



AMTA 2018

March 17 - 21, 2018

Boston, MA, USA

The 13th Conference of
The Association for Machine Translation
in the Americas

www.conference.amtaweb.org

TUTORIAL

March 17, 2018

**Corpora Quality Management for MT
- Practices and Roles**

Presenters: Silvio Picinini (*eBay*), Pete Smith (*University of Texas Arlington*)

Nicola Ueffing (*eBay*)

Quality of Data: a Data Scientist's viewpoint

Nicola Ueffing, Daniel Stein
eBay MTScience Team
2018-03-17, AMTA tutorial



Overview

**Motivation &
Background**

**What would
Moses do?**

**eBay best
practices**

**Experimental
results**

Motivation & Background

Automatic checks of data quality

Machine Translation system needs

- a lot of training data
- high-quality training data

Manual check of bilingual data not feasible in practice

- Large size of data
 - Quick turnaround
 - Data scientist might not be bilingual expert
 - Cost/benefit: worth spending manual effort on in-domain data; not worth on out-of-domain data
- => perform automatic checks to validate data quality

Motivation & Background

Overview SMT

Statistical Machine Translation

Bilingual data

Monolingual data

Cleaning / Sampling

Preprocessing / Tokenization
(and maybe splitting)

Word Alignment

Phrase
Extraction

Language
Modelling

Tuning (Automatic Evaluation)

Verification (Manual Evaluation)

Motivation & Background

Overview SMT

Statistical Machine Translation

Bilingual data

Monolingual data

Cleaning / Sampling

Preprocessing / Tokenization
(and maybe splitting)

Word Alignment

Phrase
Extraction

Language
Modelling

Tuning (Automatic Evaluation)

Verification (Manual Evaluation)

Motivation & Background

Overview SMT & NMT

Statistical Machine Translation

Bilingual data

Monolingual data

Cleaning / Sampling

Preprocessing / Tokenization
(and maybe splitting)

Word Alignment

Phrase Extraction

Language Modelling

Tuning (Automatic Evaluation)

Verification (Manual Evaluation)

