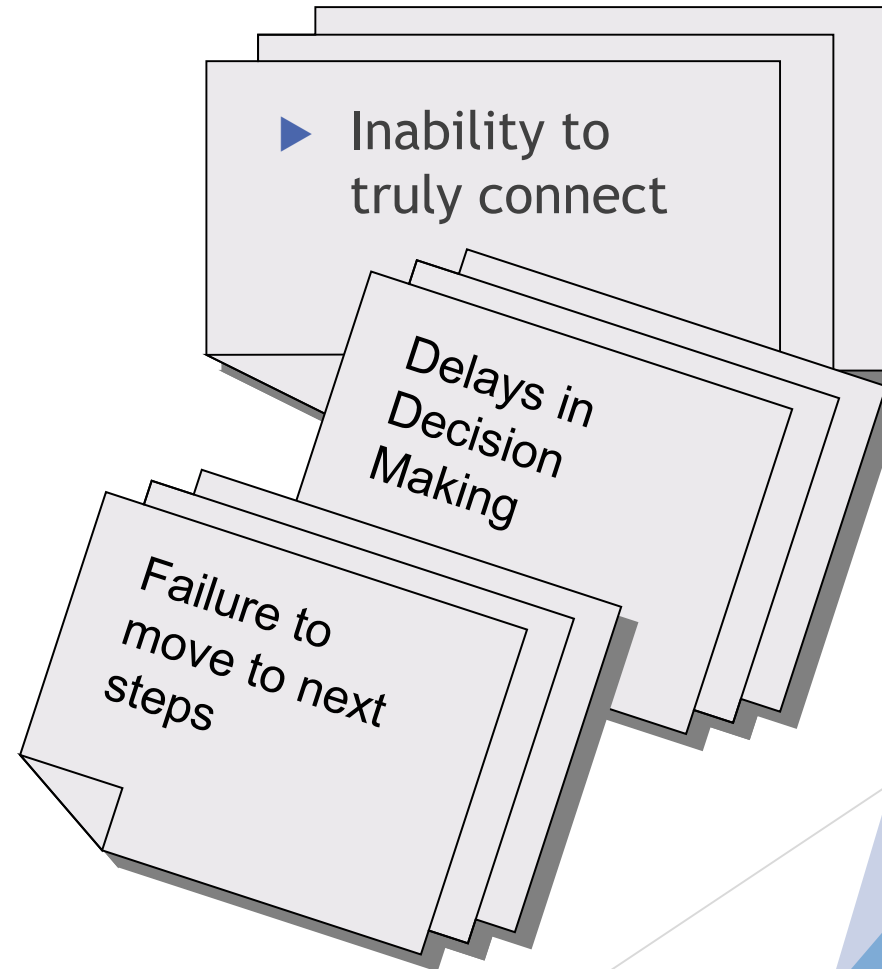
- ▶ There is a need for to enable signers to seamlessly communicate with non-signers.
- ▶ In day-to-day life, people do not have dedicated interpreters who are with them on a 24 X 7 basis.
- ▶ Obtaining an interpreter requires prior arrangements & advanced planning, sometimes weeks in advance.
- ▶ Ad-hoc meetings come up, and the chance to participate is lost if an interpreter can't be found.

Current Situation

Some of the Problems...



Some of the Consequences...



Solving the Challenge

- ▶ Develop a tool to allow an ASL signer and a speaker of English to communicate with each other...
 - ▶ Face-to-Face
 - ▶ Naturally
 - ▶ Anytime
 - ▶ Anywhere

Our Solution

- ▶ To develop wearable devices that will facilitate interactions between signers and non-signers
 - ▶ This family of solutions is called *DragonFly*
 - ▶ We envision *DragonFly* running on different types of wearables



Why?

- ▶ To help bridge the current communication barrier...
 - ▶ With *DragonFly*, a person who uses ASL and someone who does not know ASL can express themselves completely to one another.
 - ▶ The opportunity *DragonFly* creates is the ability to unlock the potential of every person to fully contribute to the mission.

Technical Approach

Technical Challenges

Sign/Signer Variability



Sensor Variability



Signal Complexity



Session Variability

e.g. observation angle



Data Availability

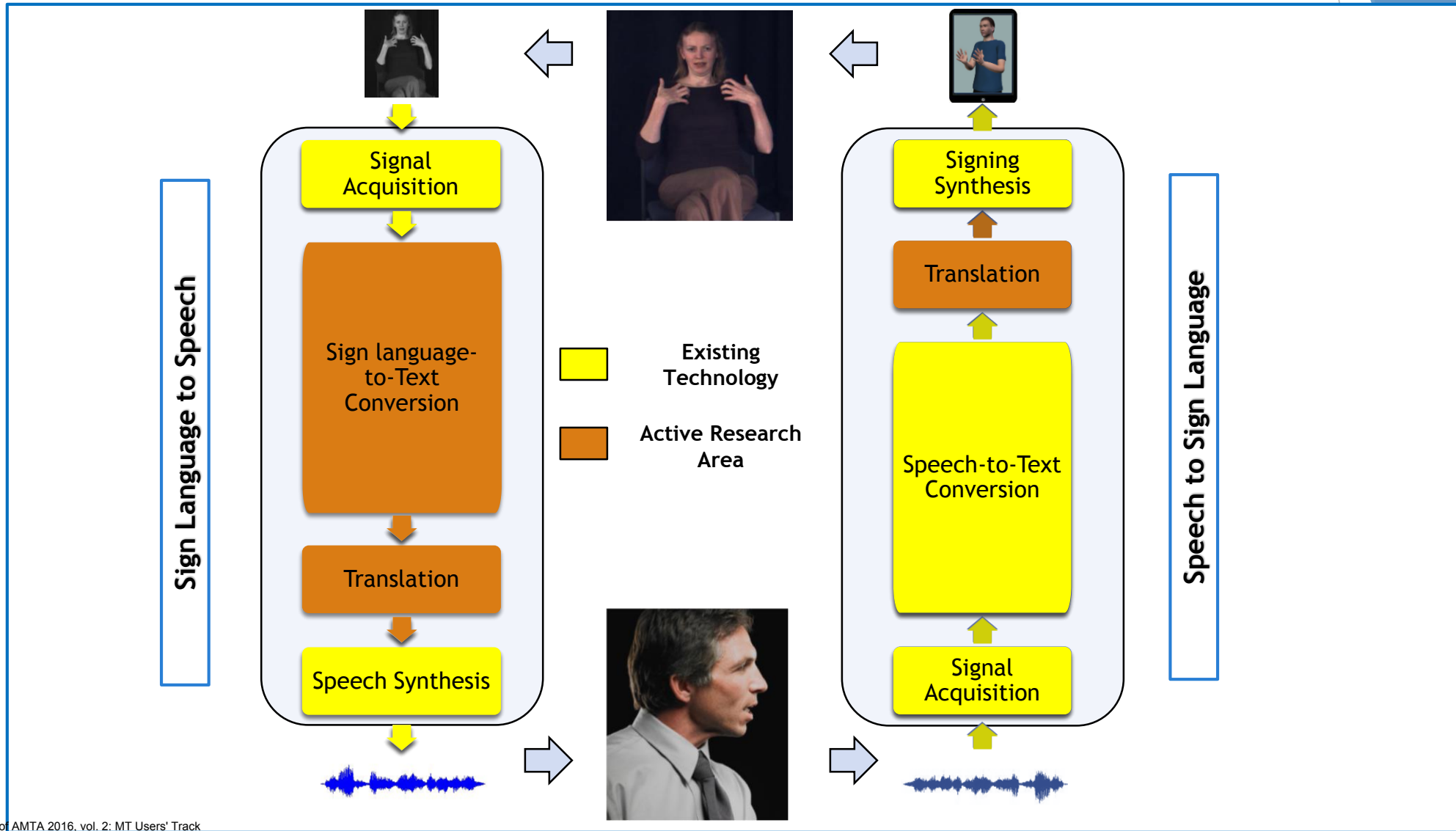
- Limited availability of well annotated ASL \leftrightarrow English content for development and evaluation (e.g. ASLLVD)
- Technical challenges remain for exploiting loosely annotated content (e.g. ASL w/ closed captioning)



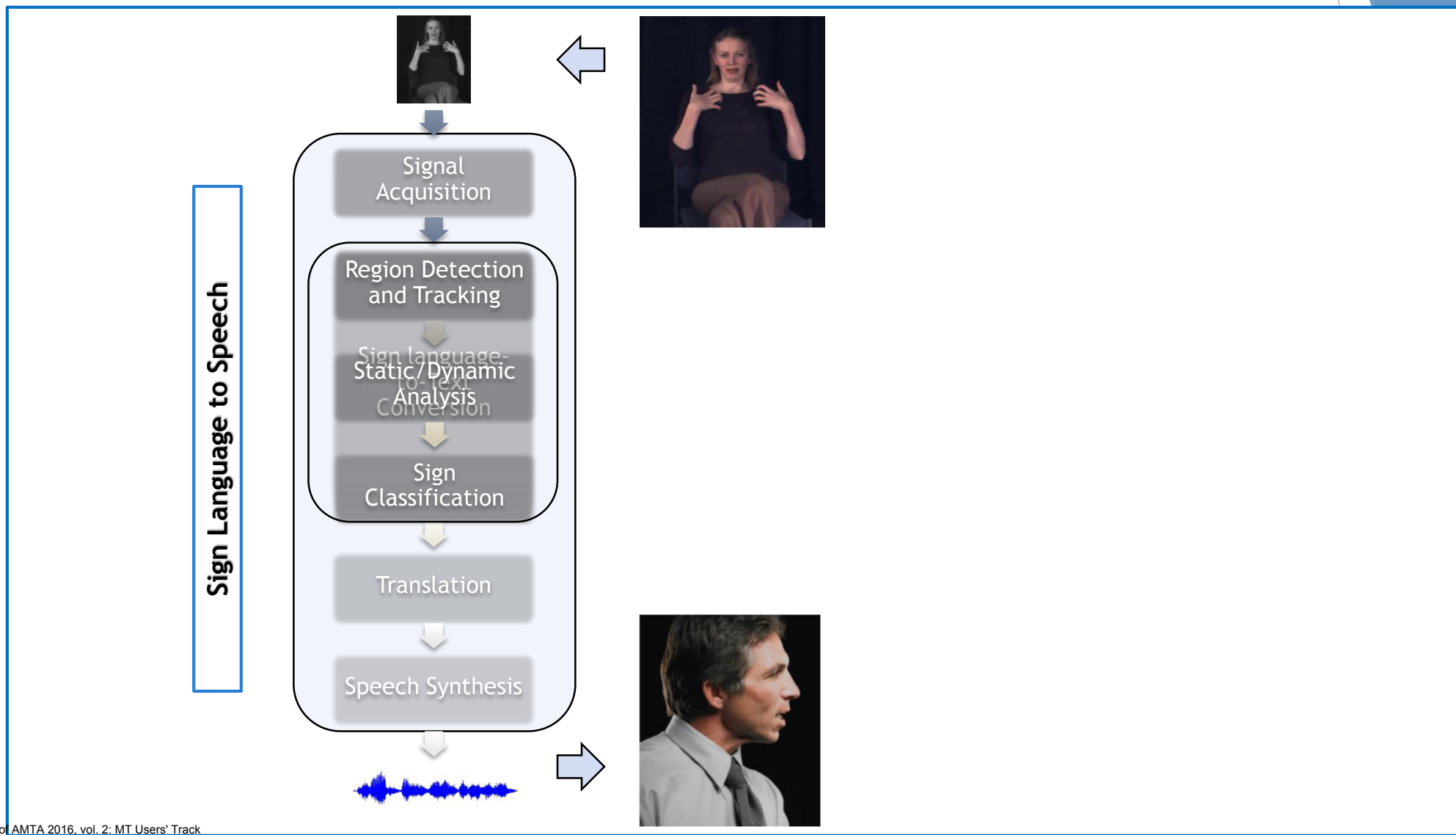
ASL: American Sign Language

ASLLVD: American Sign Language Video Dataset

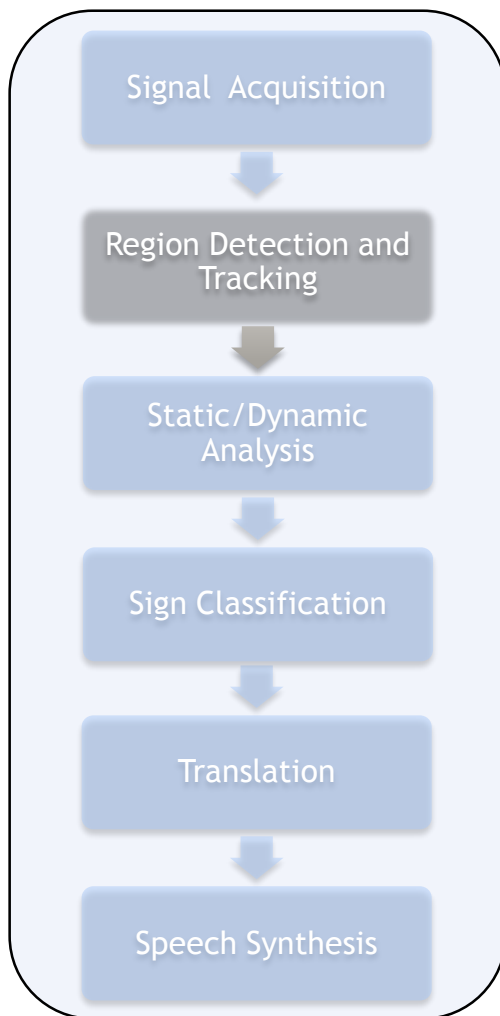
System Overview



Sign Language to Speech



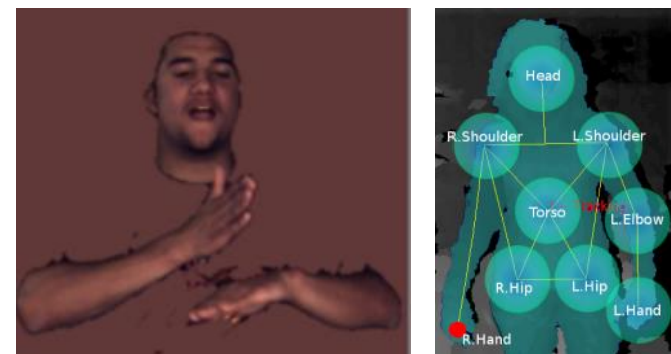
Sign Language to Speech



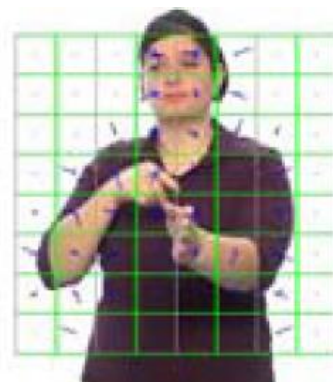
- **Signer Isolation**



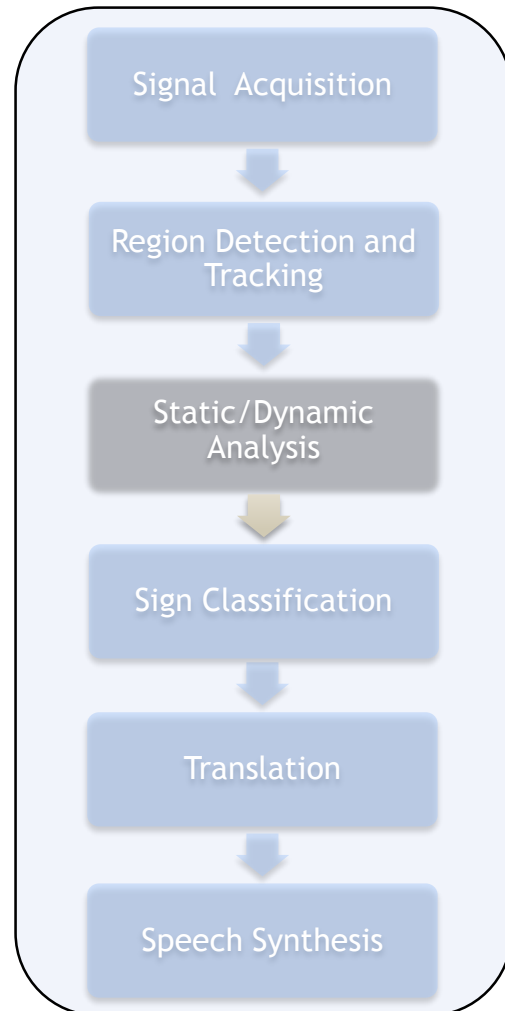
- **Feature Localization**
 - Face and Hands
 - Whole Body



- **Motion Tracking**



Sign Language to Speech



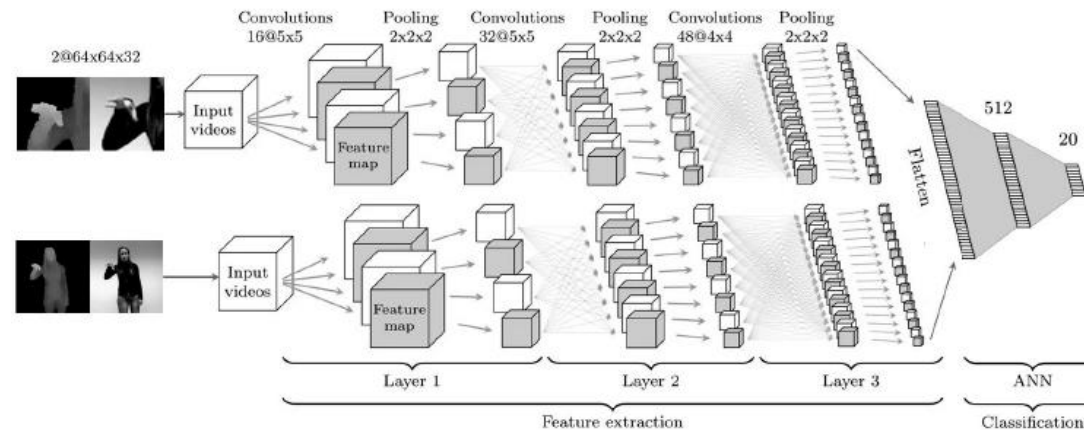
- **Classify Individual Hand Gestures**



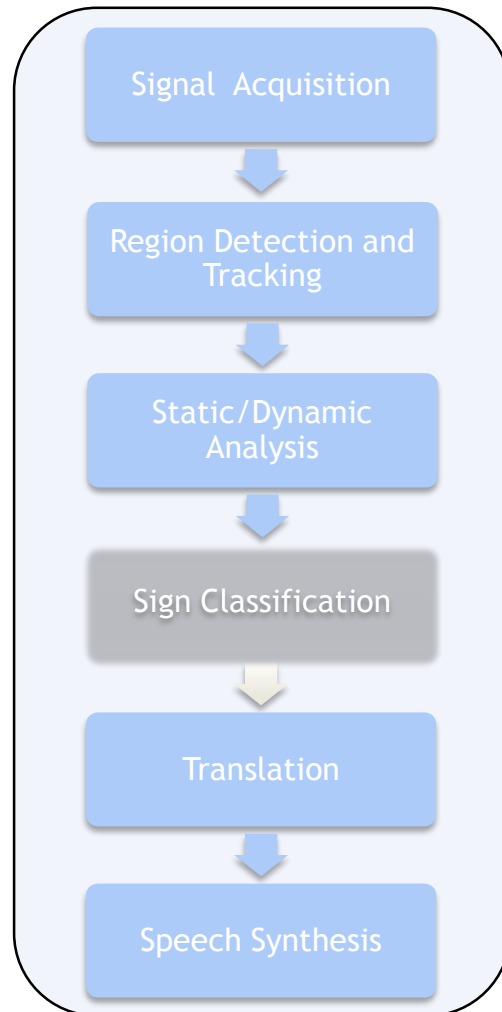
- **Evaluate Facial Cues**



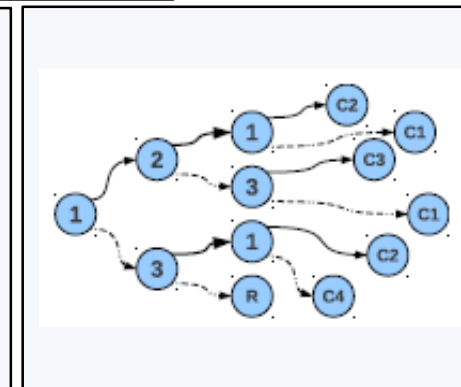
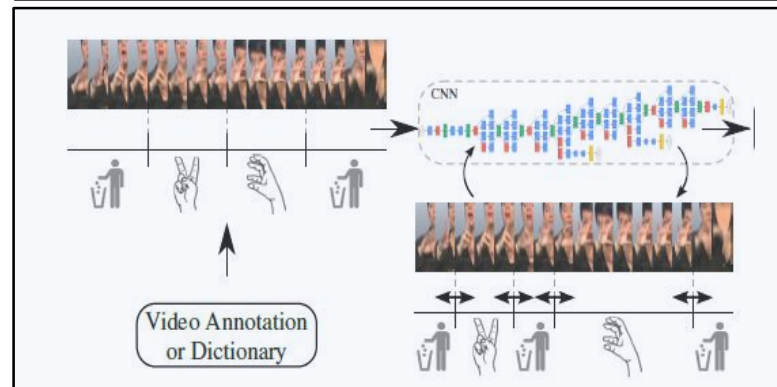
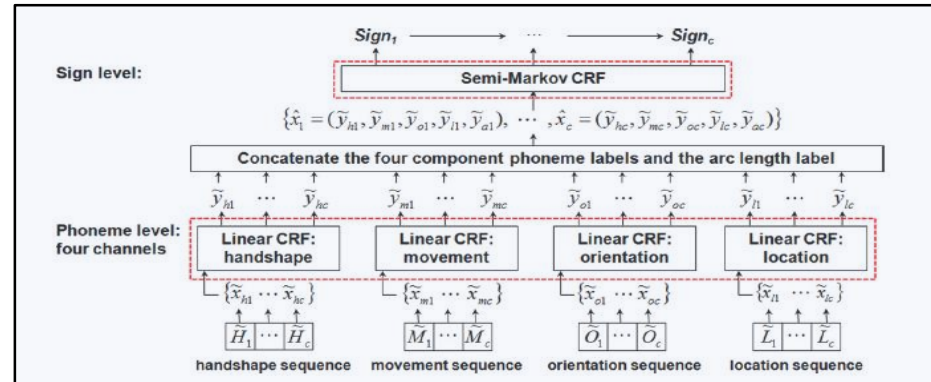
- **Performed on single frame or short series of video**
- **Typically adopt image machine learning methods.**



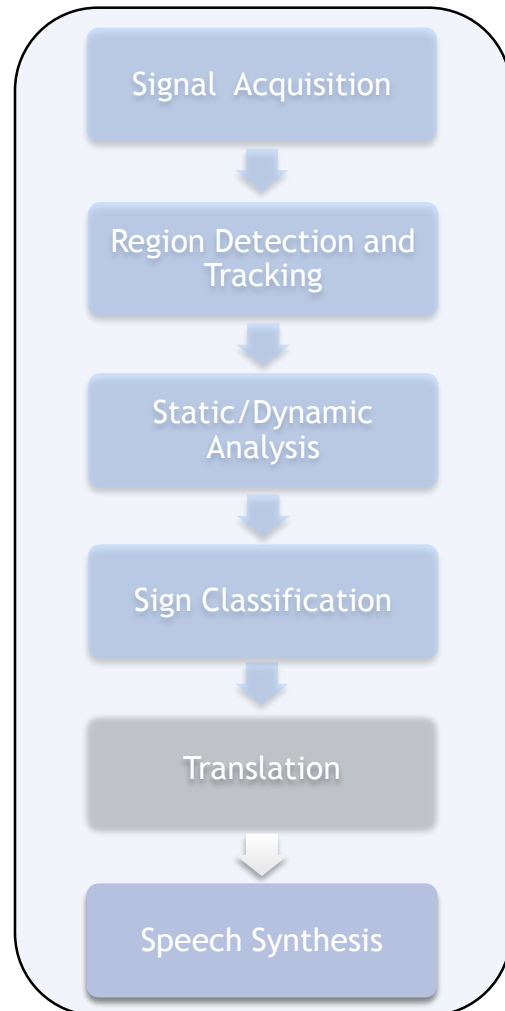
Sign Language to Speech



- Given gesture sequences, we want to identify discrete signs
- Guided by a lexicon (e.g. signing dictionary)
- Typically involves dynamic machine learning methods
 - **Markov Modeling**
 - **Neural Networks**
 - **Sequential Pattern Boosting**



Sign Language to Speech



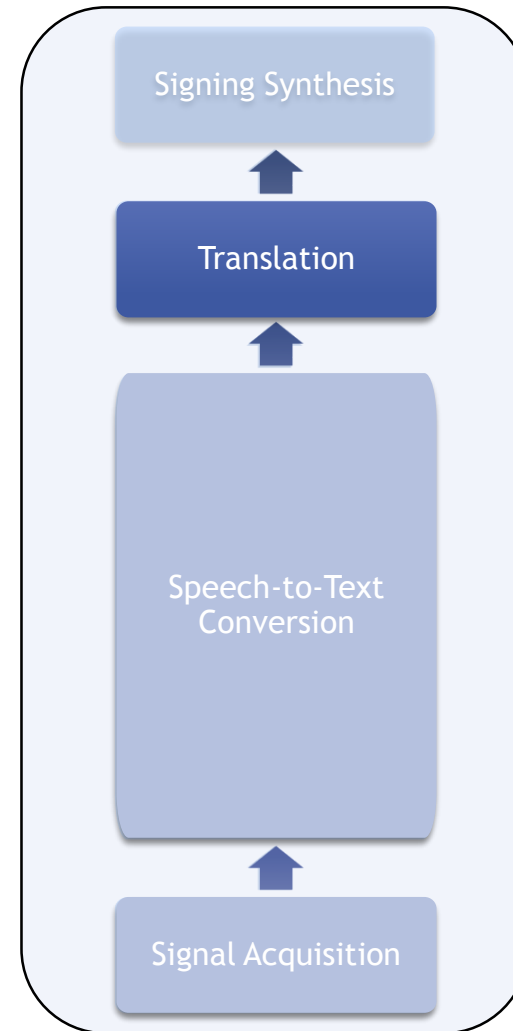
Translation:

- Convert recognized signing sequences into an English sentences.
- Example of Machine Translation (MT) Problem
- ASL is not structured like English and is more like Japanese, in that it is a Topic-Comment language. It must be ordered correctly before it is converted to speech, so that it conforms to English syntax and is readily understood by an English speaker.
- Requires sizable database of parallel ASL-English data
e.g. Television Corpus of Closed-Caption + ASL

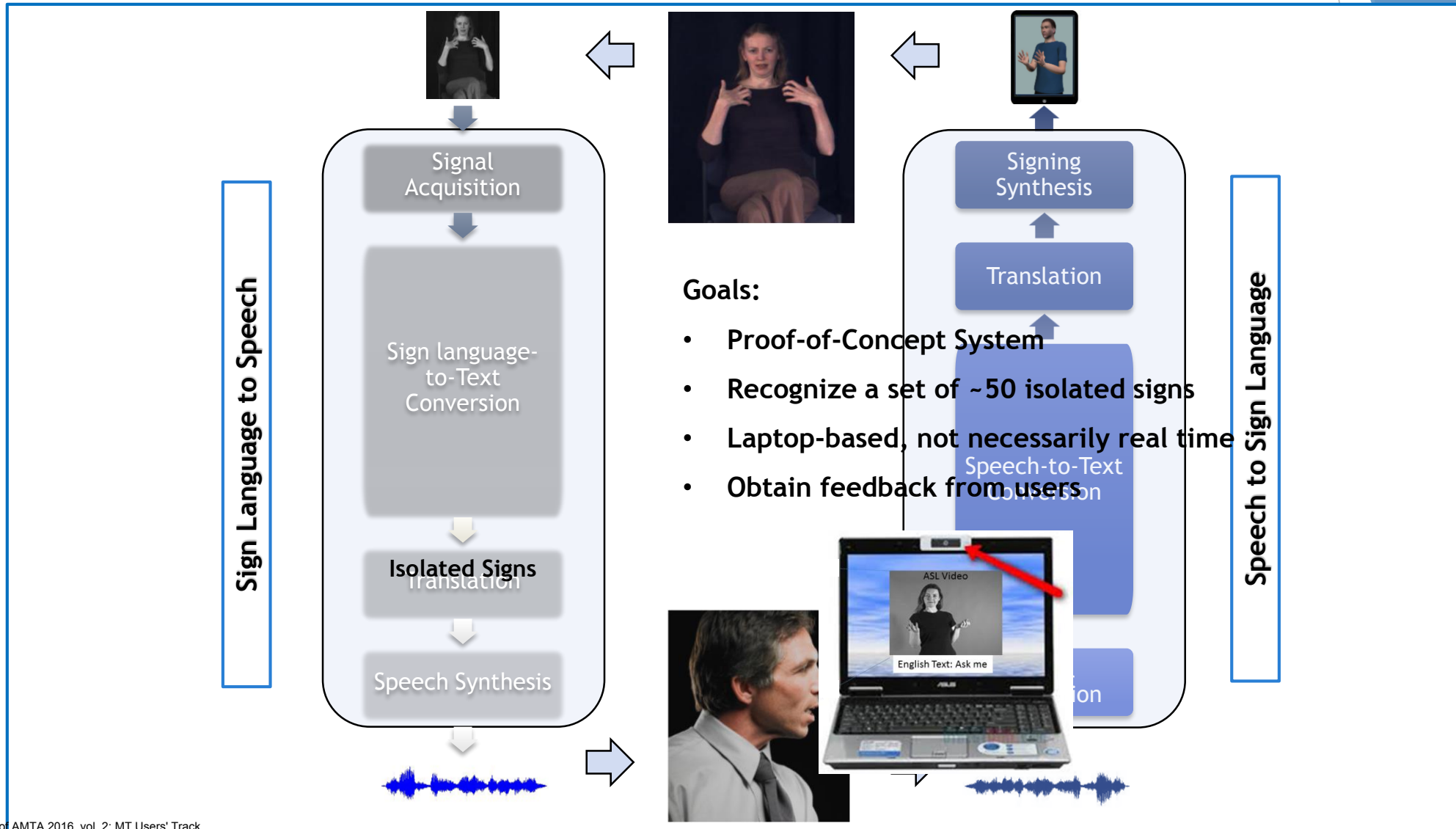
Speech to Sign Language

Translation

- Inverse operation of the MT problem
- English sentences converted into sequences of manual and non-manual gestures



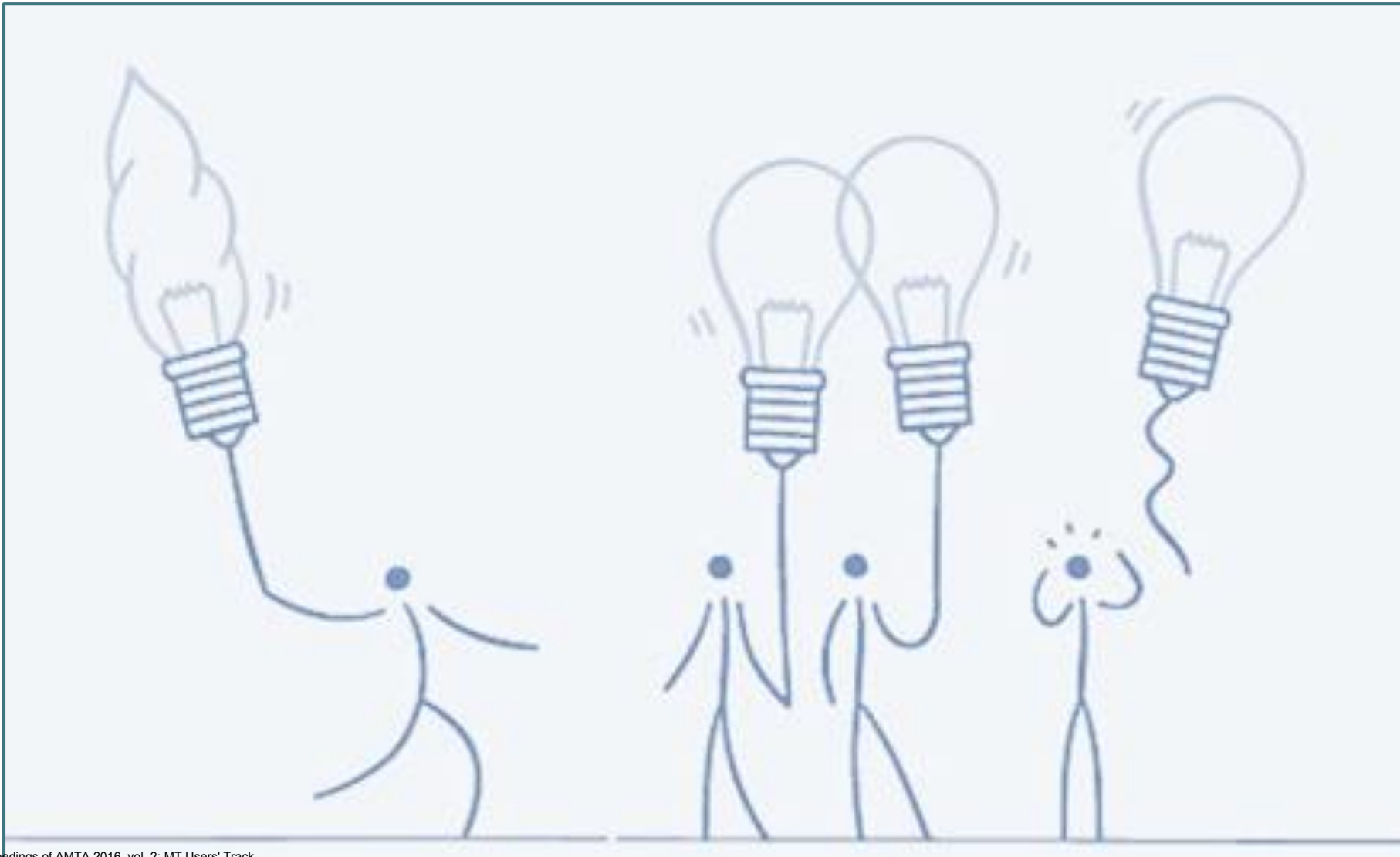
Next Steps



Project Status

- ▶ We have initiated an extensive literature search to leverage best practices and relevant research done to date.
- ▶ We are compiling data, annotation, algorithmic, and system requirements
 - ▶ Identified and aggregating annotated ASL datasets
 - ▶ ASLLVD RVL-SLLL Gallaudet
 - ▶ Identified relevant CNN models for feature extraction
 - ▶ VGG and DeepHand models
- ▶ We have begun work on a prototype for ASL recognition capabilities.

Questions?



Thank You!

Contact Info:

po17b@icloud.com

nmalyska@ll.mit.edu

Tuning Neural MT

Guido Zarrella

MITRE Corporation



**Guido
Zarrella**

**Ivan
Young**

**Becky
Marvin**



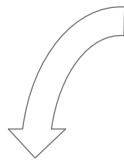
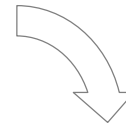
Outline

- **Tuning MT: when the system you have isn't the system you need**
- **Neural MT tuning methods differ from those for Statistical MT**
- **Genre or Domain matters (a lot):**
 - In-genre test: BLEU = **25.6**
 - Out-of-genre test: BLEU = **7.5 (-18.1)**
- **You care about NMT tuning because...**
 - Tuned w/ monolingual data only: BLEU = 10.3 (**+2.8**)
 - Trained on a small parallel set: BLEU = 13.5 (**+6.0**)
 - Tuned (transfer learning): BLEU = 15.0 (**+7.5**) to 16.9 (**+9.4**)

Tuning a system you have, to get the system you need



Tuning a system you have, to get the system you need

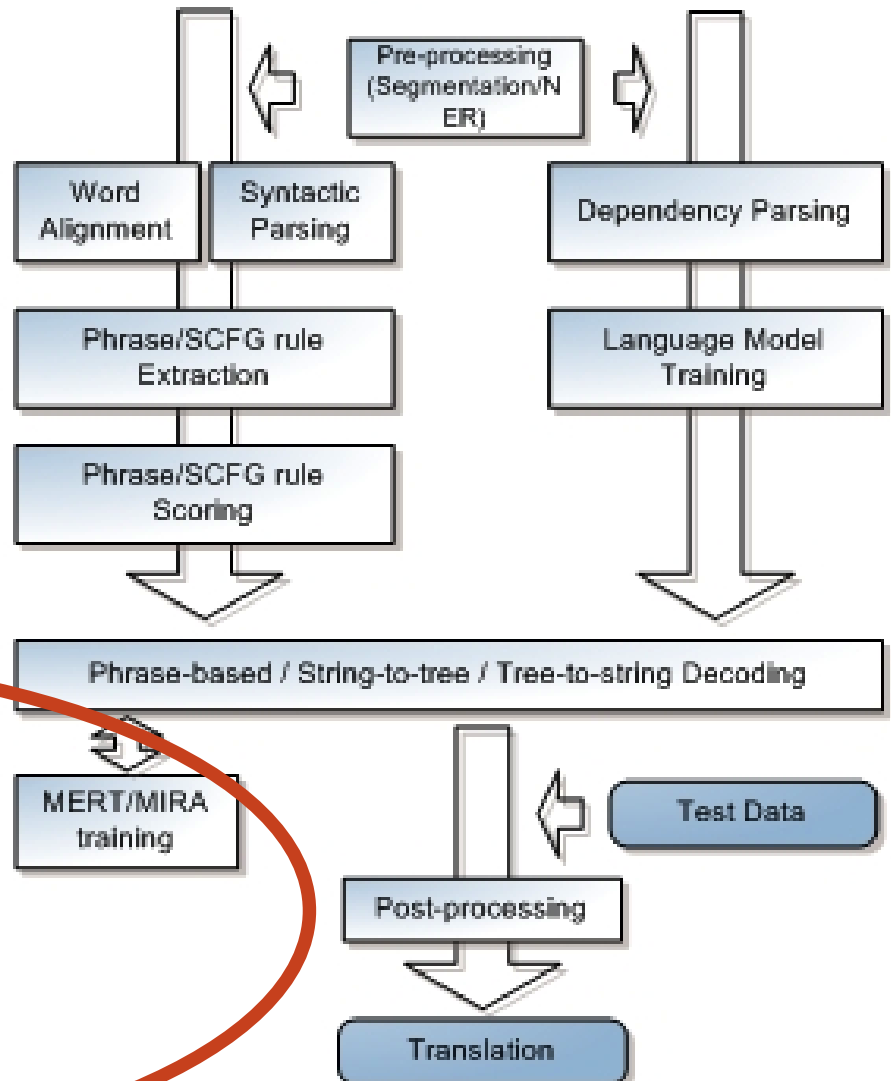


Tuning Machine Translation

In SMT, tuning involves learning a **weighted combination** of scoring features output by trained components: **translation tables, language models, reordering models, ...**

For example: **Minimum Error Rate Training (MERT)**
or **Margin-infused Relaxed Algorithm (MIRA)**

Tuning Statistical Machine Translation



Pair	System	untuned	MERT-tuned
fr-en	WMT-SMALL	28.0	29.2 (0.2)
	WMT-LARGE	29.4	32.5 (0.1)
de-en	WMT-SMALL	25.0	25.3 (0.1)
	WMT-LARGE	26.6	26.8 (0.2)

SampleRank Training for Phrase-Based Machine Translation
Barry Haddow, Abhishek Arun, Philipp Koehn 2011

Need for Domain Adaptation

News wire source

目前日本有关方面已经派出三只 巡逻艇, 协同韩国方面在出事水域开展搜寻遇难者的工作.

Semiconductor source

利用在线应力测试技术表征了掺入Pt后对镍硅化物薄膜应力性质的影响.

Human translation

Currently, Japanese authorities have three dispatched patrol boats to coordinate with the South Koreans in searching for the victims in the area of the incident.

The effect of Pt doping on the stress in the nickel silicide film has been characterized using an in-situ stress measurement.

Machine translation

Japan has dispatched three patrol boats to the area, in coordination with the South Koreans to search for the victims in the area of the incident work

Stress tests use online technology characterized by incorporation of Pt on nickel silicide films nature of the stress

Quite poor on novel domains

Need for Domain Adaptation

System	Description	Score (BLEU)	
		Semi-conductor	Chem-bio
L	Stand-alone product, statistical	9.4	9.7
S	Stand-alone product, rule-based	11.2	11.9
G	Web-based, statistical	15.1	22.8*
MITRE	Statistical	16.1	17.9

Neural MT

“With the exception of fr-es and ru-en the neural system is always **comparable or better** than the phrase-based system.”

Is Neural Machine Translation Ready for Deployment? A Case Study on 30 Translation Directions

Marcin Junczys-Dowmunt, Tomasz Dwojak, Hieu Hoang

Neural Machine Translation

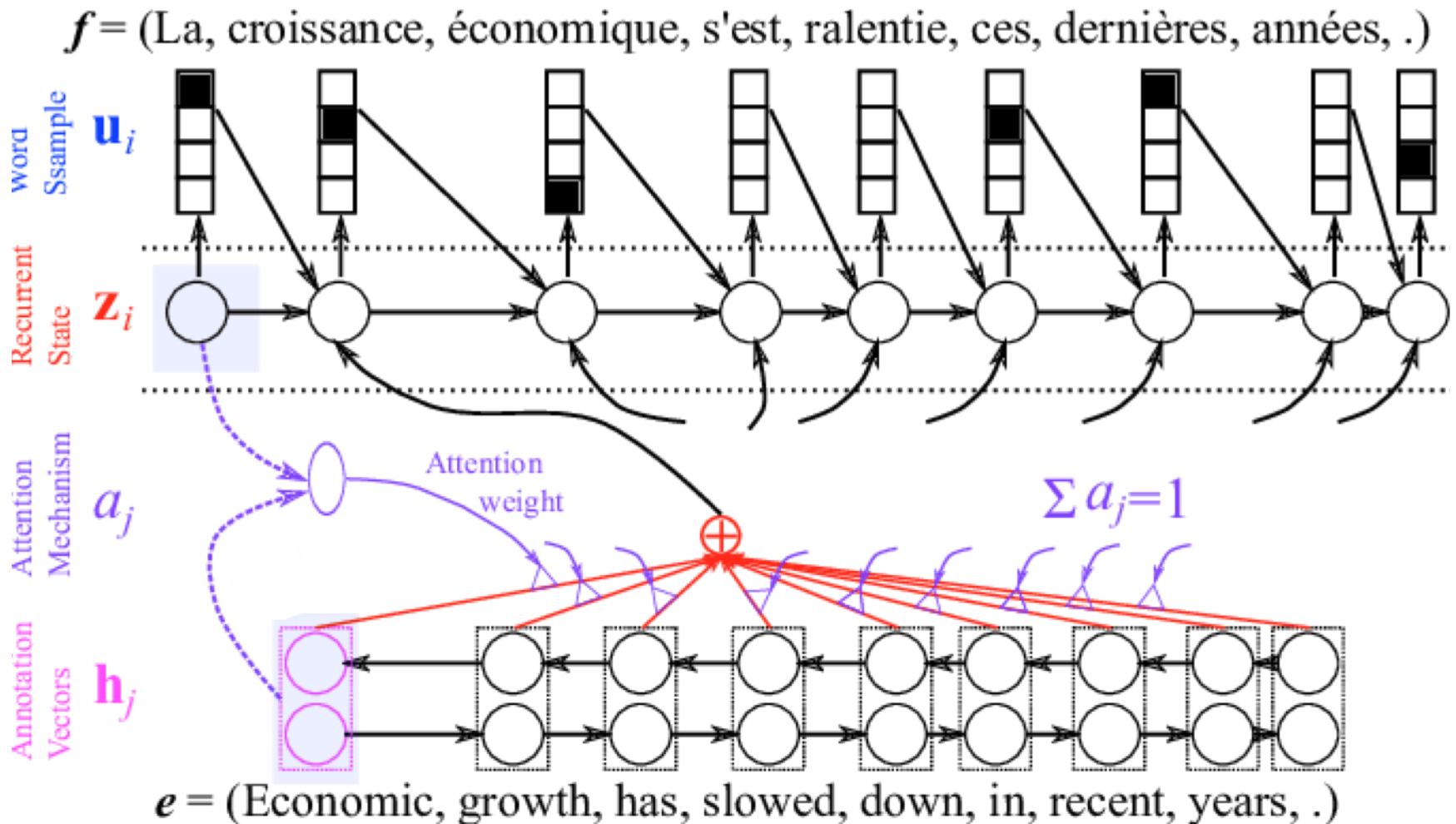


Image credit: Kyunghyun Cho: <https://devblogs.nvidia.com/parallelforall/introduction-neural-machine-translation-gpus-part-3/>

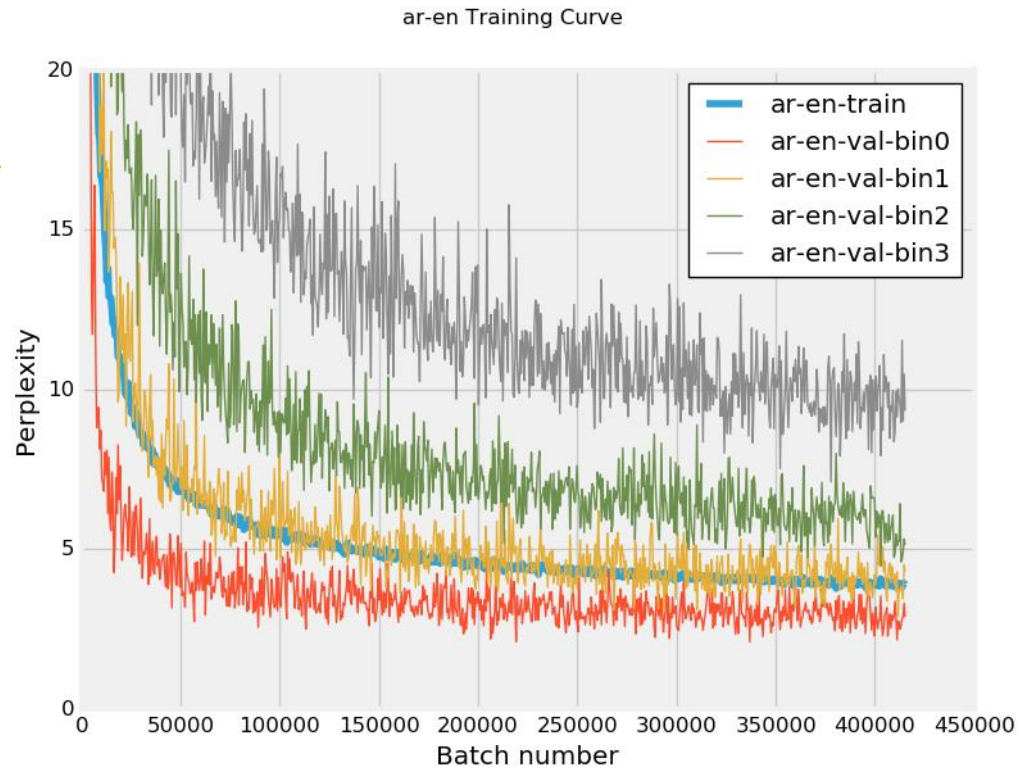
Subtitle Corpus for Discourse

Pierre Lison and Jörg Tiedemann, 2016, [OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles](#). In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016).

language	files	tokens	sentences	af	ar	bg	bn	br	bs	ca	cs	da	de	el	en	eo	es	et	eu	fa	fi	fr	g
af	32	0.2M	27.4k		6.2k	7.6k			1.8k		10.5k	6.0k	7.9k	11.6k	16.2k		12.6k	2.2k		2.1k	2.8k	7.3k	
ar	67,608	329.8M	60.8M	6.2k		16.2M	62.2k	13.0k	6.1M	0.3M	16.5M	7.5M	7.1M	15.3M	19.4M	19.4k	18.3M	6.9M	0.1M	3.0M	10.8M	14.4M	44
bg	90,376	523.4M	80.2M	7.7k	17.8M		60.7k	13.8k	7.5M	0.3M	21.1M	8.2M	8.9M	19.3M	26.4M	23.4k	24.8M	7.7M	0.1M	2.8M	13.2M	18.5M	48
bn	76	0.6M	0.1M		64.1k	62.7k			36.6k	3.1k	61.1k	58.2k	54.7k	58.5k	69.3k		65.8k	56.5k	3.1k	44.8k	56.1k	59.3k	
br	32	0.2M	23.1k		13.3k	14.1k			2.7k	5.3k	14.5k	10.0k	7.5k	14.4k	17.7k	1.1k	15.6k	15.0k	0.7k	4.4k	8.1k	15.4k	0
bs	30,511	179.5M	28.4M	1.8k	12.2M	8.5M	37.7k	2.7k		0.1M	7.5M	3.7M	3.6M	7.3M	9.5M	7.4k	9.0M	3.5M	76.3k	1.3M	5.2M	6.8M	27
ca	711	4.0M	0.5M		0.3M	0.3M	3.2k	5.5k	0.1M		0.3M	0.2M	0.2M	0.3M	0.4M		0.4M	0.2M		96.2k	0.2M	0.3M	11
cs	125,126	715.3M	112.8M	10.7k	18.1M	24.7M	63.3k	14.8k	8.5M	0.4M		8.5M	9.3M	19.8M	27.5M	31.7k	25.9M	7.9M	0.1M	2.9M	13.7M	19.1M	68
da	24,079	162.4M	23.6M	6.1k	8.0M	9.3M	60.9k	10.1k	4.0M	0.2M	9.6M		4.9M	8.1M	9.4M	11.3k	9.1M	5.0M	87.7k	2.1M	7.9M	7.6M	28
de	27,742	186.3M	26.9M	8.0k	7.6M	10.0M	57.2k	7.7k	4.0M	0.2M	10.6M	5.4M		9.1M	11.5M	24.9k	10.8M	4.3M	75.7k	1.8M	6.9M	9.2M	52
el	114,230	683.1M	101.6M	11.8k	16.8M	22.3M	60.5k	14.6k	8.1M	0.3M	23.0M	9.1M	10.2M		25.6M	24.5k	24.5M	7.5M	0.1M	2.8M	13.1M	19.6M	66
en	322,294	2.5G	336.6M	16.7k	21.9M	31.6M	75.0k	18.5k	11.1M	0.4M	33.8M	11.0M	13.4M	30.4M		49.0k	40.0M	8.6M	0.2M	3.3M	16.8M	28.0M	0.1
eo	89	0.5M	79.3k		19.9k	24.3k		1.1k	7.6k		32.8k	11.7k	25.6k	25.2k	51.1k		38.6k	17.6k		5.1k	18.9k	28.3k	0
es	191,987	1.3G	179.2M	12.9k	20.3M	29.2M	69.1k	16.0k	10.2M	0.4M	30.7M	10.4M	12.4M	28.6M	50.1M	40.2k		8.3M	0.2M	3.1M	15.7M	25.8M	0.2
et	23,515	140.7M	22.9M	2.2k	7.5M	8.9M	58.6k	15.4k	4.0M	0.2M	9.2M	5.7M	4.8M	8.6M	10.3M	18.2k	9.6M		93.3k	1.9M	6.5M	6.9M	29
eu	188	1.4M	0.2M		0.1M	0.1M	3.3k	0.7k	80.9k		0.1M	93.2k	80.1k	0.2M	0.2M		0.2M	0.1M		43.1k	0.1M	0.1M	10
fa	6,469	44.3M	7.4M	2.1k	3.1M	2.9M	46.3k	4.4k	1.4M	0.1M	3.1M	2.2M	1.9M	3.0M	3.6M	5.2k	3.3M	2.1M	44.7k		2.4M	2.5M	21
fi	44,594	208.5M	38.7M	2.8k	11.5M	14.8M	57.9k	8.3k	5.7M	0.2M	15.3M	9.0M	7.6M	14.6M	19.2M	19.5k	17.7M	7.4M	0.1M	2.5M		12.5M	40
fr	105,070	672.8M	90.9M	7.5k	15.5M	21.3M	61.4k	16.3k	7.4M	0.3M	21.8M	8.5M	10.3M	22.2M	33.5M	29.1k	30.1M	7.8M	0.1M	2.7M	13.9M		93
el	370	1.0M	0.2M		45.8k	40.7k		0.5k	28.0k	11.0k	71.3k	20.4k	54.5k	68.8k	0.2M	0.2k	0.2M	20.0k	10.5k	22.0k	40.8k	06.1k	

Arabic to English

- **Trained on 21 million conversational segments from movie subtitles**
 - 256 million training steps (sentences)
 - 19 days on K40 GPU
- **NMT BLEU = 25.6**
 - SMT BLEU = 25.3
- **Serialized as 536 MB model**
 - Deployable to laptops



“26 BLEU”

OpenSubtitles Reference

people would think that he was the terrorist. right.