Neural Machine Translation

Bilingual data

Monolingual data

Cleaning / Sampling

Preprocessing / Tokenization
Byte-Pair Encoding, Word Embedding

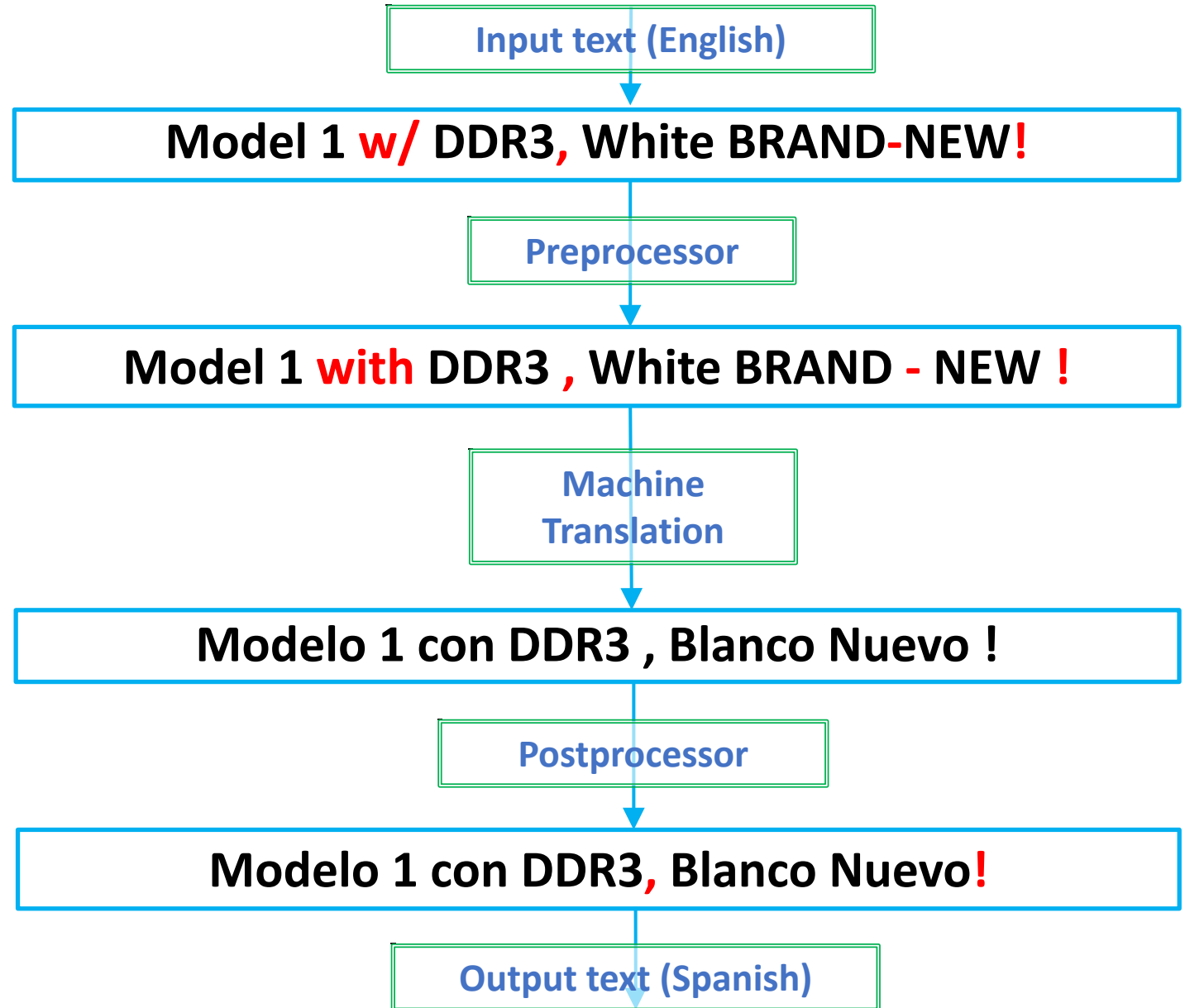
Neural Net

Tuning (Automatic Evaluation)

Verification (Manual Evaluation)