- there's a boy in the cage.

we're just here to see our friend rigby, sir.

- glass is all over the floor. - somebody broke the stereo.

like nathan.

let's get ice cream.

he's down checking a buoy in the channel.

oh, my god, please.

cervical lymph node has black flecks.

you came for your uncle's wedding.

yeah, and doctors say i should get more and more each day.

NMT Output

people say he was a terrorist . right .

- there ' s a boy in the cage .

we ' re just here to see our friend , sir .

the glass is all around someone .

like a . .

let ' s go get ice cream .

he ' s out there asking for a consult .

oh , god , please .

the black is a black .

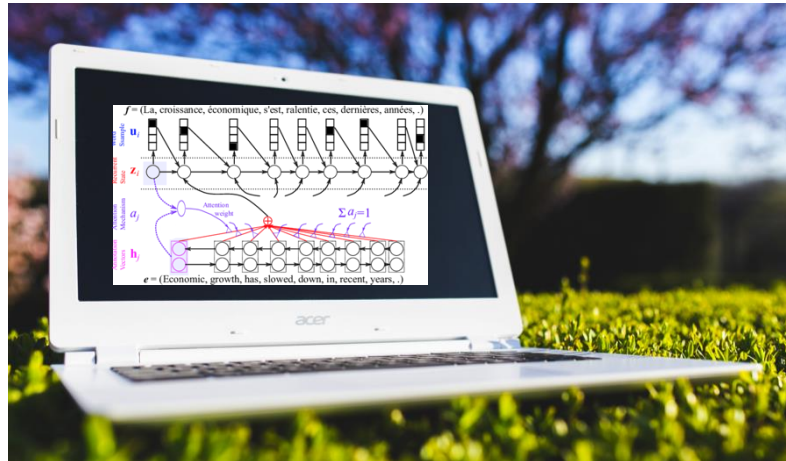
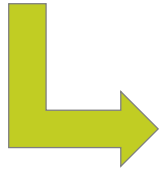
i came for your uncle ' s wedding .

yeah , the doctors said i would remember more every day .

In a new domain

“tourism accounts for almost N % of the austrian gross domestic product .”

“the industry are nearly N , of the most common population .”



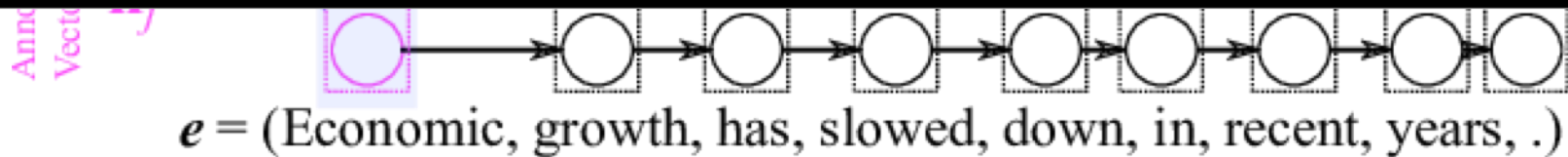
On Wikipedia:
BLEU = 7.4

Tuning NMT?

$f = (\text{La, croissance, économique, s'est, ralentie, ces, dernières, années, .})$



Black Box NMT



Transfer Learning

- Our core strategy is to employ **transfer learning** between deep neural networks pre-trained on massive datasets
- Knowledge gained in one context can be re-used to solve different but related problems



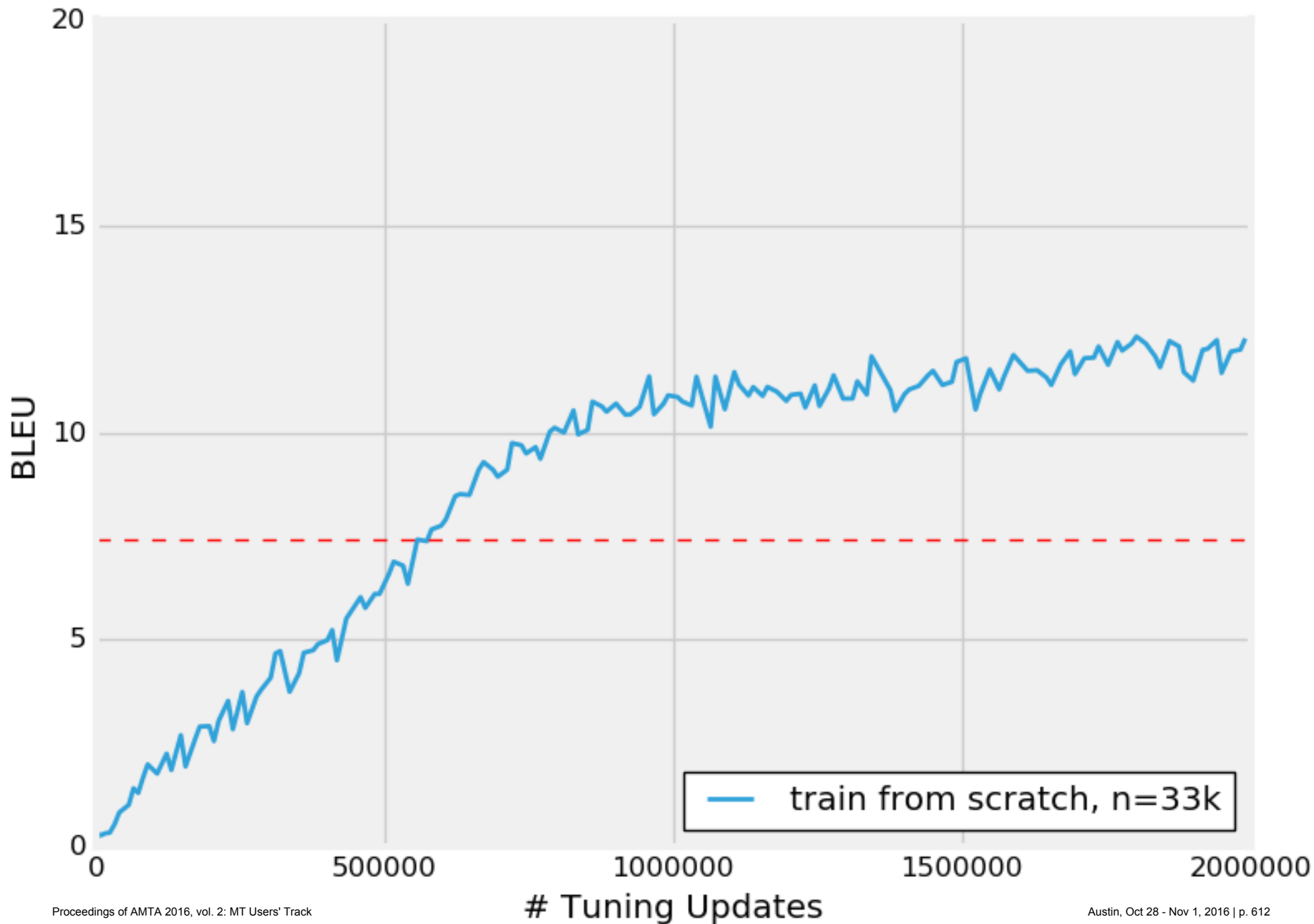
<https://edpsychexperience.wordpress.com/2013/03/25/013112-learning-learning-transfer>

Wikipedia Adaptation Experiments

- **Incremental training: we pick up where OpenSubtitles left off**
 - With tiny parallel tuning set (n=1024)
 - With small parallel training set (n=32768)
 - With full parallel training set (n=148136)
 - With varying amounts of in-domain monolingual data
 - With expanded vocabularies
- **About 22 minutes per 100k training updates**

Krzysztof Wołk and Krzysztof Marasek: Building Subject-aligned Comparable Corpora and Mining it for Truly Parallel Sentence Pairs., Procedia Technology, 18, Elsevier, p.126-132, 2014

Incrementally Adapting OpenSubtitles to Wikipedia



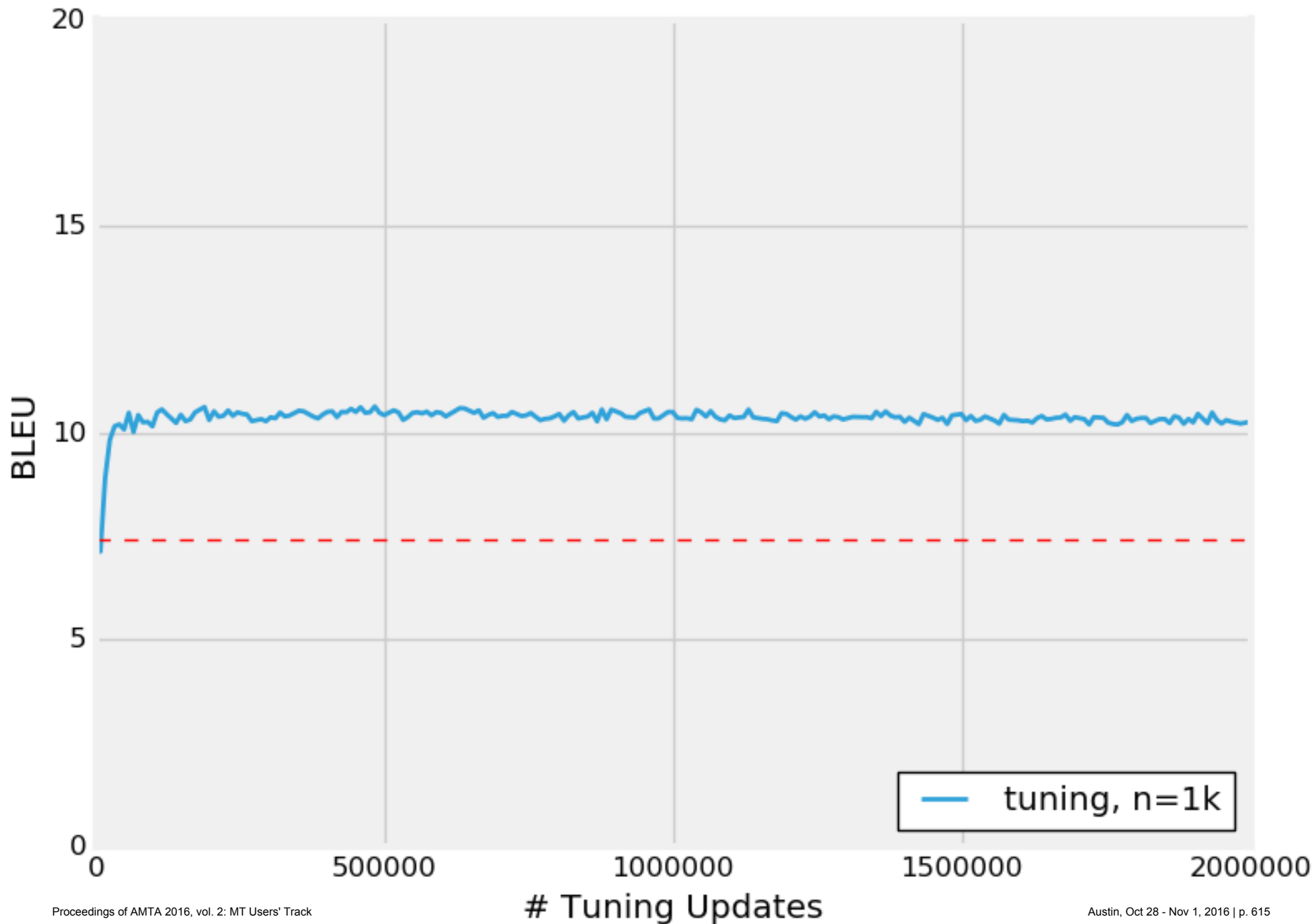
Incrementally Adapting OpenSubtitles to Wikipedia



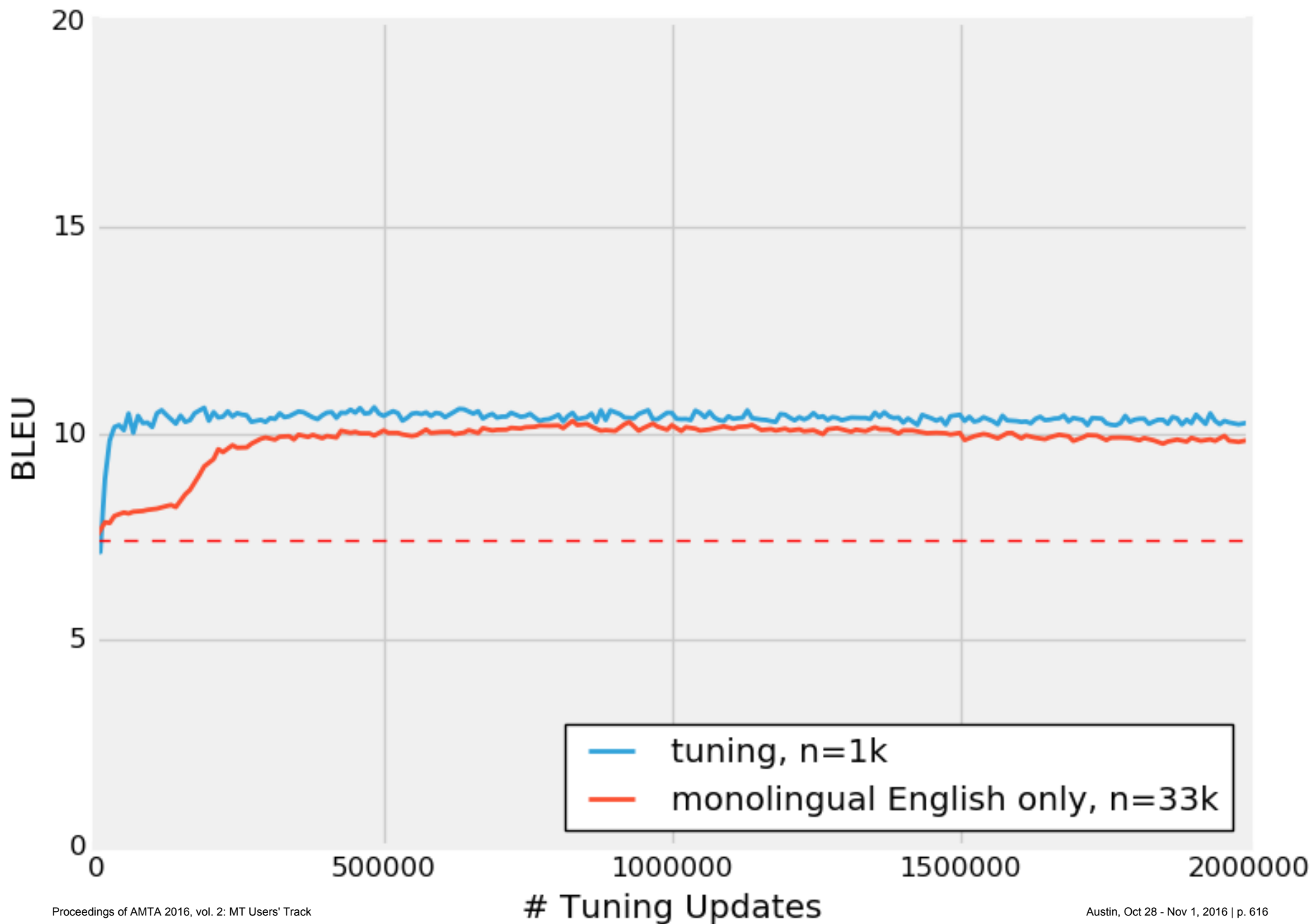
Incrementally Adapting OpenSubtitles to Wikipedia



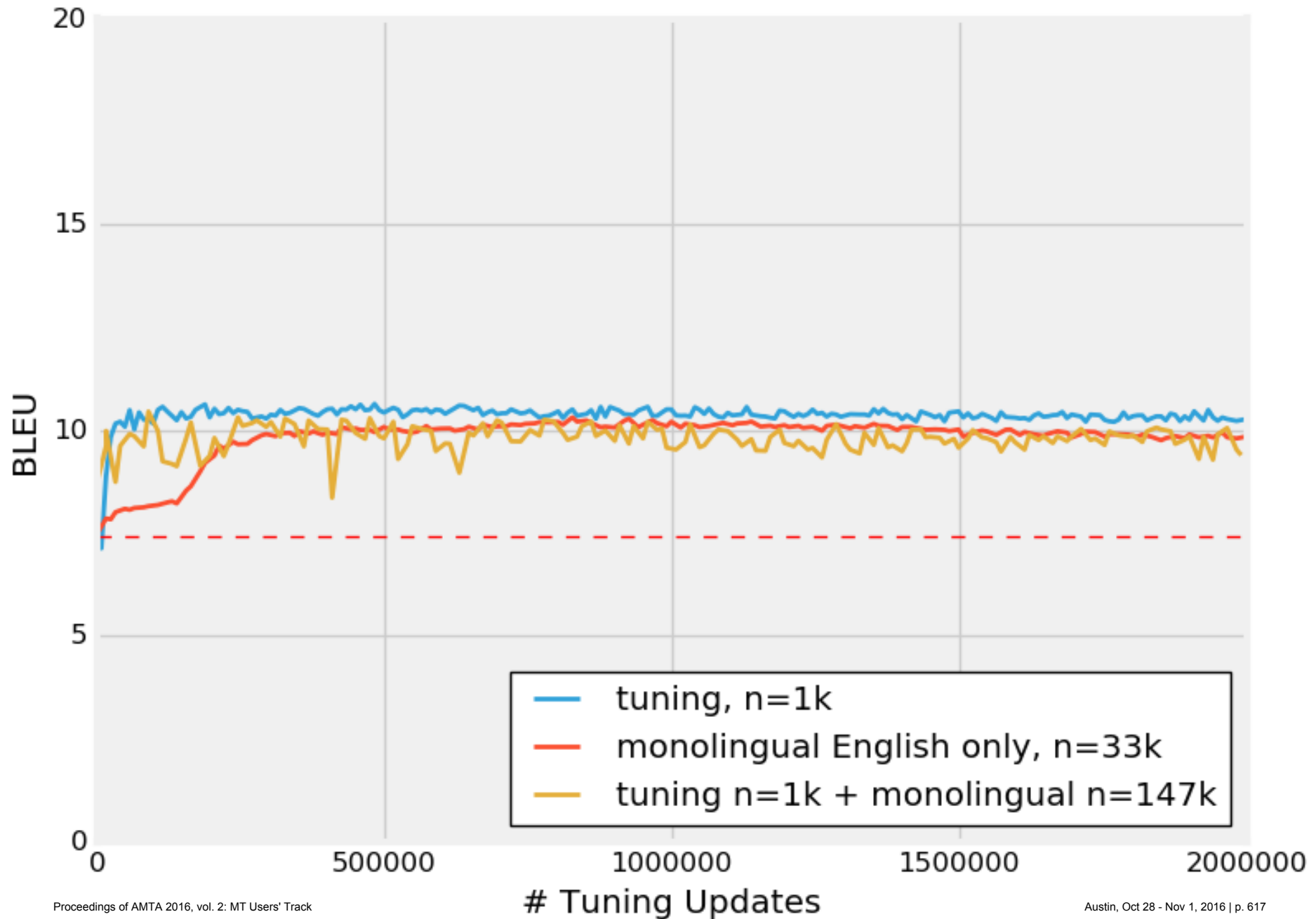
Incrementally Adapting OpenSubtitles to Wikipedia



Incrementally Adapting OpenSubtitles to Wikipedia



Incrementally Adapting OpenSubtitles to Wikipedia



Side by Side

Reference: tourism accounts for almost N % of the austrian gross domestic product .

Train from scratch, 33k: world is up for N % of the total reserves .

Untuned: the industry are nearly N , of the most common population .

1k tuning: tourism costs nearly N (of the most common population .

33k tuning: tourism often manifests approximately N % of the gdp .

... ensembling?

Results

- **Genre & domain matter (a lot)**
 - In-genre test: BLEU = **25.6**
 - Out-of-genre test: BLEU = **7.5 (-18.1)**
- **Incremental training helps**
 - Trained, parallel in-domain: BLEU = **13.5 (+6.0)**
 - Tuned, parallel in-domain: BLEU = **15.0 (+7.5)** to **16.9 (+9.4)**
- **Monolingual data helps when parallel data is scarce**
 - Tuned, 33k monolingual in-domain: BLEU = **10.3 (+2.8)**
 - Tuned, 1k parallel in-domain: BLEU = **10.6 (+3.1)**
- **Expanding vocabulary doesn't increase BLEU (yet)**

Conclusions

- **All parameters in a NMT system are tunable**
 - can create great diversity from one “well trained” seed system
 - ... in minutes or hours, with little or no additional parallel data
- **Government use cases poised to benefit most**
 - Collect many partially trained systems on the shelf?
- **Still open question how to best create systems optimized for tuning**
- **Sharing models? Share training code too.**



Thank You

Guido Zarrella
jzarrella@mitre.org
@gzco

MITRE



INVISIBLE MT

Patricia O'Neill-Brown, Ph.D
AMTA 2016

Overview

1. The Goal: Better Quality Machine Translation
2. 'Invisible MT'....a way to advance the field
3. Technical Approach
4. Questions/Discussion

INVISIBLE MT

MT processes today...

1. See paragraphs or documents you want to translate
2. Cut text
3. Open MT application
4. Paste into app
5. Wait for system to translate
6. See output & try to read/make sense of it
7. Decide what to do next
8. Maybe nothing else because your information need was met OR
9. Nothing else if your information need wasn't met OR
10. Post-edit - keep some output, discard some, revise still yet others

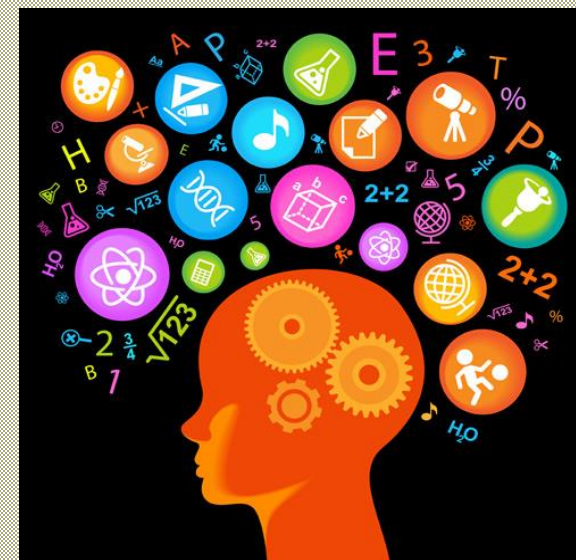
Current MT Paradigm

- Assumptions about **inputs to the system**:
 - Sentences (grammatical)
- The **output**:
 - Each system produces only one translation;
 - There is only ever one right answer
- Task **accomplishment** mode:
 - Complete Automation
 - Send document, get result

However...

What if... we changed the process & the paradigm?

- What if we designed for the input being:
 - A few short sentences
 - Fragments
 - Single words
 - Images, emoticons
- What if for the **output**, we said:
 - The system's goal is to convey meaning;
 - Translation is one technique & you can use others;
 - There can be multiple 'right' ways of saying things.



- How about for the task **accomplishment** mode, it were designed as:
 - Human-in-the-MT-loop

Our Concept

Invisible MT



Invisible MT...

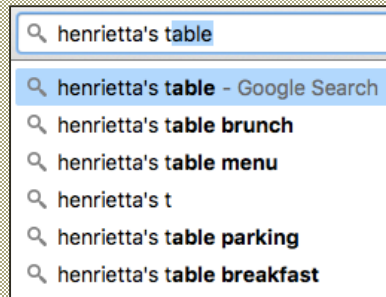
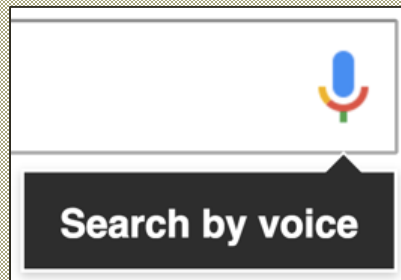
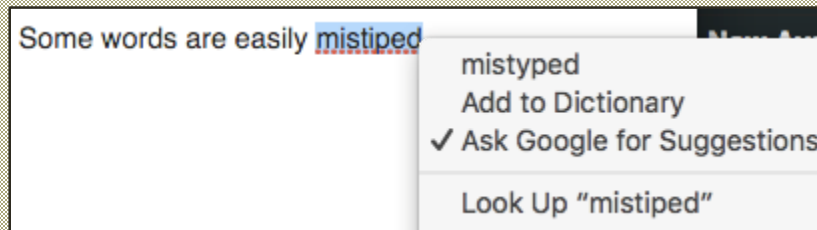
The design is driven by the *form factor* of the hardware used & the way people are used to interacting with it already



Invisible MT...

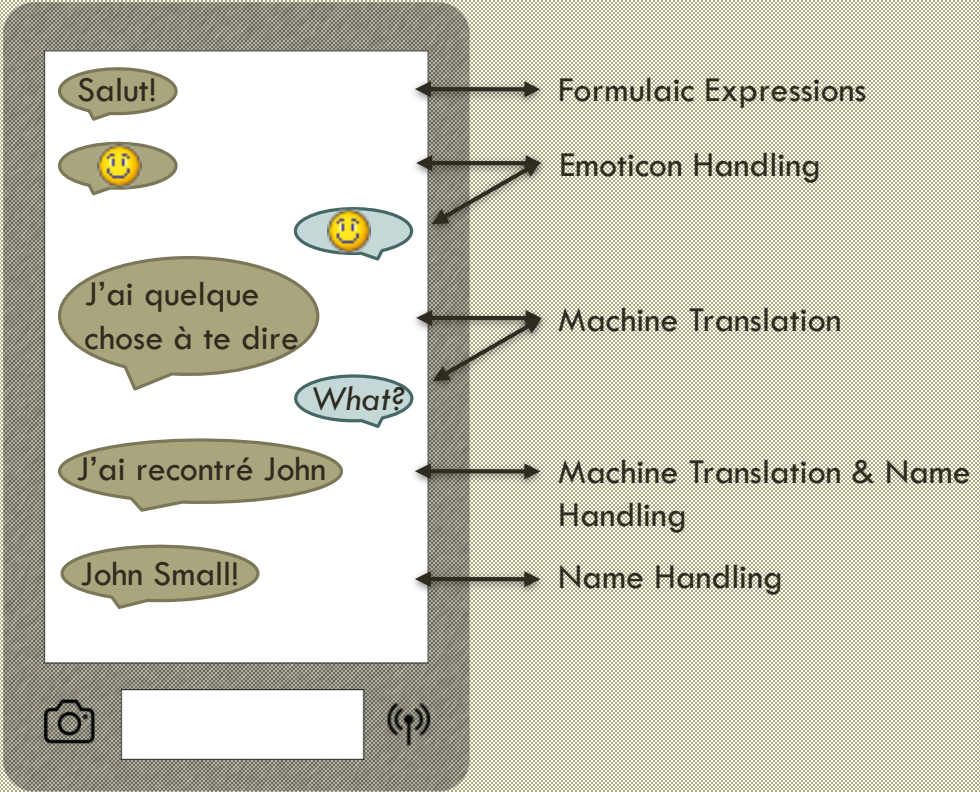
The design is also driven by the *characteristics* of the data for the different use cases

'SEARCH' TECHNOLOGY IS *INVISIBLE* TO THE USER



- Autocomplete
- Predictive Typing
- Spelling Correction
- Voice Authentication
- Information Retrieval
- Interactive Voice Response
- Automatic Speech Recognition

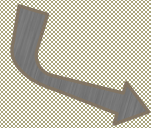




Non-Invasive

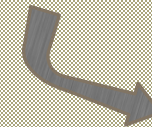


Natural



'autosuggest'

Easy



No having to
go to another
app

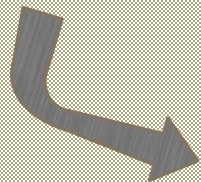
User
In The Loop

Adaptive



Interactive

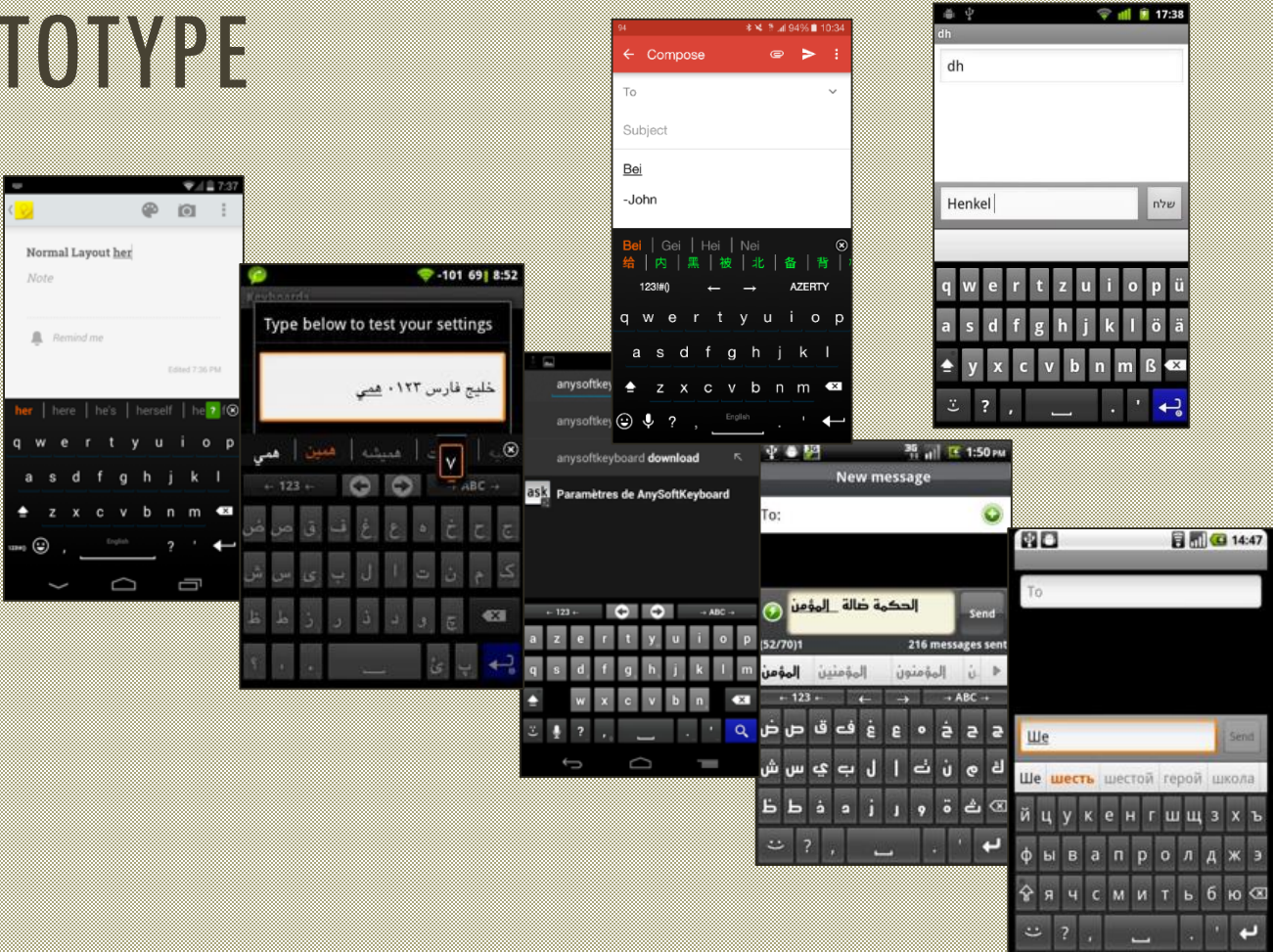
Invisible MT



More Accurate

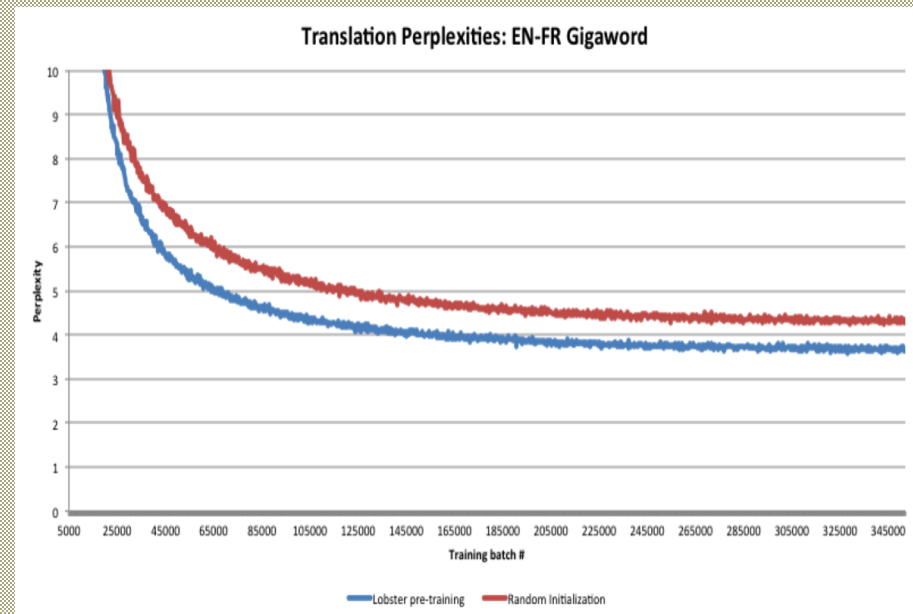
Work to date...

PROTOTYPE

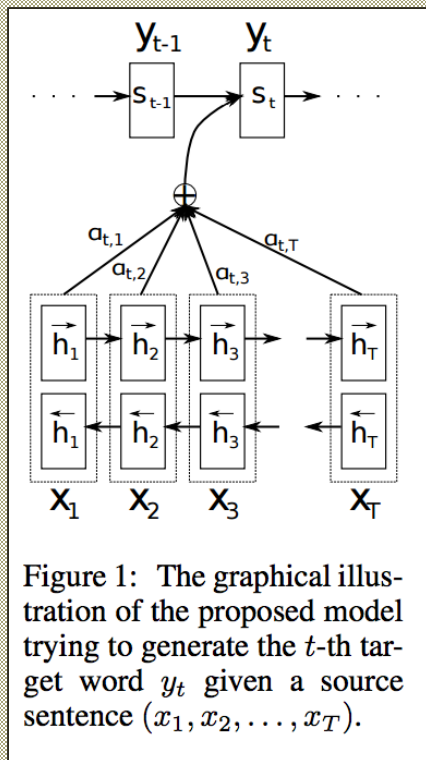


Deep Learning for MT

- Train a recurrent neural network to **understand** input and **generate** translation
- These models run on the mobile device
- Tunable to specific domains
 - Leverages monolingual text data
 - Can be retrained



DEEP LEARNING FOR MT



Bahdanau, D., Cho, K. and Bengio, Y., 2014. [Neural machine translation by jointly learning to align and translate](#). arXiv preprint arXiv:1409.0473.

We're currently following the standard sequence-to-sequence model for MT.

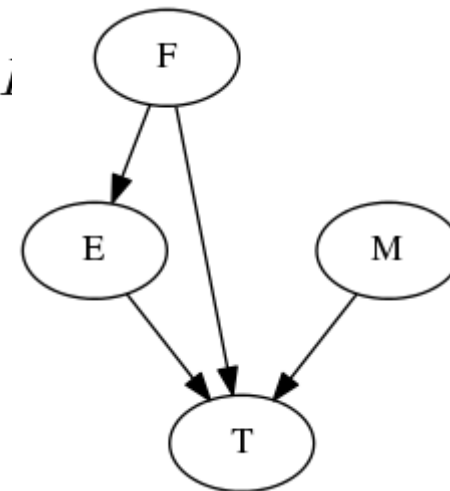
One part of model learns where to look in the source sentence when trying to produce the next word. *Attention*

Other part of the model decides what word to produce given where the first part is looking.

THE GRAPHICAL MODEL

$$\begin{aligned} P(F, E, T, M) &= \sum_{F, E, M} P(F)P(E|F)P(T|F, E, M)P(M) \\ &= P(M = 1) \sum_F P(T|F)P(F) \\ &+ P(M = 0) \sum_{E, F} P(T|E)P(F|E)P(F) \end{aligned}$$

$P(E)$ English Language Model
 $P(F)$ Foreign Language Model
 $P(F|E)$ Translation Model
 $P(T|E), P(T|F)$ Editing Models
 $P(M)$ Modality (1=typing Foreign, 0=typing English)



TRAINING & TESTING DATA

Europarl Data

Not the right fit – formal

Not the type of language used for
texting - informal

Online Movie Database –
OpenSubtitles






Good fit







Conversational











Pierre Lison and Jörg Tiedemann, 2016, *OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles*. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016).








language	files	tokens	sentences	af	ar	bg	bn	br	bs	ca	cs	da	de	el	en	eo	es	et	eu	fa	fi	fr	g
af	32	0.2M	27.4k		6.2k	7.6k			1.8k		10.5k	6.0k	7.9k	11.6k	16.2k		12.6k	2.2k		2.1k	2.8k	7.3k	
ar	67,608	329.8M	60.8M	6.2k		16.2M	62.2k	13.0k	6.1M	0.3M	16.5M	7.5M	7.1M	15.3M	19.4M	19.4k	18.3M	6.9M	0.1M	3.0M	10.8M	14.4M	44
bg	90,376	523.4M	80.2M	7.7k	17.8M		60.7k	13.8k	7.5M	0.3M	21.1M	8.2M	8.9M	19.3M	26.4M	23.4k	24.8M	7.7M	0.1M	2.8M	13.2M	18.5M	48
bn	76	0.6M	0.1M		64.1k	62.7k			36.6k	3.1k	61.1k	58.2k	54.7k	58.5k	69.3k		65.8k	56.5k	3.1k	44.8k	56.1k	59.3k	
br	32	0.2M	23.1k		13.3k	14.1k			2.7k	5.3k	14.5k	10.0k	7.5k	14.4k	17.7k	1.1k	15.6k	15.0k	0.7k	4.4k	8.1k	15.4k	0
bs	30,511	179.5M	28.4M	1.8k	12.2M	8.5M	37.7k	2.7k		0.1M	7.5M	3.7M	3.6M	7.3M	9.5M	7.4k	9.0M	3.5M	76.3k	1.3M	5.2M	6.8M	27
ca	711	4.0M	0.5M		0.3M	0.3M	3.2k	5.5k	0.1M		0.3M	0.2M	0.2M	0.3M	0.4M		0.4M	0.2M		96.2k	0.2M	0.3M	11
cs	125,126	715.3M	112.8M	10.7k	18.1M	24.7M	63.3k	14.8k	8.5M	0.4M		8.5M	9.3M	19.8M	27.5M	31.7k	25.9M	7.9M	0.1M	2.9M	13.7M	19.1M	68
da	24,079	162.4M	23.6M	6.1k	8.0M	9.3M	60.9k	10.1k	4.0M	0.2M	9.6M		4.9M	8.1M	9.4M	11.3k	9.1M	5.0M	87.7k	2.1M	7.9M	7.6M	28
de	27,742	186.3M	26.9M	8.0k	7.6M	10.0M	57.2k	7.7k	4.0M	0.2M	10.6M	5.4M		9.1M	11.5M	24.9k	10.8M	4.3M	75.7k	1.8M	6.9M	9.2M	52
el	114,230	683.1M	101.6M	11.8k	16.8M	22.3M	60.5k	14.6k	8.1M	0.3M	23.0M	9.1M	10.2M		25.6M	24.5k	24.5M	7.5M	0.1M	2.8M	13.1M	19.6M	66
en	322,294	2.5G	336.6M	16.7k	21.9M	31.6M	75.0k	18.5k	11.1M	0.4M	33.8M	11.0M	13.4M	30.4M		49.0k	40.0M	8.6M	0.2M	3.3M	16.8M	28.0M	0.1
eo	89	0.5M	79.3k		19.9k	24.3k		1.1k	7.6k		32.8k	11.7k	25.6k	25.2k	51.1k		38.6k	17.6k		5.1k	18.9k	28.3k	0
es	191,987	1.3G	179.2M	12.9k	20.3M	29.2M	69.1k	16.0k	10.2M	0.4M	30.7M	10.4M	12.4M	28.6M	50.1M	40.2k		8.3M	0.2M	3.1M	15.7M	25.8M	0.2
et	23,515	140.7M	22.9M	2.2k	7.5M	8.9M	58.6k	15.4k	4.0M	0.2M	9.2M	5.7M	4.8M	8.6M	10.3M	18.2k	9.6M		93.3k	1.9M	6.5M	6.9M	29
eu	188	1.4M	0.2M		0.1M	0.1M	3.3k	0.7k	80.9k		0.1M	93.2k	80.1k	0.2M	0.2M		0.2M	0.1M		43.1k	0.1M	0.1M	10
fa	6,469	44.3M	7.4M	2.1k	3.1M	2.9M	46.3k	4.4k	1.4M	0.1M	3.1M	2.2M	1.9M	3.0M	3.6M	5.2k	3.3M	2.1M	44.7k		2.4M	2.5M	21
fi	44,594	208.5M	38.7M	2.8k	11.5M	14.8M	57.9k	8.3k	5.7M	0.2M	15.3M	9.0M	7.6M	14.6M	19.2M	19.5k	17.7M	7.4M	0.1M	2.5M		12.5M	40
fr	105,070	672.8M	90.9M	7.5k	15.5M	21.3M	61.4k	16.3k	7.4M	0.3M	21.8M	8.5M	10.3M	22.2M	33.5M	29.1k	30.1M	7.8M	0.1M	2.7M	13.9M		93
el	370	1.9M	0.2M		15.8k	19.7k		0.5k	28.0k	11.0k	71.3k	29.4k	54.5k	68.8k	0.2M	0.3k	Autism	0.28	Nov 14, 2016	16:49	06:11		

EXAMPLE TRANSLATION MODEL ENTRIES

almost		
presque		39%
près		19%
presque		9%
quasi		5%
quasiment		4%

some		
certain		28%
certaines		11%
une		7%
quelques		6%
des		6%
un		5%

border		
frontière		24%
frontières		20%
des		10%
aux		10%
frontalières		6%
les		5%
frontaliers		5%
frontalière		3%
la		3%
frontalier		3%

youth		
jeunesse		25%
jeunes		22%
la		19%
des		13%
les		6%
pour		3%
de		2%

DEEP LEARNING FOR MT

Arabic-English BLEU

- 15.1 Baseline
- 18.7 +UNK, some post-processing
- 22.4 +gradient clipping, longer training
- 24.5 +pretrained word embeddings
- ...

System is now near state-of-research performance

- Time to switch to E-A for MCT

DEEP LEARNING FOR MODEL COMPRESSION



The amount of storage on phones is smaller compared to that on servers

- Neural methods can be used to make small models as effective as traditional, larger models

T-TABLE COMPRESSION

Input: English word

```
"Feast":
```

Output: Probability it translates
to each of 43K French words
(most ~0%)

```
{  
  "festin": 0.42,  "fête":  
  0.57  
}
```

Challenge: Model table with
fewer parameters than needed
to store (table entries)

Score: Cross Entropy (CE)

- * Lower means less precision lost during compression.

T-TABLE COMPRESSION, *WORK IN PROGRESS*

Source side: English word

Target side: French word

Numeric entry: $p(f|e)$

Compression of source side

5.3 CE Recurrent Neural Net (RNN)

0.75 CE Convolutional Neural Net (CNN)

Target side compression (all infeasible for use in an actual system)

0.73 CE Word Index

5.4 CE RNN

5.2 CE CNN

T-TABLE COMPRESSION, BOTTOM LINE

Can we reduce data storage requirements while still predicting MCT suggestions with enough accuracy to be useful?

- Yes

Best approach we've found, so *far*:

- Character-based input encoder, one character per keystroke typed
- Hierarchical softmax output decoder produces the distribution of numbers over the output vocabulary

Next Steps

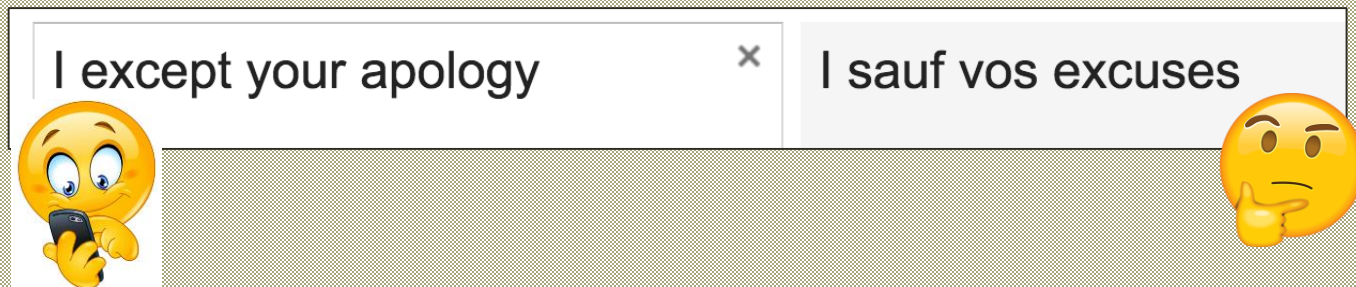
1. Obtain user Feedback
2. Determine if data used designing the system is adequate
4. Look at if there are areas to optimize

Questions?

DEEP LEARNING FOR ERROR CHECKING AND HINTING

Rapid typing encourages certain types of mistakes

- * Some, e.g. **typos**, can be corrected immediately
- * Others require context: **determiners, inflections, homonyms**



Magic Punctuation: when the author types a period, can we identify if part of the sentence doesn't match the user's intent?

DEEP LEARNING FOR ERROR CHECKING: MANDARIN

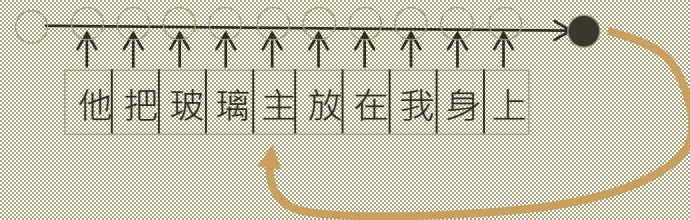
Goal: identify when user selects the **wrong character** from a **list of phonetically similar options** with same pinyin transcription

Our noise model samples from a pinyin / character frequency table to corrupt clean Mandarin sentences

- * Allows us to cheaply build large training data from monolingual text

Our denoiser is a **recurrent neural network** that reads the sequence of characters, then “points” at a position in the input

- * Very compact model (3 MB)
- * Corrupt characters are detected with over 85% accuracy in initial tests



You got your StatMT
on my rules!

You got your rules
in my StatMT!

MoJo

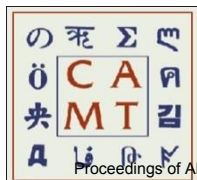
Bringing Hybrid MT to the
Center for Applied Machine Translation

Marianna J. Martindale,
Center for Applied Machine Translation (CAMT)



MT in Research & Industry vs CAMT

- Research is almost entirely StatMT (now Neural)
 - Emphasis on new techniques
 - Most research on high-resource language pairs (except LORELEI & MATERIAL)
 - Not concerned with operational constraints
- In industry StatMT is the norm (for now)
 - Primarily commercially viable language pairs (high-resource)
 - Speed is important, compute resources may or may not be
- CAMT's GOTS MT is (currently) rule-based
 - Many languages **regardless of resource availability**
 - Speed is important, compute resources limited (server OR laptop)
 - Fidelity is more important than fluency



Why not StatMT before?

- Technical issues with StatMT
 - Speed
 - Memory
 - Well-engineered systems not readily available
 - Can be tricky to build right

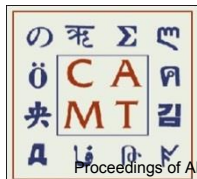


Why not StatMT before?

- Technical issues with StatMT

- Speed
- Memory
- Well-engineered systems not readily available
- Can be tricky to build right

OPEN SOURCE



Why not StatMT before?

- Technical issues with StatMT

- Speed
- Memory
- Well-engineered systems not readily available
- Can be tricky to build right

OPEN SOURCE

- Domain needs

- Many languages (often low-resource)



Languages Supported in CyberTrans

AFRIKAANS	DANISH	JAPANESE	PERSIAN	TAJIK
ALBANIAN	DARI	JAVANESE	PERSIAN Romanized	TAJIK Romanized
AMHARIC	DUTCH	KAZAKH	POLISH	TAUSUG
ARABIC	ESTONIAN	KOREAN	PORTUGUESE	TETUM
ARABIC Romanized	FINNISH	KURDISH (Kurmanji)	PUNJABI	THAI
ARMENIAN	FRENCH	KURDISH (Sorani)	ROMANIAN	TOK PISIN
AZERBAIJANI	GALICIAN	KYRGYZ	RUSSIAN	TURKISH
BALUCHI	GEORGIAN	LAO	RUSSIAN Romanized	TURKMEN
BASQUE	GEORGIAN Romanized	LATVIAN	SERBIAN	TWI*
BELARUSIAN	GERMAN	LINGALA	SERBIAN Cyrillic	UKRAINIAN
BULGARIAN	GREEK	LITHUANIAN	<i>SHONA*</i>	UKRAINIAN Romanized
BULGARIAN Romanized	GREEK Romanized	MACEDONIAN	SLOVAK	URDU
CATALAN	HAITIAN CREOLE	MACEDONIAN Romanized	SLOVENE	URDU Romanized
CEBUANO	HAUSA	MAGUINDANAON	SOMALI	UYGHUR
CHAVACANO	HEBREW	MALAGASY	SPANISH	UZBEK Cyrillic
CHECHEN	HINDI	MALAYSIAN	SRANAN	UZBEK Romanized
CHINESE Simplified	HMONG	NORWEGIAN	SUNDANESE	VIETNAMESE
CHINESE Traditional	HUNGARIAN	PAPIAMENTO	SWAHILI	WOLOF
CROATIAN	INDONESIAN	PASHTO	SWEDISH	YAKAN
CZECH	ITALIAN	PASHTO Romanized	TAGALOG	YORUBA



Why not StatMT before?

- Technical issues with StatMT

- Speed
- Memory
- Well-engineered systems not readily available
- Can be tricky to build right

OPEN SOURCE

- Domain needs

- Many **languages** (often low-resource)
- Little or no ***in-domain*** parallel text
- Frequent sometimes urgent **updates**
- **Fidelity** as priority (accurate, traceable)



Why not StatMT before?

- Technical issues with StatMT

- Speed
- Memory
- Well-engineered systems not readily available
- Can be tricky to build right

OPEN SOURCE

- Domain needs

- Many **languages** (often low-resource)
- Little or no ***in-domain*** parallel text
- Frequent sometimes urgent **updates**
- **Fidelity** as priority (accurate, traceable)

MoTrans

- Human instead of bitext
- Updated based on actual text submitted
- Easy to trace input to output
- Caveat: Sacrifice fluency for fidelity



Features of Rule-based and StatMT

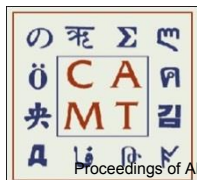
Rule-based

- Rules are composed by language experts
- Performs a deep source language analysis
- Easy to update, adapt to new domains
- Easy to trace input to output
- Very fast

StatMT

- Learns automatically from example translations
- Doesn't require language-specific knowledge
- Leverages Big Data
- More fluent translations
- Recent engineering advances make adoption easier

Best of both worlds?



Best of both worlds

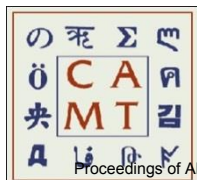
MoTrans

Human constructed
 Domain focused
 Knowledge-rich
 CAMT linguistic and
 technical investment

Statistical

Learned automatically
 Generic
 Language-agnostic
 Commercially dominant
 Open source

Hybrid



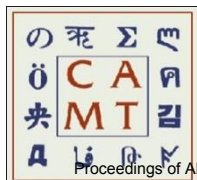
Example (Russian)

System	Output
Motrans	He noted, that presidential pre-election campaign provoked “discrepant and often frequently vulgar rhetoric,” eating away democracy and society.
StatMT	He noted that the presidential electoral campaign has provoked “inconsistent and often vulgar rhetoric,” разъедающую democracy and society .
Hybrid	He noted that the presidential electoral campaign has provoked “inconsistent and often vulgar rhetoric,” eating away democracy and society.
Human	He said the presidential campaign has brought “divisive and often vulgar rhetoric” that corrodes democracy and society.