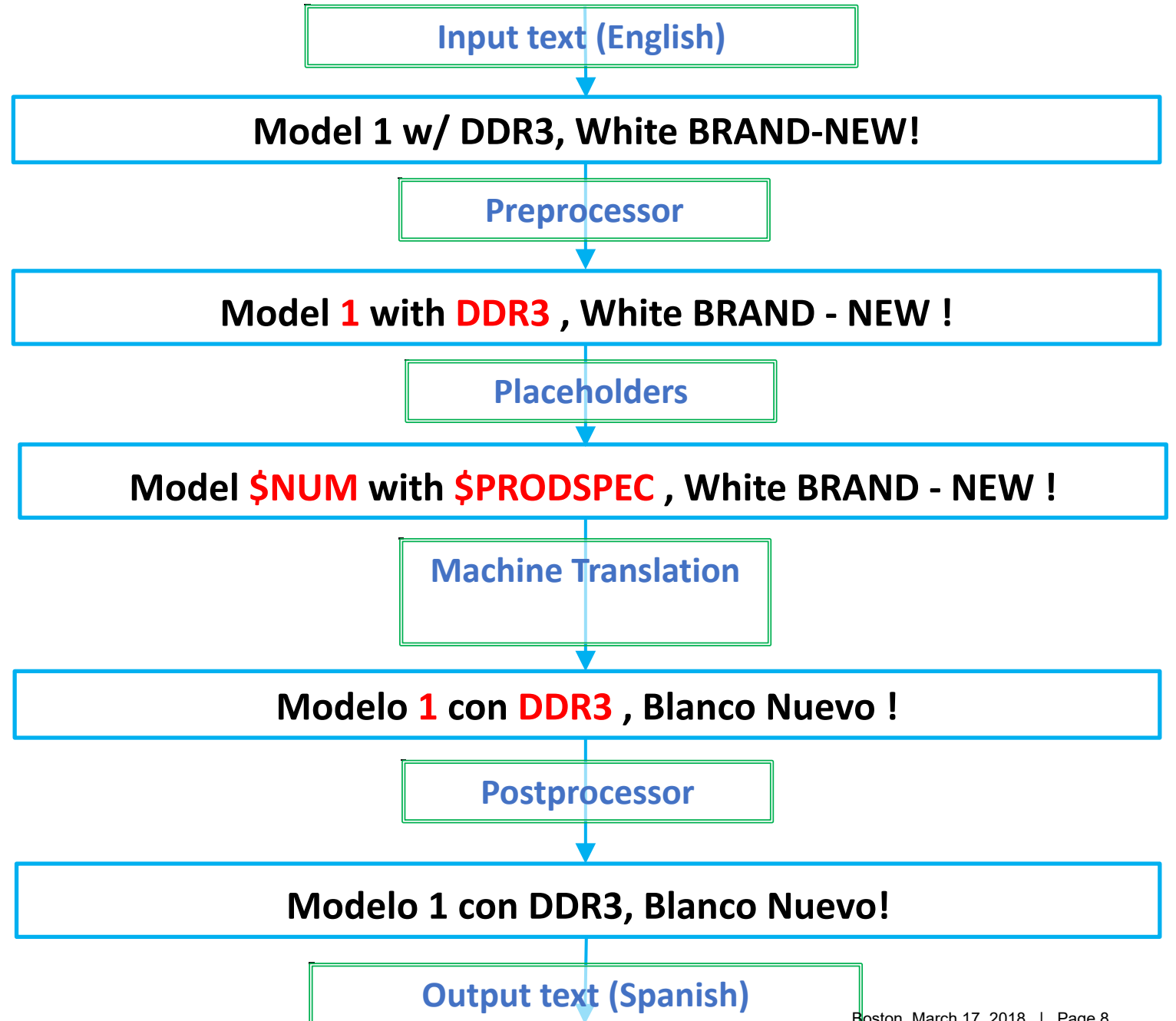
Motivation & Background

Preprocessing



Motivation & Background

Preprocessing





CUTE HELLO KITTY Stuffed Plush 12" so CUUUUUUTE!!!!(FREE SHIPPING in USA)

Condition: **New**

Quantity:

Last one / 2 sold

Price: **US \$9.99**

Buy It Now

Add to cart

[Add to watch list](#)

[Add to collection](#)

Last item available

More than 66% sold

Longtime member

Shipping: **\$16.55** USPS First Class Mail International / First Class Package International Service | [See details](#)

See details about international shipping here. [?](#)

Item location: Sun Prairie, Wisconsin, United States

Ships to: United States, Europe, Canada, Australia | [See exclusions](#)

Delivery:  Estimated between **Mon. Mar. 5 and Fri. Mar. 23**
Please note the delivery estimate is **greater than 7 business days.**

Payments: **PayPal** | [See details](#)

Returns: 14 day returns. Buyer pays for return shipping | [See details](#)

Seller info

100% Positive f

[Save this \\$](#)

Contact seller

Visit store: 

See other item

Any C
So Ma
Possi

Give a d
eBay Gif

Get It No

Title:

CUTE HELLO KITTY Stuffed Plush 12" so CUUUUUUTE!!!!(FREE SHIPPING in USA)