Example (Russian)

System	Output
Motrans	From Moscow to Sochi on the train about two days! Really? Is it possible? You want to lead two days in the uncomfortable train?
StatMT	From Moscow to Sochi to train about two days! Do you want to spend two days in awkward train?
Hybrid	From Moscow to Sochi on the train about two days! Do you want to spend two days in uncomfortable train?
Human	From Moscow to Sochi by train is close to 2 days! Do you really want to spend two days in an uncomfortable train?



Example (Swahili)

System	Output
Motrans	LABLA America in/at what region? America is big.
StatMT	“Maybe America in what state? The United States is the greatest.
Hybrid	Maybe America in what region? The United States is big.
Human	To be more precise, which state in America? America is vast.

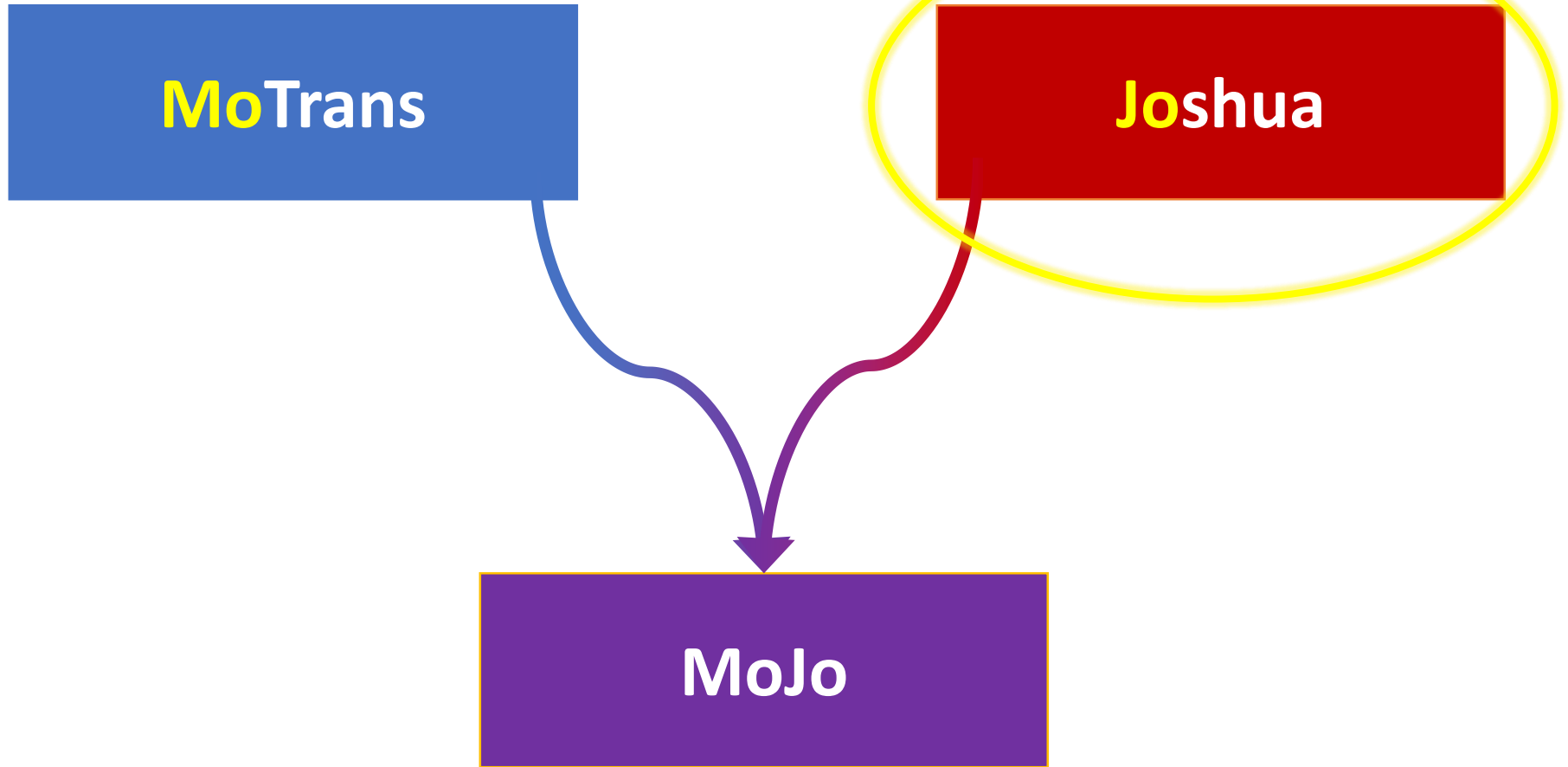


Example (Swahili)

System	Output
Motrans	Pat: say (you/it) succeeded to get children and MUMEO Mohammed HAMIS?
StatMT	Pat: is ulifanikiwa to children and mumeo Mohamed Hamis?
Hybrid	Pat: <i>Have you</i> succeeded in <i>getting</i> children mumeo Mohamed Hamis?
Human	Pat: Were you successful at having children by that husband Mohamed Hamis?



Best of both worlds

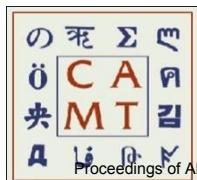


Apache Joshua



- Open-source Java statistical machine translation system
- Apache project currently in Incubation Stage
- Provides both phrase-based and hiero StatMT
- Pre-built language packs available for download
- **Written in Java** (like CyberTrans)
- **Development lead conveniently located at JHU HLT COE in nearby Baltimore**

<http://joshua.apache.org/>



Best of both worlds

MoTrans

Joshua

MoJo



MoTrans Translator

Settings	Lexical Entries	Grammar
- , ; / . ' (@ \$ \u20AC % + < = 0 1 2 3 4 5 6 7 8 9 A Á À Ã Ä Å		
+ - × ÷ ↵ ↶ ↷		
Source *	POS	Target
gabarit^ de fraisage	N	milling jig^
gabarit^ de membrure	N	frame mould^
gabarit^ de mécanicien	N	engineer's jig^
gabarit^ de montage	N	assembly jig^
gabarit^ de perçage	N	drill template^
gabarit^ de traçage	N	contour template^
gabarit^ de vérification	N	inspection gauge^
gabarit-obstacle	N	minimum dimensions
gabarre	N	lighter
gabbro	N	gabbro
gabegie	N	intrigue
gabelage	N	time^ during which the salt was in store t
gabelleur	N	customs official^
gabelier	N	salt-tax officer^
gabelle	N	salt tax^
gaber	V	joke



- Morphological Translator
- Fast
- Deep morphological analysis
- Expressive lexicon and grammar
- Continually updated by lexicographers
- Quick “better than nothing” for Low Resource Languages
- Currently over 40 languages
- Many users, positive feedback

MoTrans Lexicon Example

- Lexical entries

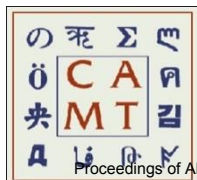
- puntilla |N| lace
- pas/ar |V.AR| pass
- de |DE| of
- pas de puntillas |V.AR| sidestep |S--

- Rules

PROCLITIC:V.CONJ:ha:has %s

SUFFIX:V.AR:ado:%s:+pastp:V.CONJ

SUFFIX_PATTERN:N:(*V)(X-0{s}):%s:+plural



MoTrans Lexicon Example

- Lexical entries

Source

- **puntilla** |N| lace
- **pas/ar** |V.AR| pass
- **de** |DE| of
- **pas de puntillas** |V.AR| sidestep |S--

Usually lemma

But not always

- Rules

PROCLITIC:V.CONJ:ha:has %s

SUFFIX:V.AR:ado:%s:+pastp:V.CONJ

SUFFIX_PATTERN:N:(*V)(X-0{s}):%s:+plural



MoTrans Lexicon Example

- Lexical entries

- puntilla | **N** | lace
- pas/ar | **V.AR** | pass
- de | **DE** | of
- pas de puntillas | **V.AR** | sidestep | S--

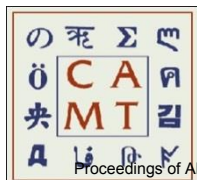
Part of Speech

- Rules

PROCLITIC:V.CONJ:ha:has %s

SUFFIX:V.AR:ado:%s:+pastp:V.CONJ

SUFFIX_PATTERN:N:(*V)(X-0{s}):%s:+plural



MoTrans Lexicon Example

- Lexical entries

- puntilla|N| lace
- pas/ar |V.AR| pass
- de |DE| of
- pas de puntillas|V.AR| sidestep|S--

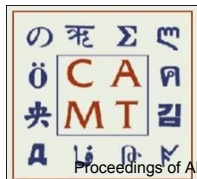
Target

- Rules

PROCLITIC:V.CONJ:ha:has %s

SUFFIX:V.AR:ado:%s:+pastp:V.CONJ

SUFFIX_PATTERN:N:(*V)(X-0{s}):%s:+plural



MoTrans Lexicon Example

- Lexical entries

- puntilla | N | lace
- pas/ar | V.AR | pass
- de | DE | of
- pas de puntillas | V.AR | sidestep | S--

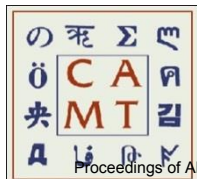
Stem indicator

- Rules

PROCLITIC:V.CONJ:ha:has %s

SUFFIX:V.AR:ado:%s:+pastp:V.CONJ

SUFFIX_PATTERN:N:(*V)(X-0{s}):%s:+plural



MoTrans Lexicon Example

- Lexical entries

- puntilla |N| lace
- pas/ar |V.AR| pass
- de |DE| of
- pas de puntillas |V.AR| sidestep |S--

Affix Location Pattern

- Rules

PROCLITIC:V.CONJ:ha:has %s

SUFFIX:V.AR:ado:%s:+pastp:V.CONJ

SUFFIX_PATTERN:N:(*V)(X-0{s}):%s:+plural



MoTrans Lexicon Example

- Lexical entries

- puntilla|N| lace
- pas/ar |V.AR| pass
- de |DE| of
- pas de puntillas|V.AR|sidestep|S--

- Rules

PROCLITIC:V.CONJ:ha:has %s

SUFFIX:V.AR:ado:%s:+pastp:V.CONJ

SUFFIX_PATTERN:N:(*V)(X-θ{s}):%s:+plural

Type



MoTrans Lexicon Example

- Lexical entries

- puntilla|N| lace
- pas/ar |V.AR| pass
- de |DE| of
- pas de puntillas|V.AR|sidestep|S--

- Rules

PROCLITIC:**V.CONJ**:ha:has %s

SUFFIX:**V.AR**:ado:%s:+pastp:V.CONJ

SUFFIX_PATTERN:**N**:(*V)(X-θ{s}):%s:+plural

Part-of-Speech



MoTrans Lexicon Example

- Lexical entries

- puntilla |N| lace
- pas/ar |V.AR| pass
- de |DE| of
- pas de puntillas |V.AR| sidestep |S--

- Rules

PROCLITIC:V.CONJ: **ha**:has %s

SUFFIX:V.AR: **ado**:%s:+pastp:V.CONJ

SUFFIX_PATTERN:N: **(*V)(X-θ{s})**:%s:+plural

Source
Transformation



MoTrans Lexicon Example

- Lexical entries

- puntilla |N| lace
- pas/ar |V.AR| pass
- de |DE| of
- pas de puntillas |V.AR| sidestep |S--

- Rules

PROCLITIC:V.CONJ:ha:has %s
 SUFFIX:V.AR:ado:%s:+pastp:V.CONJ
 SUFFIX_PATTERN:N:(*V)(X-θ{s}):%s:+plural

Target
Transformation



MoTrans Lexicon Example

- Lexical entries

- puntilla |N| lace
- pas/ar |V.AR| pass
- de |DE| of
- pas de puntillas |V.AR| sidestep |S--

- Rules

PROCLITIC:V.CONJ:ha:has %s

SUFFIX:V.AR:ado:%s:**+pastp**:V.CONJ

SUFFIX_PATTERN:N:(*V)(X-θ{s}):%s:**+plural**

Target
Conjugation



MoTrans Lexicon Example

- Lexical entries

- puntilla |N| lace
- pas/ar |V.AR| pass
- de |DE| of
- pas de puntillas |V.AR| sidestep |S--

- Rules

PROCLITIC:V.CONJ:ha:has %s

SUFFIX:V.AR:ado:%s:+pastp:**V.CONJ**

SUFFIX_PATTERN:N:(*V)(X-θ{s}):%s:+plural

New
Part-of-Speech



MoTrans Lexicon Example

- Lexical entries

- **puntilla** |N| lace
- pas/ar |V.AR| pass
- de |DE| of
- pas de puntillas |V.AR| sidestep |S--

- Rules

PROCLITIC:V.CONJ:ha:has %s

SUFFIX:V.AR:ado:%s:+pastp:V.CONJ

SUFFIX_PATTERN:N:(*V)(X-θ{s}):%s:**+plural**

For a noun
that ends
in a vowel

Add an 's'

Pluralize
the English

puntillas ->

lace



MoTrans Lexicon Example

- Lexical entries

- puntilla |N| lace
- **pas/ar** |V.AR| **pass**
- de |DE| of
- pas de puntillas |V.AR| sidestep |S--

- Rules

PROCLITIC:V.CONJ:ha:has %s

SUFFIX:V.AR:ado:%s:+pastp:V.CONJ

SUFFIX_PATTERN:N:(*V)(X-0{s}):%s:+plural

puntillas ->

lace

ha pasado ->

has passed



MoTrans Lexicon Example

- Lexical entries

- puntilla |N| lace
- pas/ar |V.AR| pass
- de |DE| of
- **pas de puntillas |V.AR| sidestep |S--**

- Rules

PROCLITIC:V.CONJ:ha:has %s

SUFFIX:V.AR:ado:%s:+pastp:V.CONJ

SUFFIX_PATTERN:N:(*V)(X-θ{s}):%s:+plural

puntillas ->

lace

ha pasado ->

has passed

ha pasado de puntillas ->

has sidestepped



MoTrans Lexicon Example

- Lexical entries

- puntilla |N| lace
- pas/ar |V.AR| pass
- de |DE| of
- pas de puntillas |V.AR| sidestep |S--**

- Rules

PROCLITIC:V.CONJ:ha:has %s

SUFFIX:V.AR:ado:%s:+pastp:V.CONJ

SUFFIX_PATTERN:N:(*V)(X-θ{s}):%s:+plural

puntillas ->

PUNTILLA

(SUFFIX_PATTERN:N:(*V)(X-θ{s}):%s:+plural):N: **lace**

ha pasado ->

PAS (PROCLITIC:V.CONJ:HA:has %s:+pastp)

(SUFFIX:V.AR:ADO:%s:+pastp):

V.CONJ: **has passed**

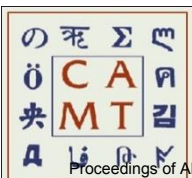
ha pasado de puntillas ->

PAS DE PUNTILLAS

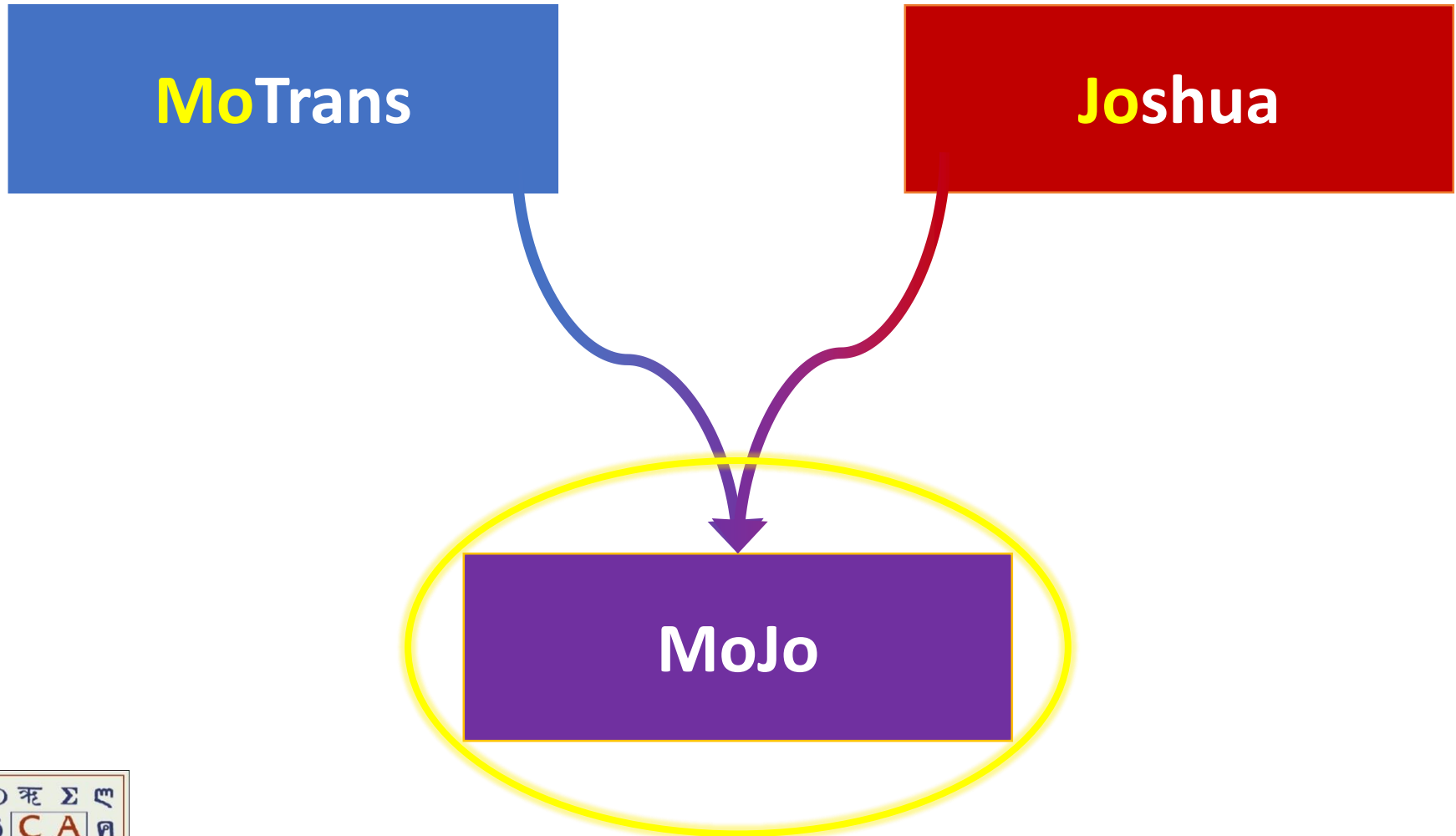
(PROCLITIC:V.CONJ:HA:has %s:+pastp)

(SUFFIX:V.AR:ADO:%s:+pastp):

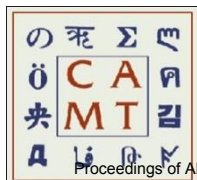
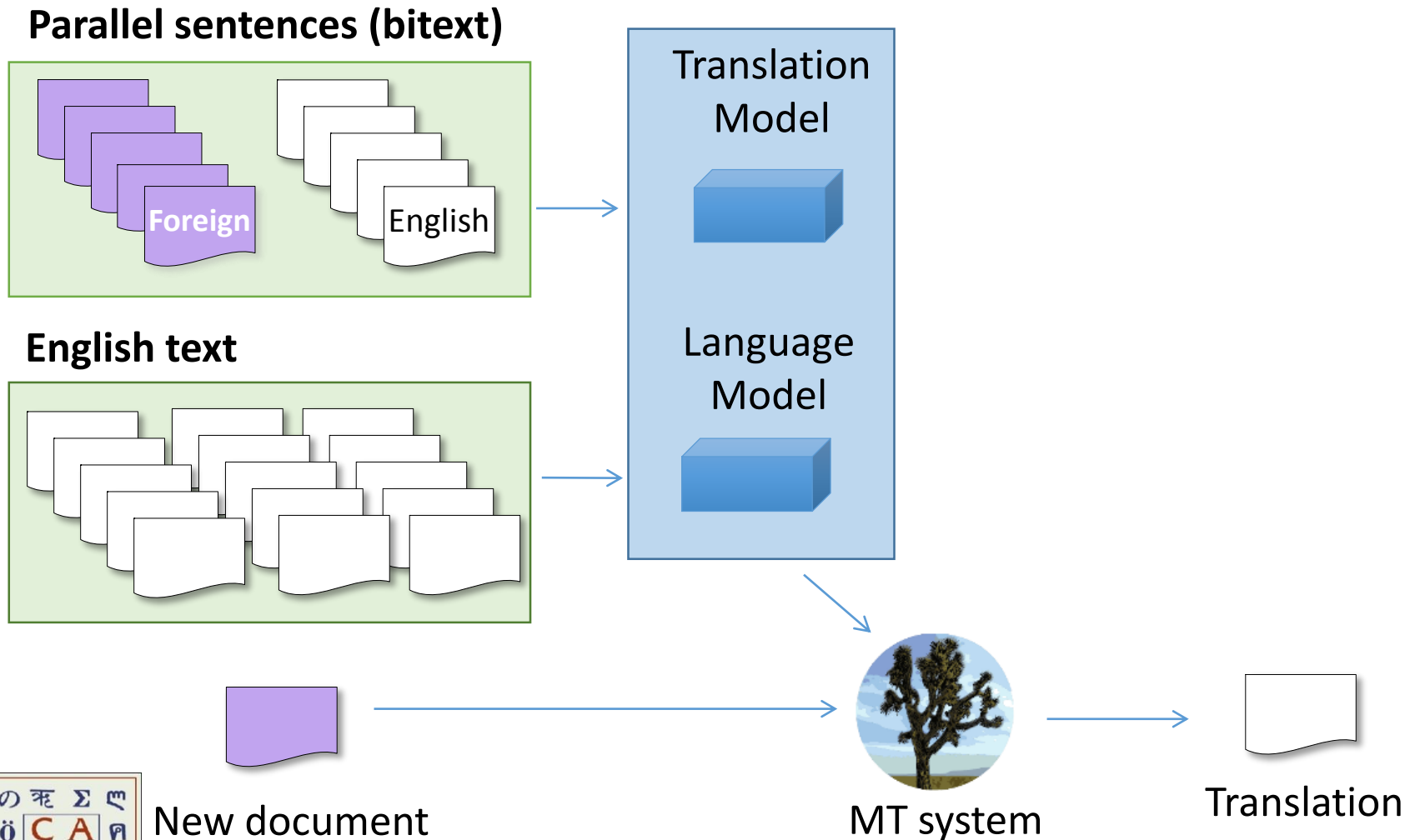
V.CONJ: **has sidestepped**



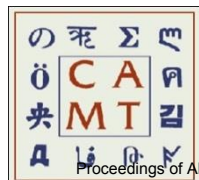
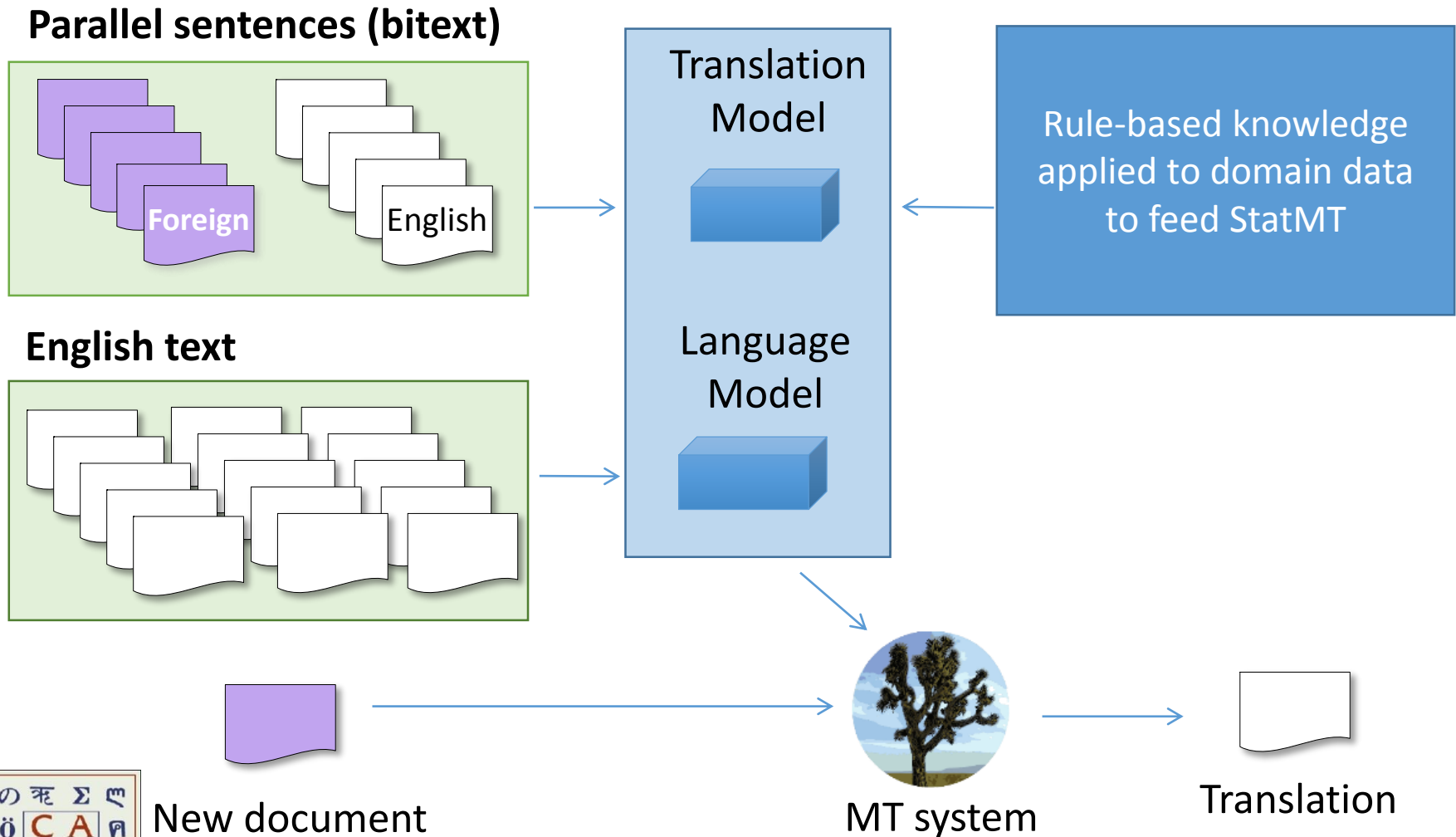
Best of both worlds



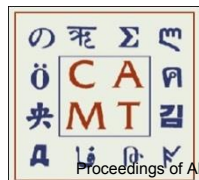
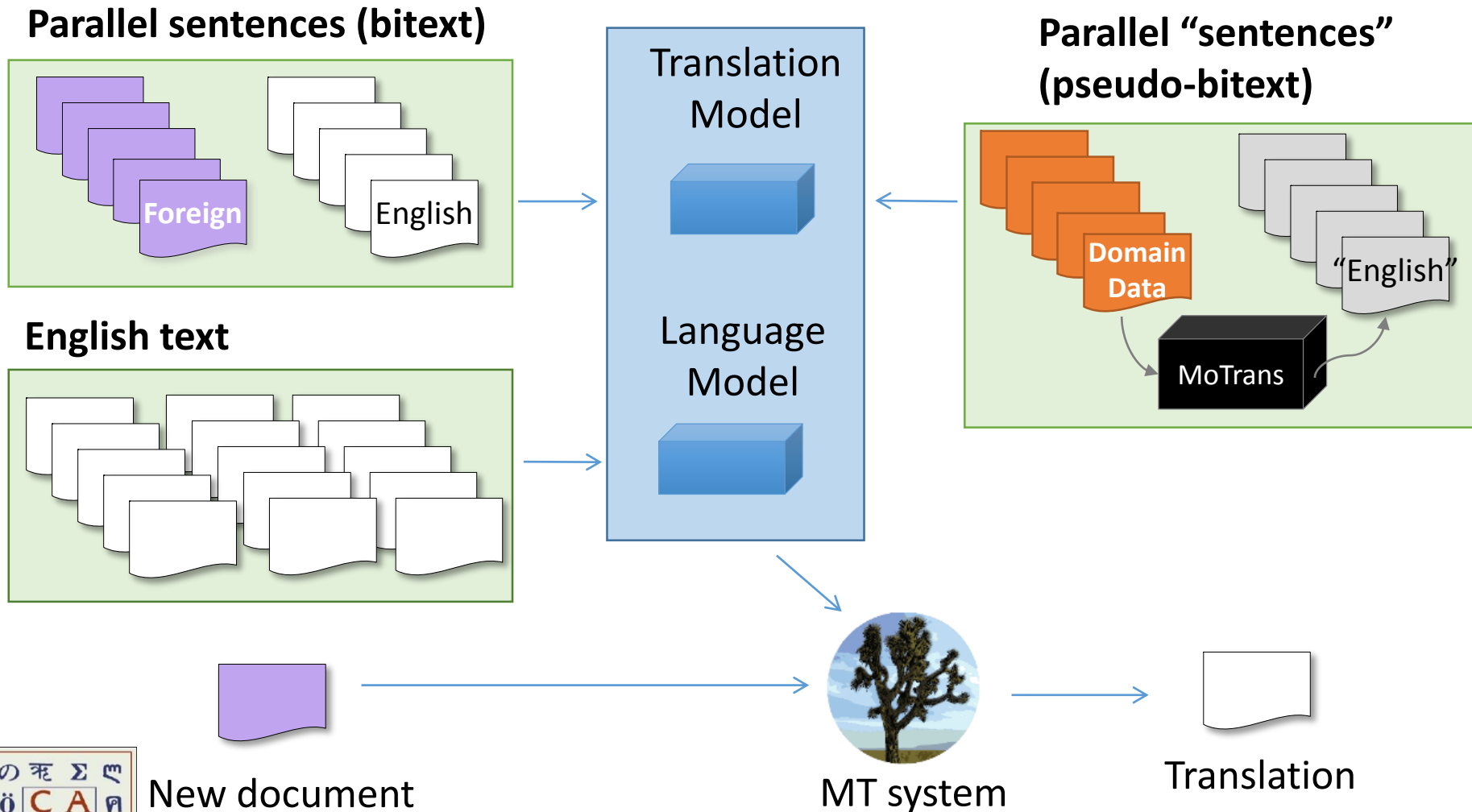
Building the Hybrid: Base StatMT



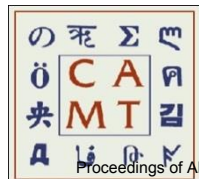
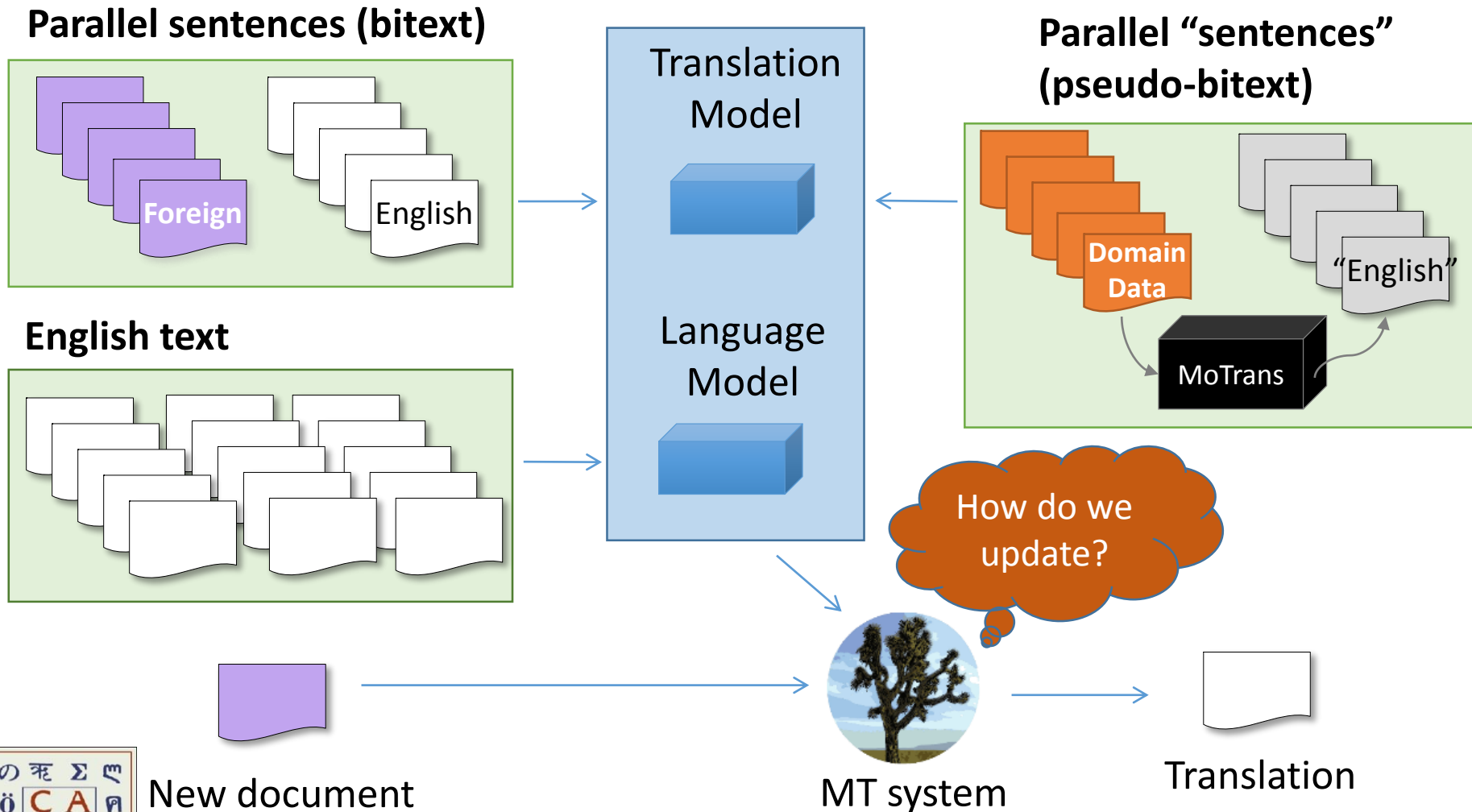
Building the Hybrid: Black Box



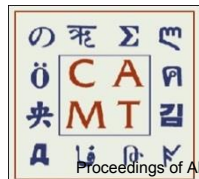
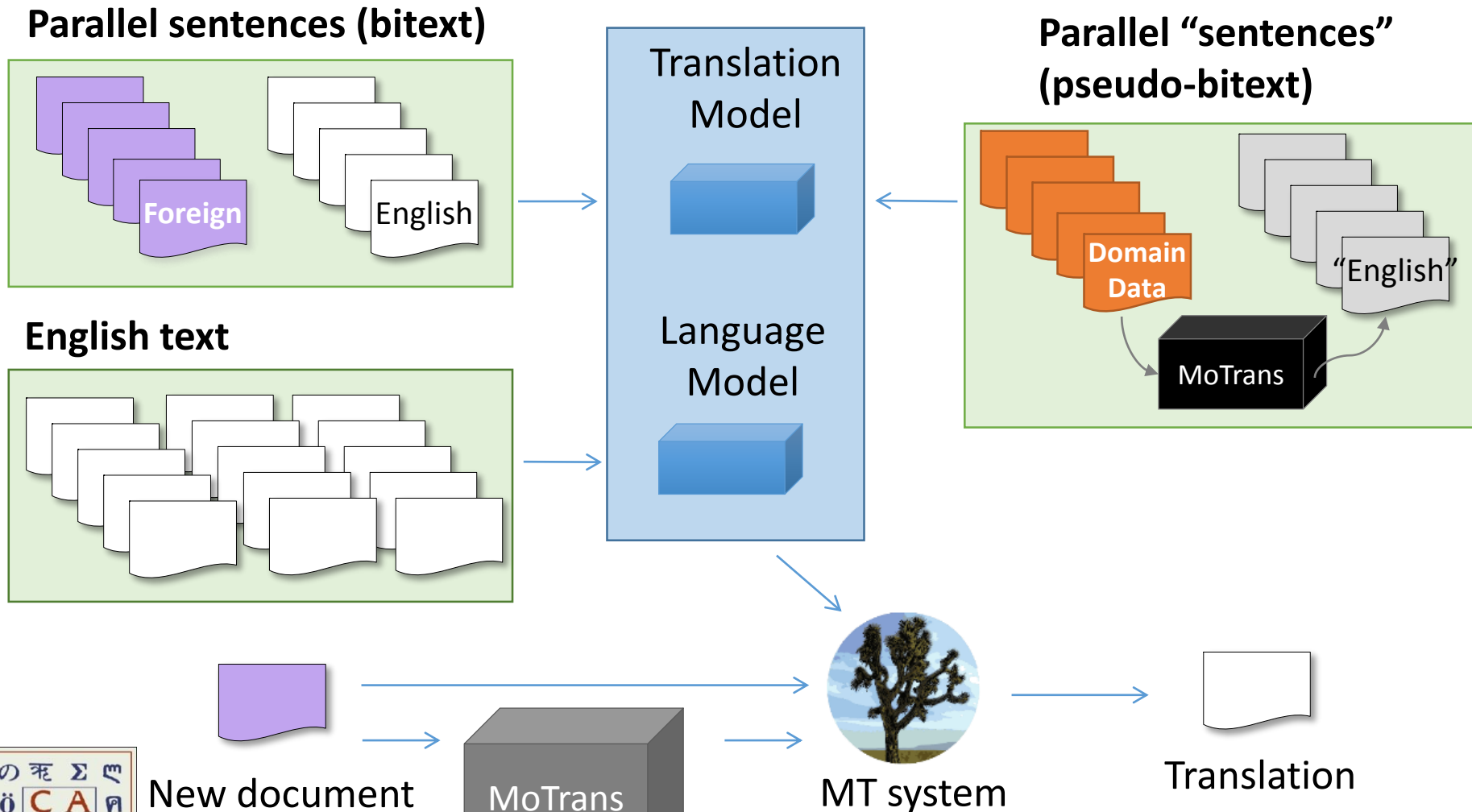
Building the Hybrid: Black Box



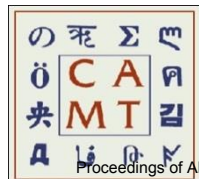
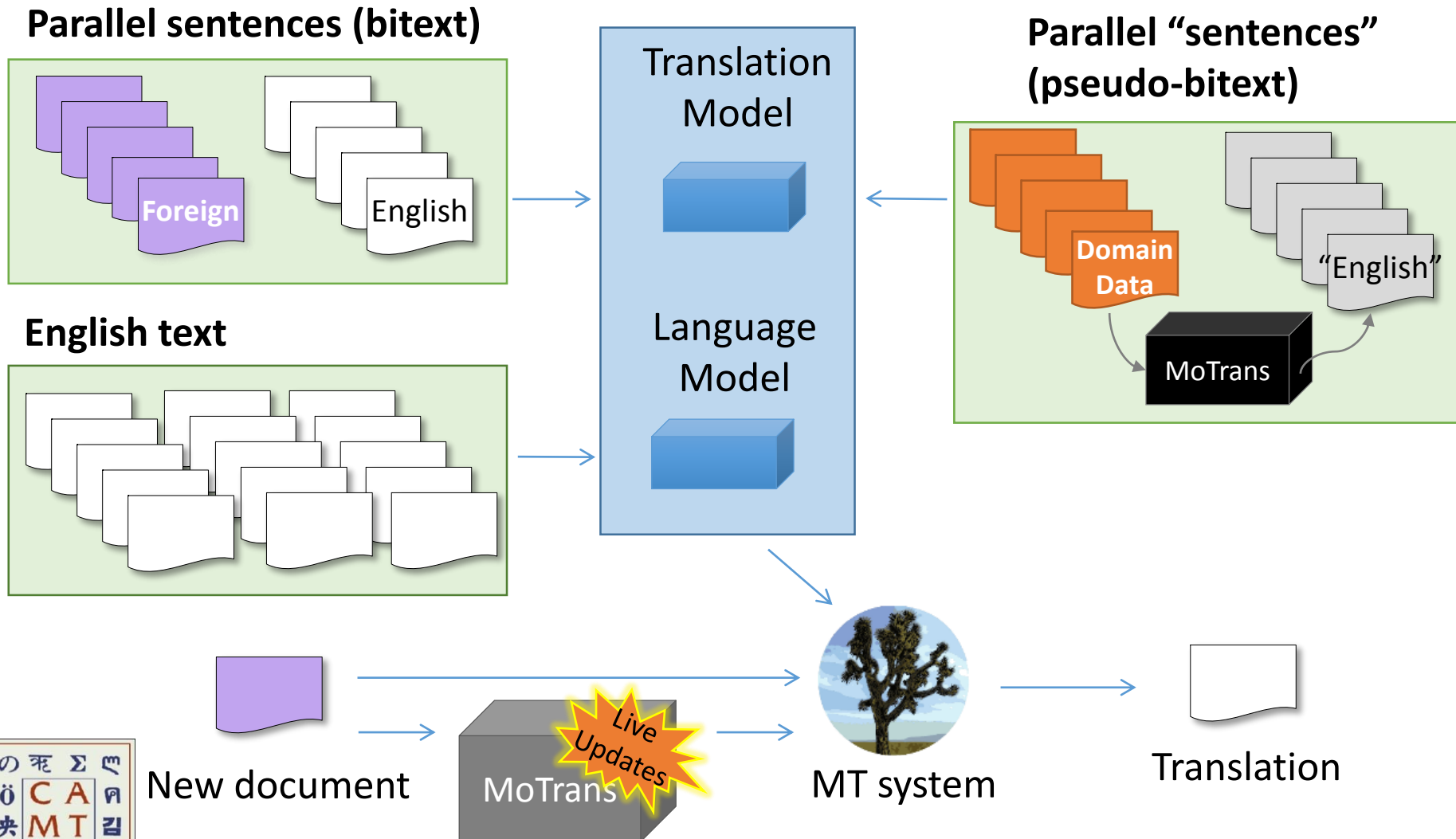
Building the Hybrid: Black Box



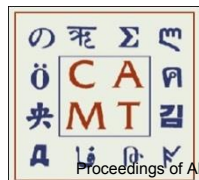
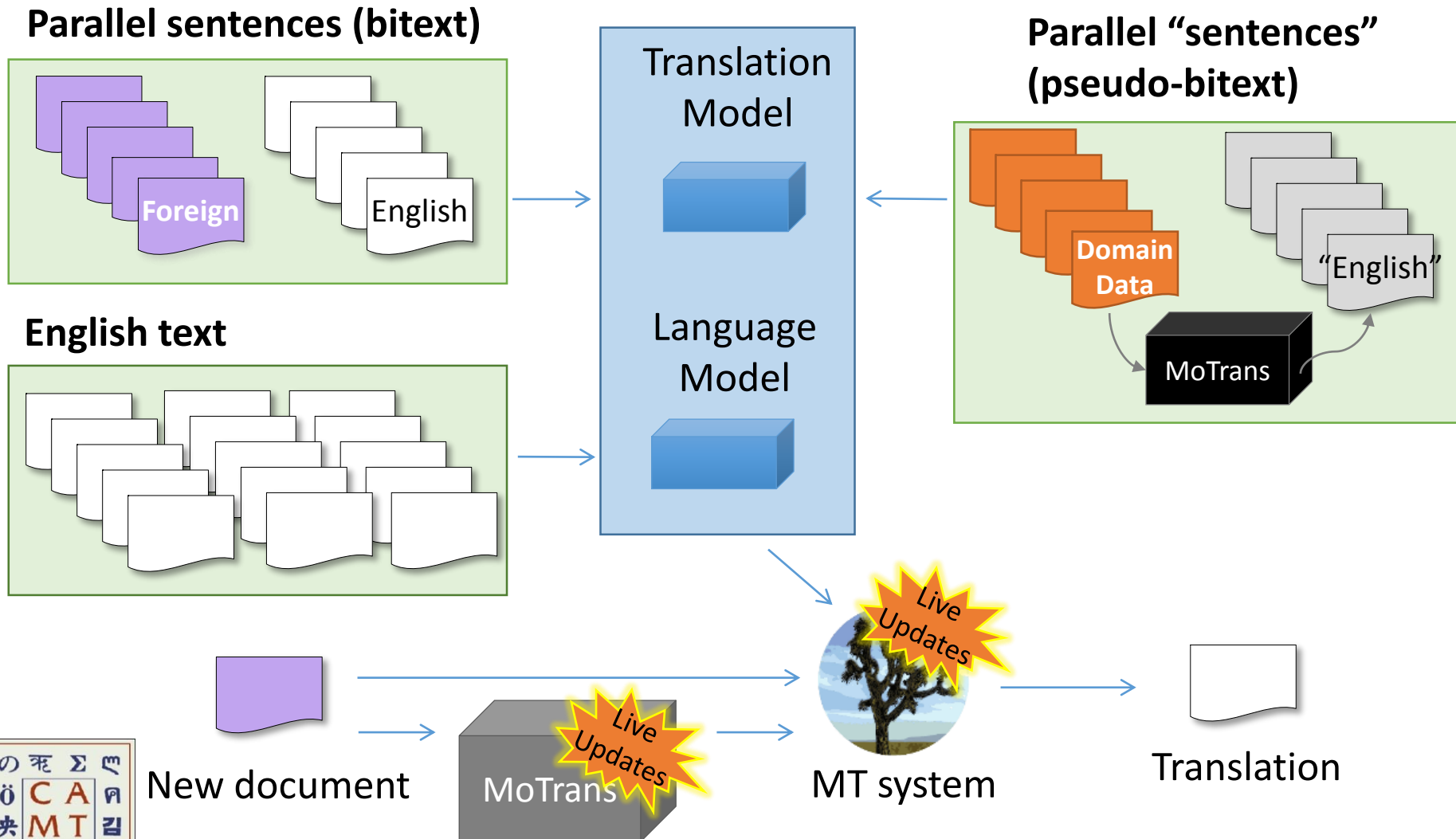
Building the Hybrid: On Demand



Building the Hybrid: Live Updates



Building the Hybrid: Live Updates



User View

Translation Options

- Reset Add 2nd Pass User Settings
- Input Language: Spanish
- Output Language: English
- Encoding: Unknown
- Translator: Recommended (D)
- Topic Dictionary: n/a
- User Dictionary: n/a
- Output Format: Auto Select
- Annotations: [Dropdown]
- Multi-Pass Options: (0) [Dropdown]
- Output Format: Tab Window File

Text to Translate Translate-1

Original Text	Motrans	MoJo
<p>En el primer debate de los precandidatos republicanos a la presidencia de los Estados Unidos organizado y transmitido por la cadena estadounidense FOX News el 7 de diciembre de 2015, el Precandidato mantuvo una postura que fue objeto de una gran polémica. Destacó que el sistema político de su país se encuentra presuntamente «roto» y que él y los Estados Unidos «no tienen tiempo de ser políticamente correctos», argumento sustentado en que el país ha perdido protagonismo y competitividad en el escenario global, esto según declaraciones del propio magnate. Él también destacó que no descartaría la posibilidad de ser un candidato independiente a la presidencia del referendo país si no llegase a ser nominado formalmente como "Candidato Presidencial por el Partido Republicano" lo cual fue objeto de críticas dentro de las filas de dicho partido.</p>	<p>In the first debate of the republican primary candidates to the presidency of the organized United States and transmitted by the American chain FOX News the 7 December 2015, the maintained primary candidate a posture that was l object of a great controversy. Emphasized that the political system of their country is presumably «l rotale» and that he and the United States «they haven't time of to be politically correct», argument sustained in which El País has lost prominence and competitiveness in the global stage, this according to statements of the own magnate. He also emphasized that wouldn't reject the possibility of to be an independent candidate to the presidency of the referenced country or else came to be nominated formally as "Presidential Candidate by/for the Republican party" which was l object of criticism within the strings of sad left.</p>	<p>In the first debate of the republican pre-candidates to the presidency of the United States organized and transmitted by the American chain Fox News the 7 of December of 2015, the Pre-candidate maintained a position that was object of a great controversy. It emphasized that the political system of its country is presumably «broken» and that it and the United States «do not have time to be politically correct», argument sustained in which the country is lost protagonism and competitiveness in the global scene, this according to declarations of the own tycoon. It also emphasized that he would not discard the possibility of being an independent candidate to the presidency of the referred country if he did not arrive to be name formally as "Presidential Candidate by the Republican Party", which was object of critics within the rows of this party.</p>

Translation Options: Tab Window File

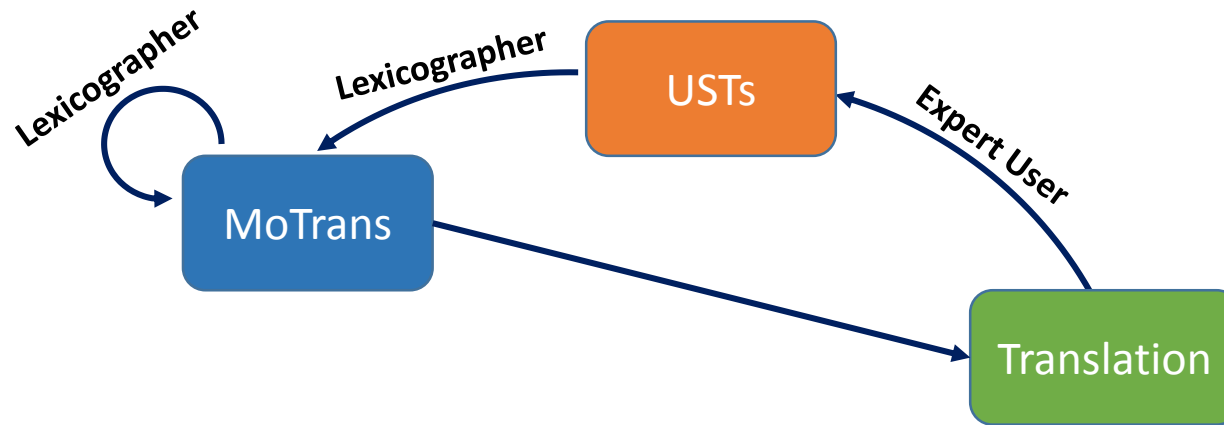
Annotations: [Dropdown]

Multi-Pass Options: (0) [Dropdown]

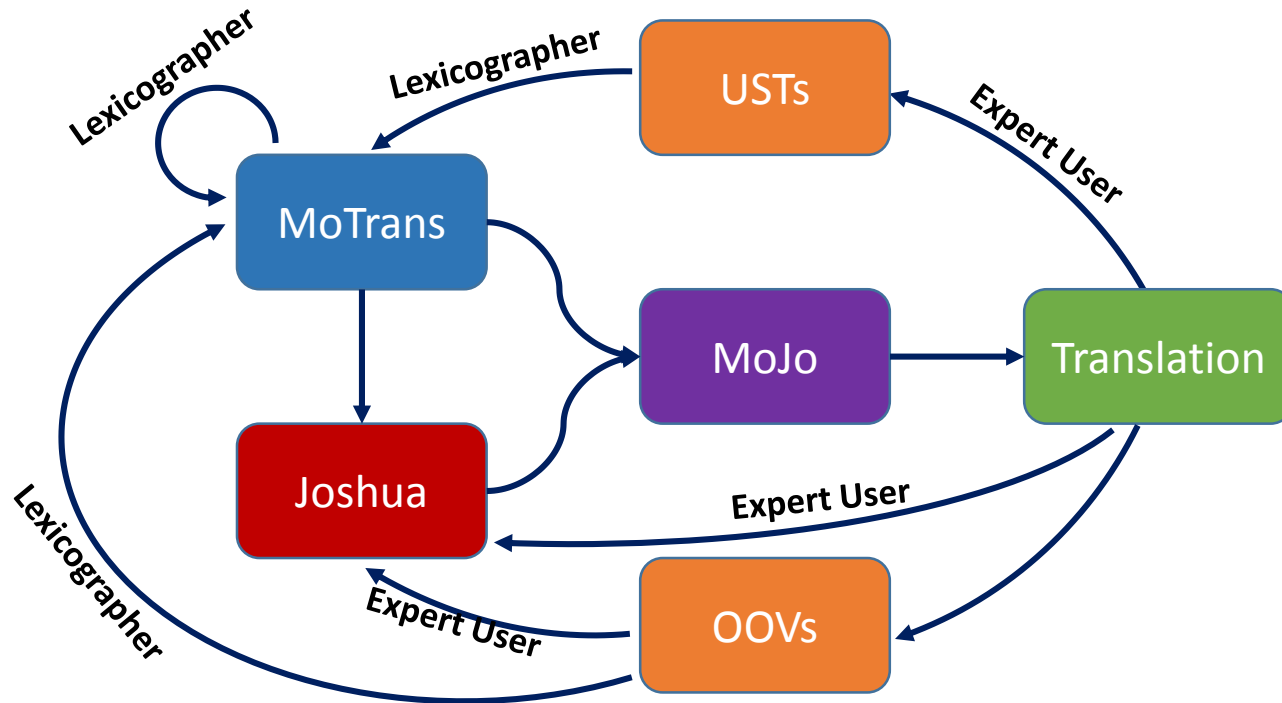
Language Tools: 你好 hello, AaE, 翻译 fan yi, windows, RQE



Workflow



Workflow



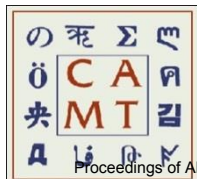
Conclusions

- Mojo online in CyberTrans “soon”
 - Productization in progress
 - Starting with Spanish
 - Other languages will follow
 - Shortly thereafter will be available as add-on for CyberTrans distributions

Questions?



Backup Slides



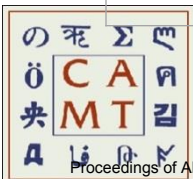
Example (Portuguese)

System	Output
Motrans	however, the balloting already finished but not yet there are results ends.
StatMT	Meanwhile, the ballot finished but there is still no final results.
Hybrid	However, the ballot has finished but there is still no final results.
Human	However, the audit is over but there still are no final results.



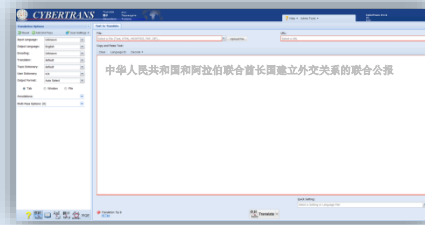
Example (Arabic)

System	Output
Motrans	And adds Dr. Syrian/Suri that this is reason viewing it to some the patients/al-Murdi who come to him after the passing of the time of their treatment, in addition to the resistance of their bodies to drugs specific/Mu'inah.
StatMT	He said d. Sorry, this is a reason to see some patients who come to him after it was too late for treatment, as well as to resist their bodies certain drugs.
Hybrid	D. Syrian adds that this is the reason see some patients who come to him after the passing of time their treatment, as well as to resist their bodies specific drugs.
Human	Dr. Sory also stated that is why he sometimes had diseased people come in when it was too late for treatment and why there was resistance to certain drugs.



Center for Applied Machine Translation

中华人民共和国和阿拉伯联合酋长国建立外交关系的联合公报



The Peoples Republic of China and the United Arab Emirates established diplomatic relations with the joint communique

- DoD-recognized Center of Excellence for Machine Translation
 - Serving the US Government for over 14 years.
- Flagship product: **CYBERTRANS**
 - Integrated suite of automated tools for MT, language and encoding identification, spelling and text enhancement, and encoding conversion.



CyberTrans Usage

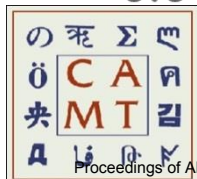
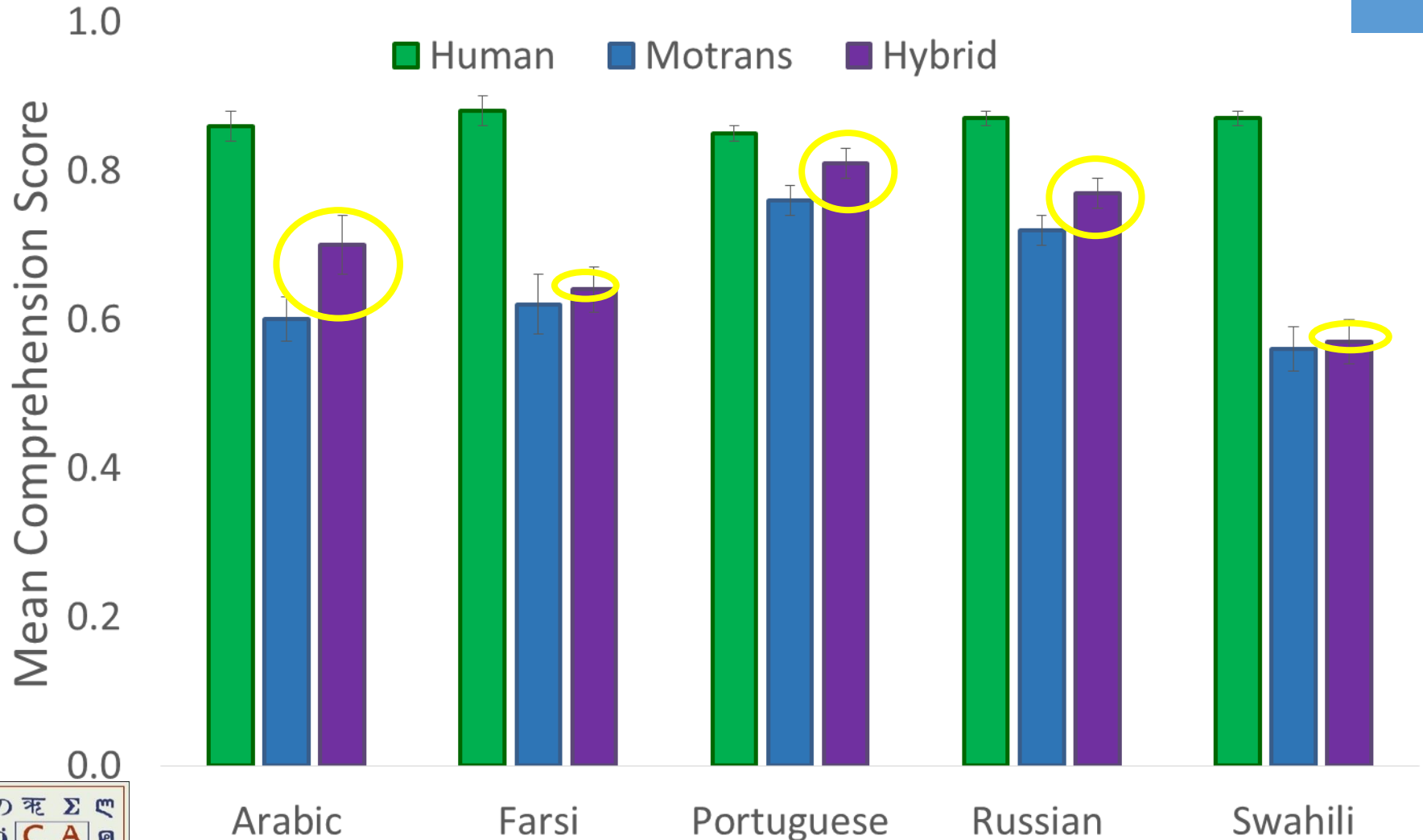
- Primary customers **don't** know language
 - Triage, filtering, selection
 - Free translators from spending time on low value material
- Secondary customers **do** know language
 - Gisting
 - Seed translation



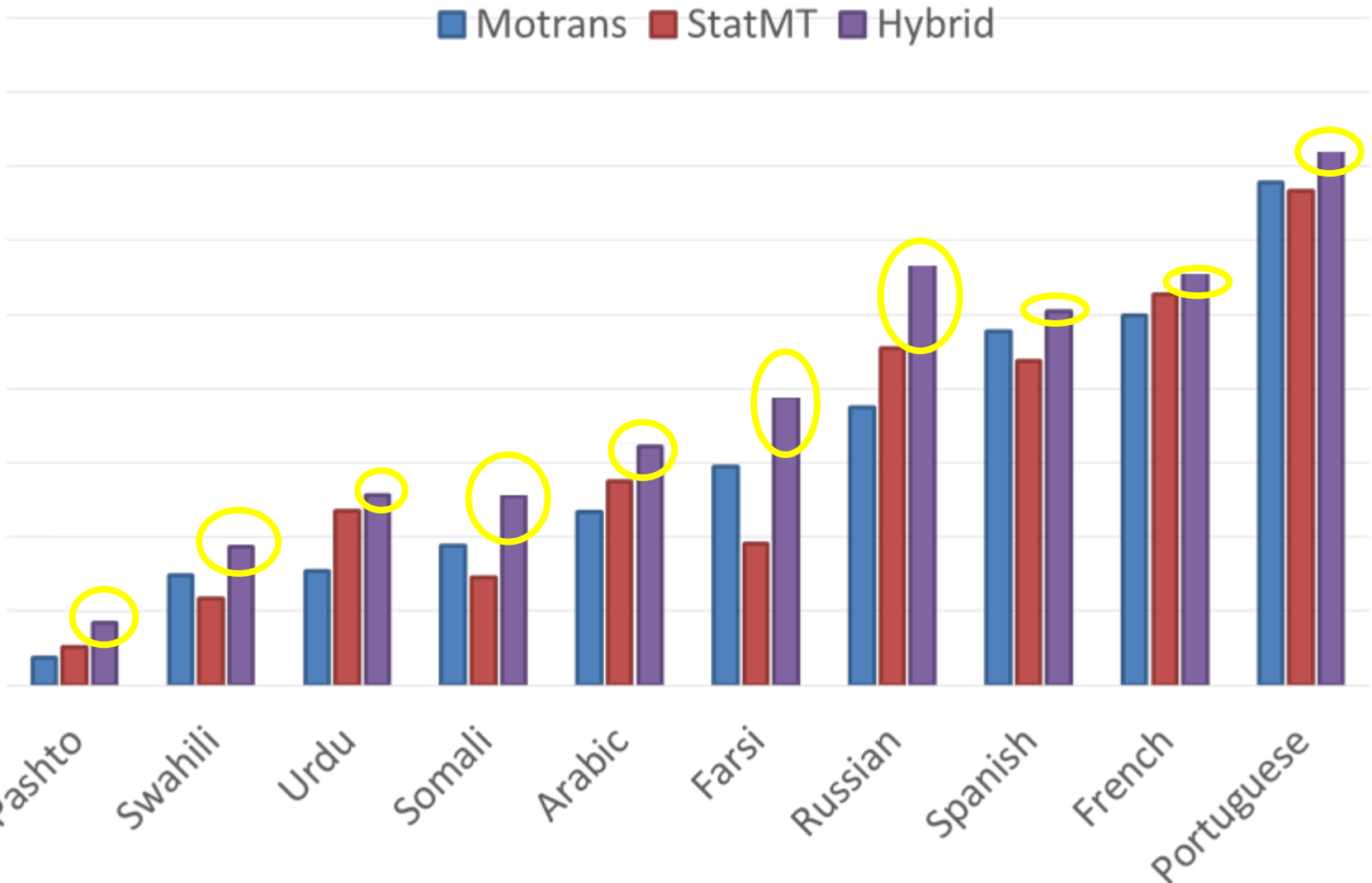
Hybrid Approaches



Human Comprehension Results



BLEU Scores (In-domain)

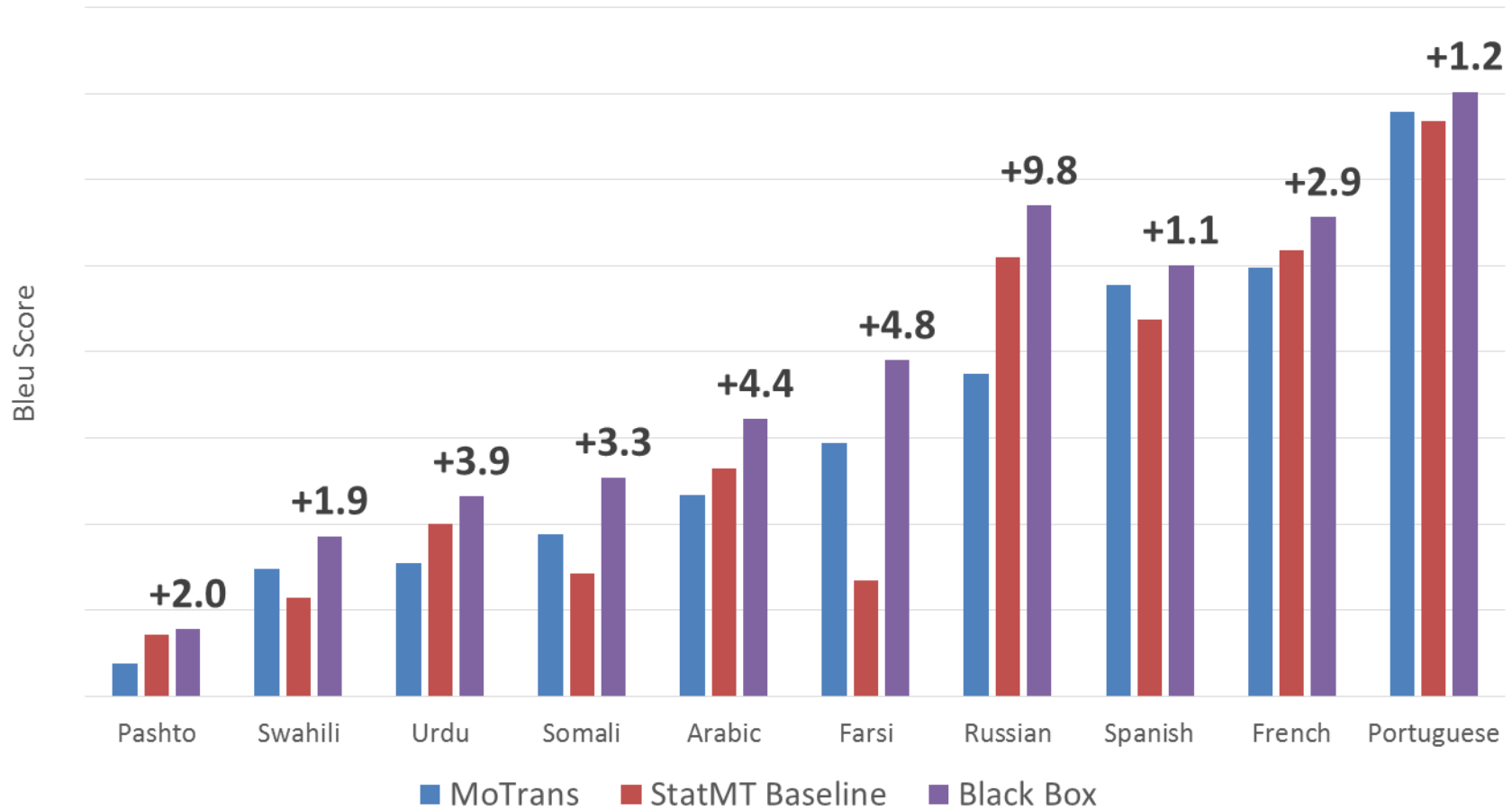


Three Hybrid Approaches

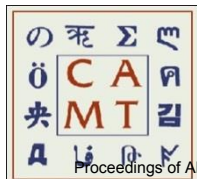
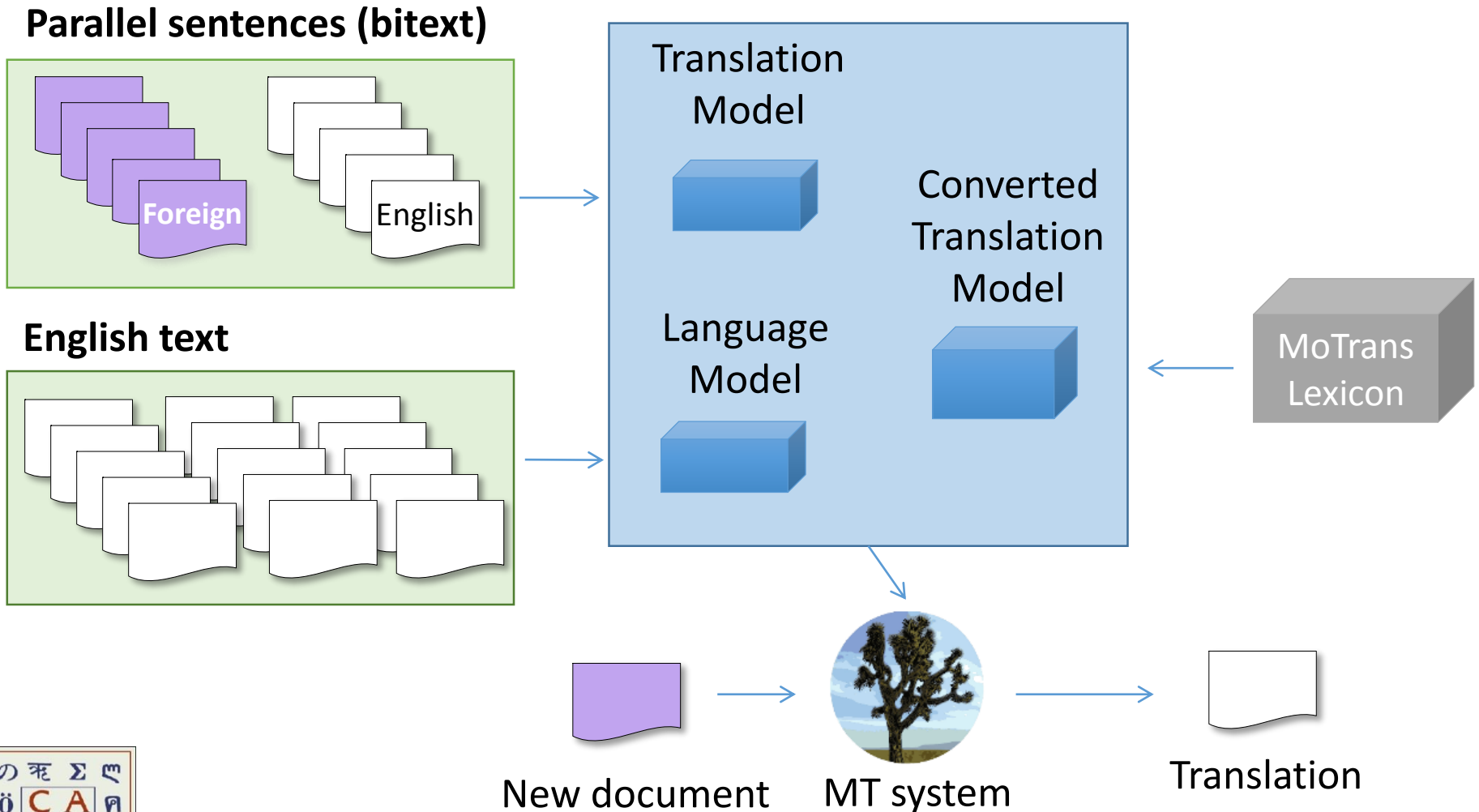
<p>Black Box</p>	<p>Run Motrans on large text, build a regular StatMT model</p>
<p>Direct Conversion</p>	<p>Convert Motrans rules directly to StatMT phrase pairs</p>
<p>On Demand</p>	<p>StatMT system queries Motrans, incorporates its suggestions</p>



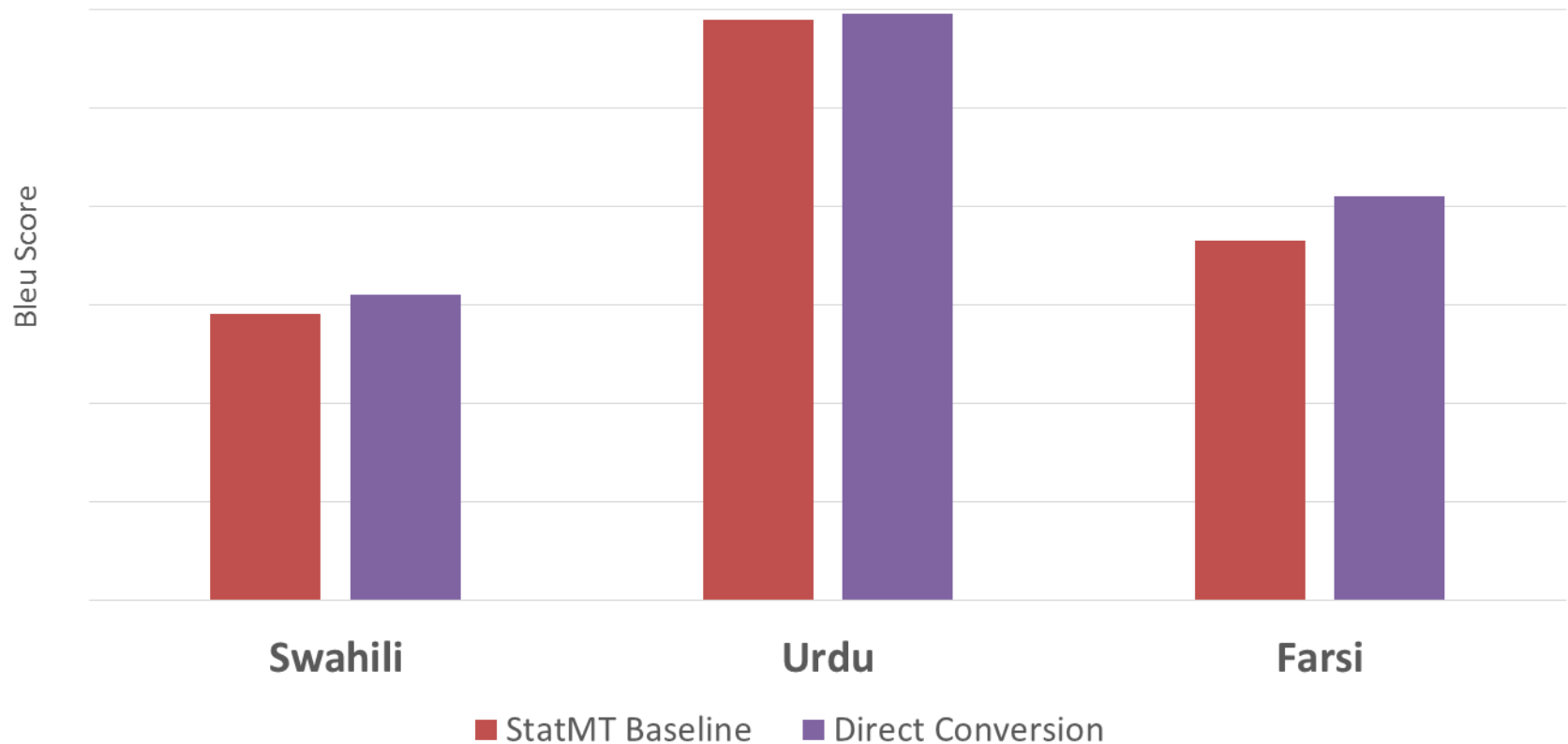
Black Box results (in-domain)



Direct Conversion Approach

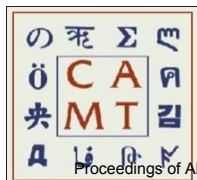


Direct Conversion Results

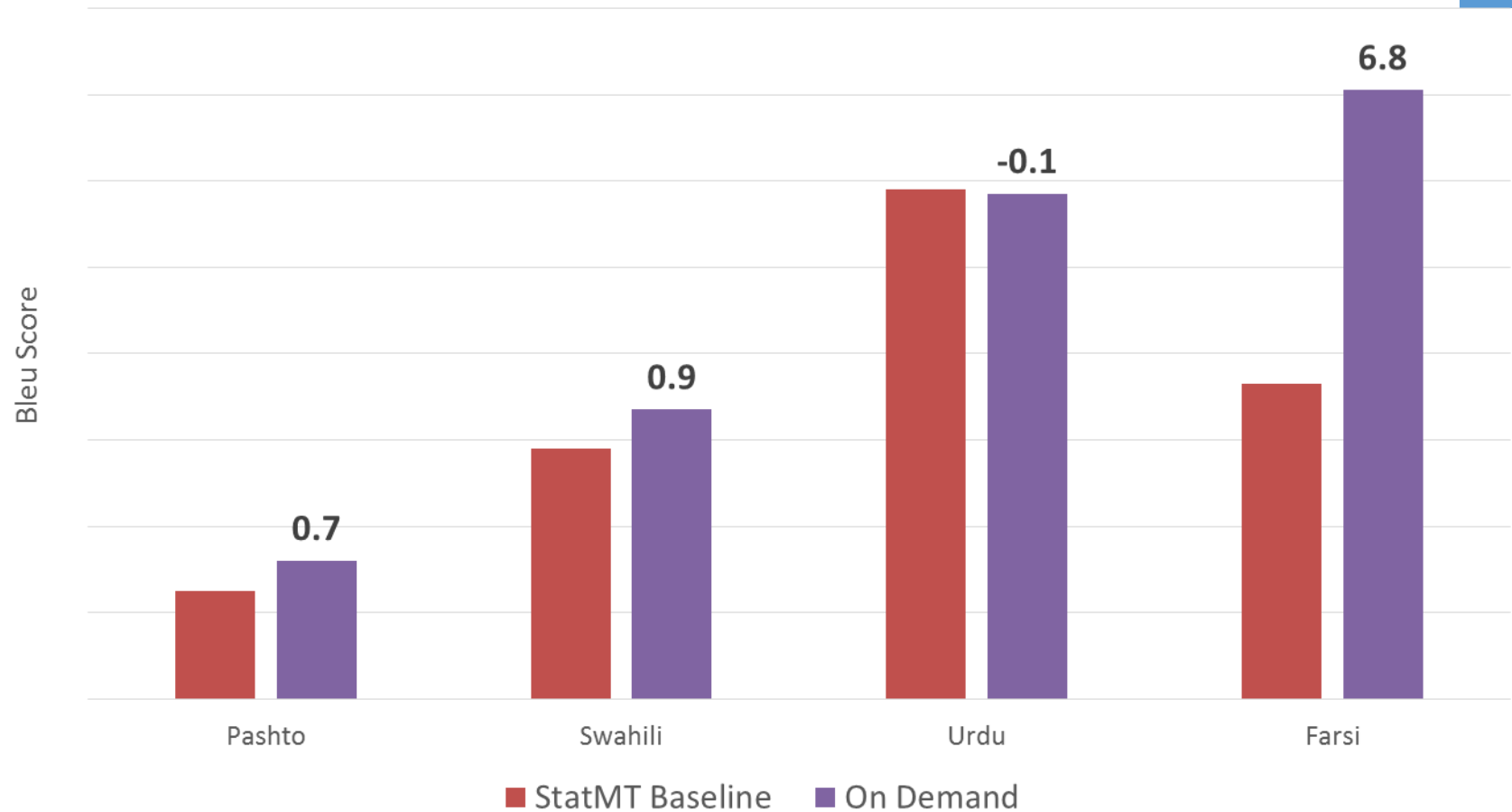


Challenges of the Conversion Approach

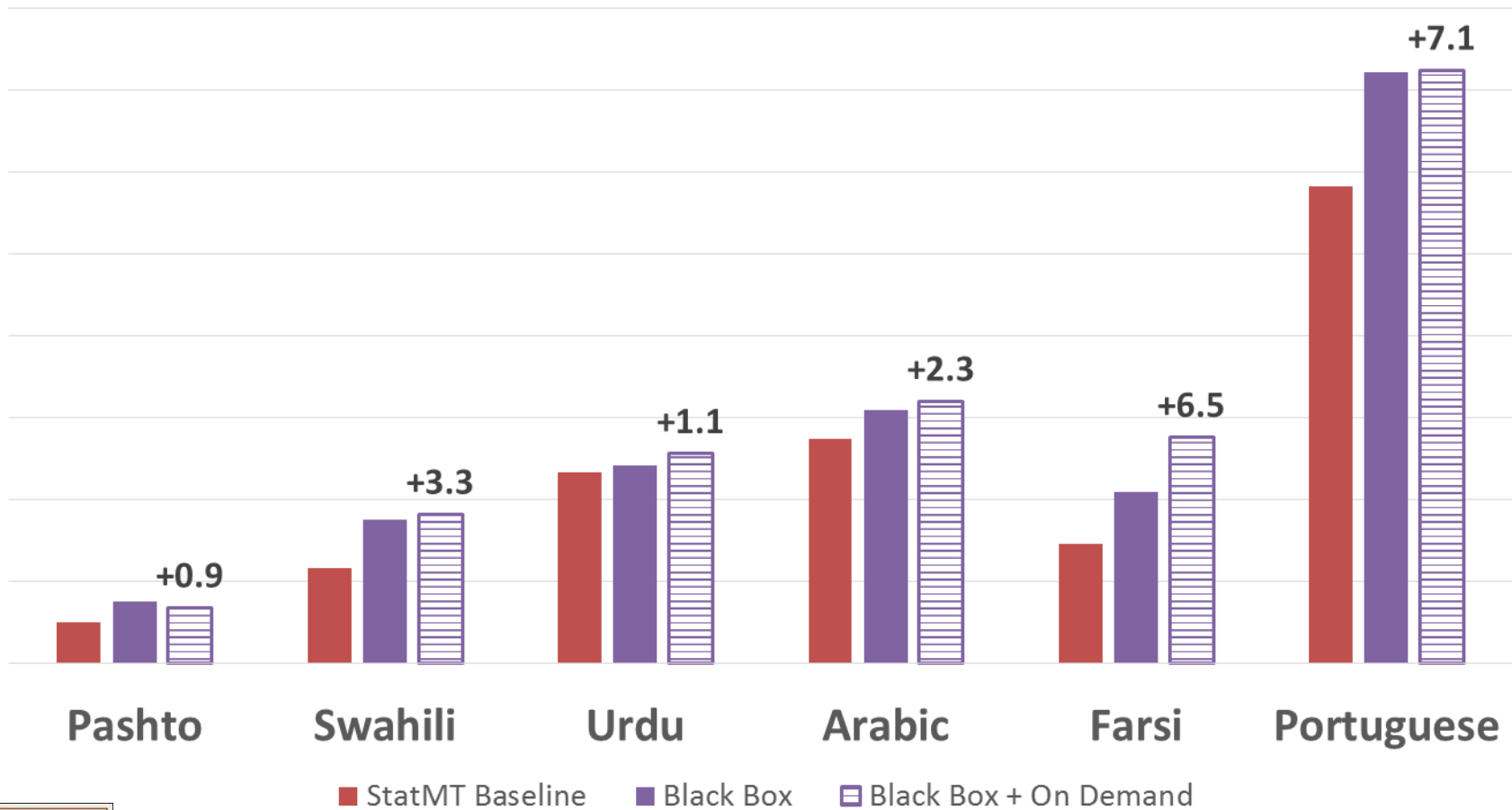
- MoTrans rules have complex, unique syntax
 - Parsing requires in-depth knowledge of MoTrans
 - Some rule types not yet handled
- Not feasible to expand all rules
 - Some rules apply to full sentence, not phrases
 - Rule chaining creates exponential possibilities
- Cybertrans does pre- and post-processing that's not described in lexicon



On Demand Results



Combined Hybrid Results





SYSTRAN Software

Intelligent Language Technologies

**Building Renewable
Language Assets in
Government Domains**

Agenda

- Why SYSTRAN
- Background of Government Work
- Current Direction

Why SYSTRAN



Expertise

- Leadership role for 40+ years in machine translation
- 200+ language technology professionals in the US, France and Korea



Best Fit Technology

- Multiple deployment options (on-premise, desktop, mobile, cloud)
- Specialized engines for domains



Continuous Improvement

- Re-investment of 25% revenue into R&D
- Partnerships with academia, currently focused in the area of Neural MT



Trusted Industry Partner

- Organic growth through government partner referrals
- Flexible approach to meet mission requirements

Background of Government Work

IT Systems



- IC Networks
- Command, Control, Communication, Computers & Intelligence (C4I) Systems

Language expansion resulting from world events, starting with Russian

Language Requirements



Technology Evolution



1. Migration off the mainframe
2. Arrival of the internet
3. Statistical MT
4. **Neural MT**

What have we done in the past

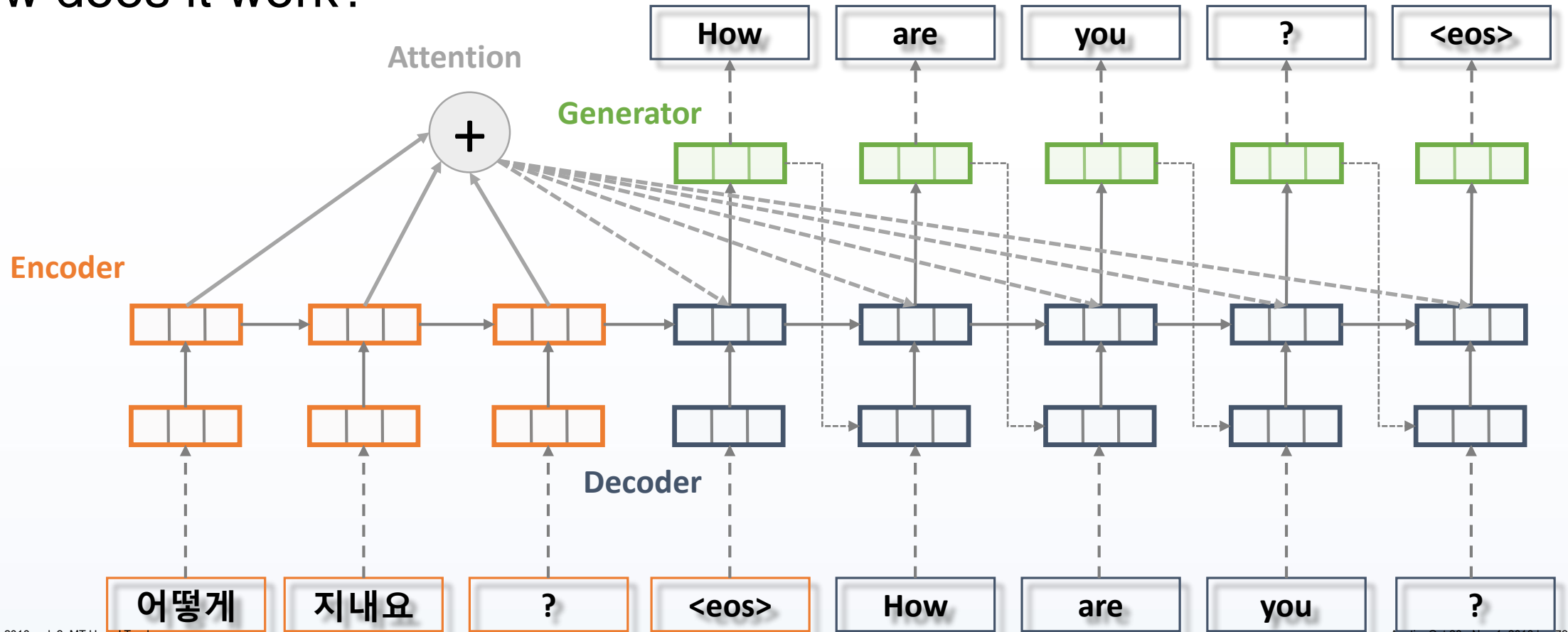
- Full Rule-Based Backbone for English target
 - Morphology, part of speech, normalization
 - Mix rules and statistical decision models
- Statistical post-editing (SPE)
 - Statistical layer to fine tune output based on bilingual corpus
 - Allows for greater customization
 - Allows for quality translations in domains with little bilingual corpus
- Dictionaries
 - Extensive dictionaries – user defined
 - Cover terminology not found in bilingual corpus
 - Cover low resource domains and languages – Science and Technology
- Entity recognition

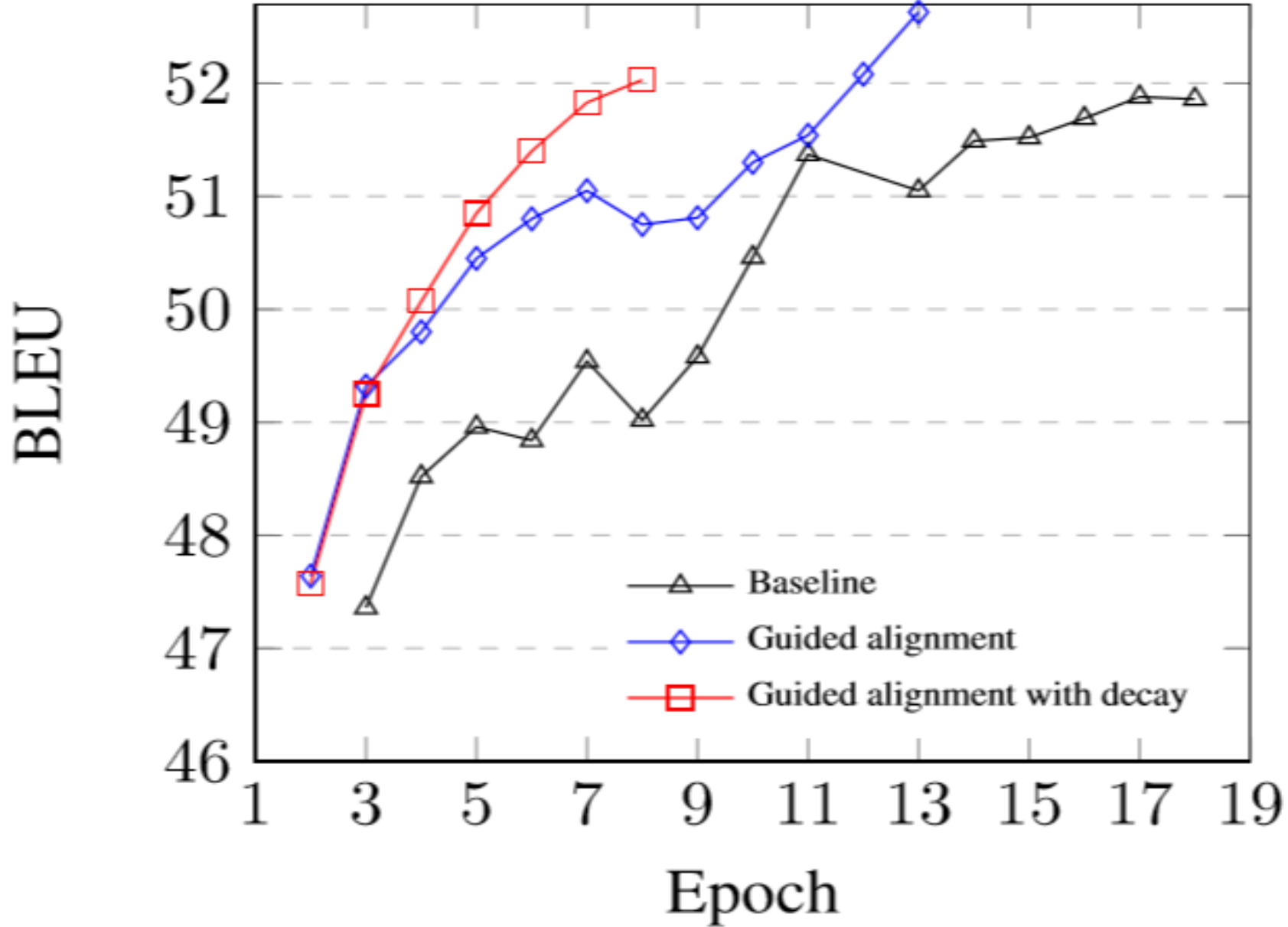
Pure Neural Machine Translation

- <http://demo-pnmt.systran.net>
 - 14 language pairs with more on the way
- <https://arxiv.org/pdf/1610.05540v1.pdf>
- Systran specialization
 - Architecture
 - Pre-processing
 - Features
 - Post-processing
 - Domain specialization
- Incorporate standard customization features
 - User Dictionaries, Translation Memory etc.

Neural Machine Translation

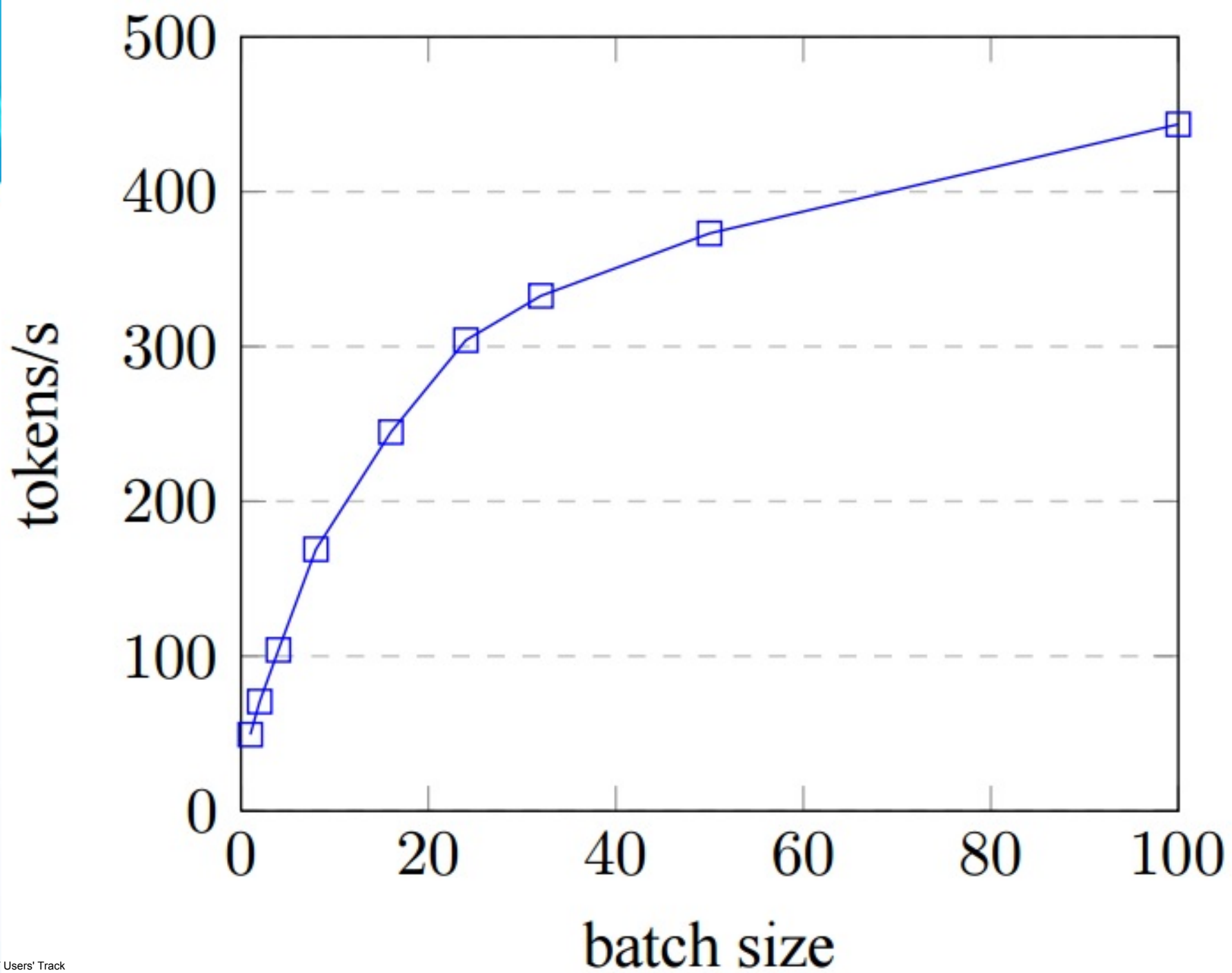
- How does it work?





Architecture

- seq2seq-attn (<https://github.com/harvardnlp/seq2seq-attn>)
 - Harvard NLP group
 - Sequence to sequence RNN
 - Guided attention with decay
 - Features in source and target
 - Open source with ability to tune several parameters
- Run time can now be done on CPU
 - 4 threads on a desktop Intel i7 CPU
 - Pruning
 - Distillation
 - Batch modes with beam size



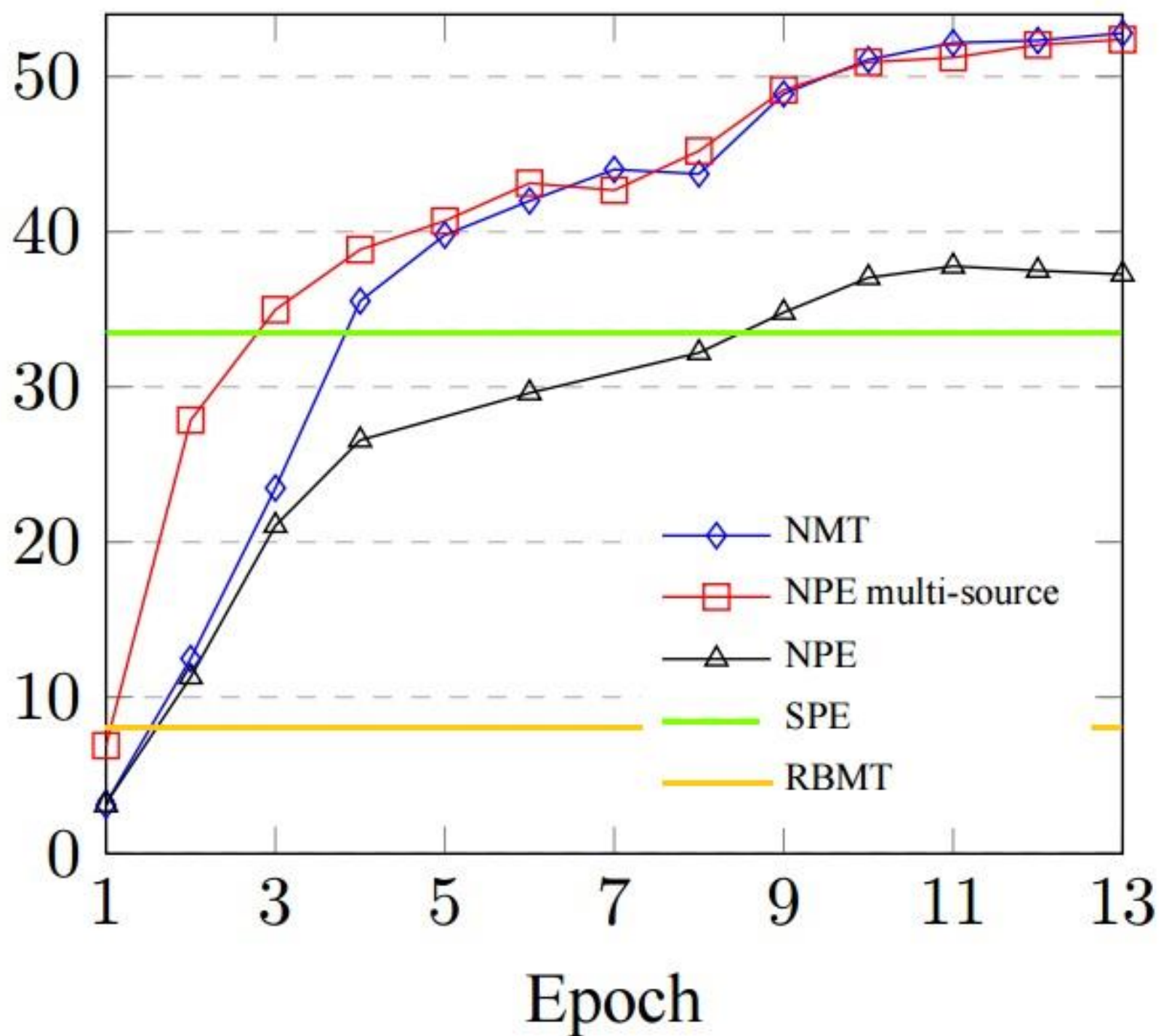
Pre-processing

- Normalization
- Tokenization
 - Word based
 - Special tokenization for CJK, German and Arabic
 - More experimentation (BPE, character, combo)
- Entity recognition – replace with token (`__ent_numeric`)
 - Replacement needs to be in both to learn

Linguistic Features

- Dictionaries, entities, capitalization
 - Maybe include part of speech, parse
- Formality mode – Korean
 - Domain mode
- Include traditional translation
 - Smart "Neural Post-Editing"
- Can't neural networks learn everything, given enough data?
 - Depends on complexity of language and amount of data
 - We know NMT has issues with OOV

BLEU

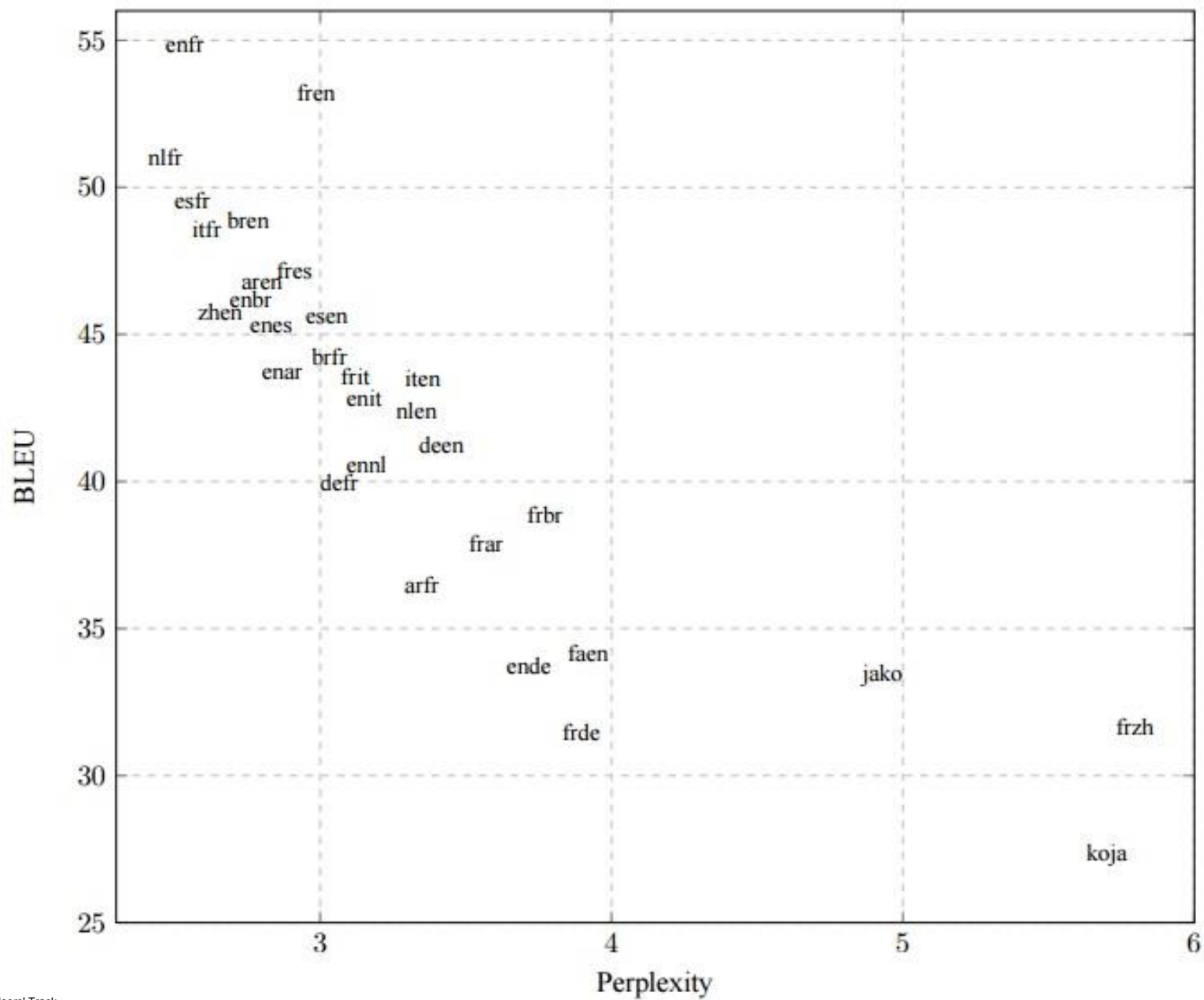


Post-processing

- Restore entities and dictionary terminology
- Apply features (capitalization)
- Restore punctuation
- Out of Vocabulary
 - Look them up with dictionary
 - Use SMT-style phrase table
 - Allow NMT to transliterate

Domain specialization

- Stack neural networks
 - Adds another neural network trained on in-domain corpus
 - Requires some tweaks on vocabulary
 - Enable full specialization in a few hours
- Marker for domain
- Synthetic corpus
 - Makes use of well-written, in-domain sentences from target language
- Inject terminology
 - Full support for User Dictionaries
 - Explicit recognition of entities
 - Other linguistic knowledge?



Examples

- FARSI: فرهنگ هر کشور دارای هویت و ویژگی های خاص خود می باشد.
- SPE: Culture of each country has identity and **its** special features.
- PNMT: **The** culture of each country has **its own** identities and particularities.

- FARSI: به دلیل اختلافات ایدئولوژیکی و عدم توافق بر سر اهداف مورد حمله، بسیاری از هسته ها از این سازمان تازه بیرون کشیدند یا هیچوقت به آن نپیوستند.
- SPE: **For an ideological reason for disagreements and discordance**, goals of **case of attack** pulled out **many of nuclei** of this new organization over or **they never joined that**.
- PNMT: **Many cells** pulled out of the new organization or **never joined it because of ideological differences and disagreements** over the **targeted targets**.

QUESTIONS

Beth Flaherty
Beth.flaherty@systrangroup.com

Joshua Johanson
Joshua.johanson@systrangroup.com