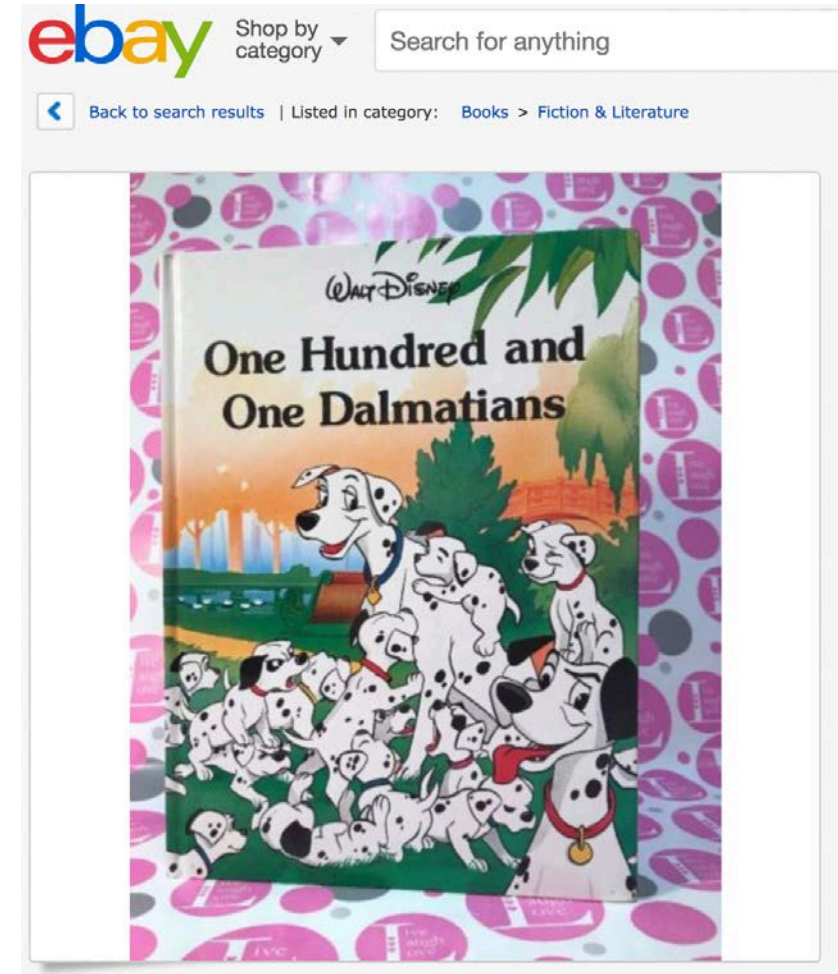
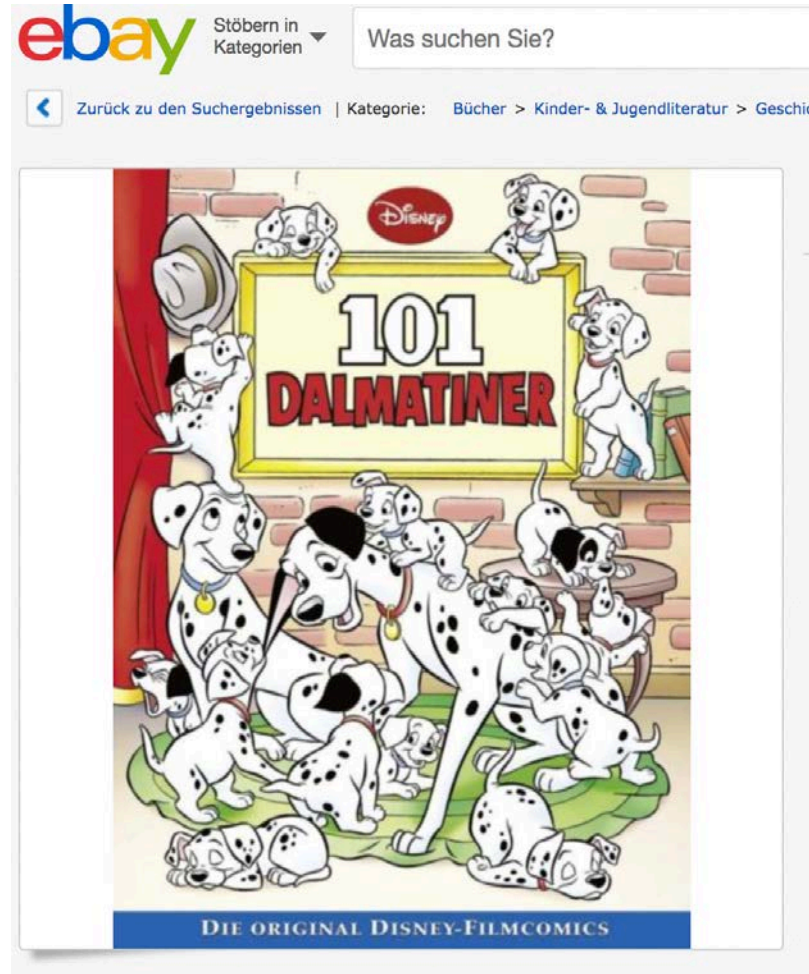
Preprocessed:

CUTE HELLO KITTY Stuffed Plush 12" so CUUUUUUTE !!! (FREE SHIPPING in USA)

Example idea: „what if I remove all sentences with number mismatches?“

Automatic Filtering

... careful there



Digits or numbers?

English: All-season tires

German: Ganzjahresreifen

Italian:

Automatic Filtering

... careful there!

The screenshot shows an eBay search result for tires. At the top, the eBay logo is visible with a dropdown menu for 'Scegli la categoria' and a search bar containing 'Cerca qualsiasi cosa'. Below the search bar, there are navigation links: 'Torna alla homepage', 'In vendita nella categoria: Veicoli: ricambi e accessori > Auto: cerchi e pneumatici > Pn...', and 'Altro 1x Pneumatico 4 stagioni Fortuna Ecoplus 4s 175/6...'. The main content area is titled 'Chi ha visto questo oggetto ha visto anche' and displays three recommended tire products. The first product is 'Pneumatici 4 stagioni 175/65/13...' priced at EUR 40,82 with free shipping. The second is 'Pneumatici 4 stagioni 155/65/13...' priced at EUR 29,66 with free shipping. The third is 'Pneu gioni' with a price of EUR. Below this, the main product listing is for 'Pneumatici 4 stagioni 175/65/13 80'. The product title is highlighted with a blue box. It features a star rating and the text 'Scrivi una recensione per primo.'. The condition is listed as 'Nuovo'. The quantity is set to 1, with 4 items available and 4 sold. The price is EUR 32,79. There are buttons for 'Compralo S' and 'Aggiungi al c'. At the bottom, there are links to 'Aggiungi a Oggetti che c' and 'Aggiungi alla collezione'.

Automatic Filtering

.... seriously, be careful there!

Digits or numbers?

The Associated Press Stylebook:

Spell out the numbers **one through nine**; for 10 and up, use Arabic numerals. For ages and percentages, always use Arabic numerals, even for numbers less than 10.

Dublin City University: „Numbers in academic writing“

For general academic writing, you need to write these numbers in words: **all numbers under one hundred** (e.g. ninety-nine) rounded numbers (e.g. four hundred, two thousand, six million) and ordinal numbers (e.g. third, twenty-fifth).

Duden (German) „Schreibung von Zahlen“

Eine früher gültige Buchdruckerregel, nach der generell die Zahlen von 1 bis 12 in Buchstaben und die Zahlen ab 13 in Ziffern zu schreiben sind, gilt heute nicht mehr!

A former rule from book printing, where all numbers between **1 and 12** have to be written as words, and all numbers starting from 13 as digits, is no longer valid!

What would Moses do?

Standard cleaning procedures in MT community

Moses: open-source toolkit for statistical phrase-based MT

Clean training data and remove ...

- all words exceeding length threshold, e.g. 1000 characters
- all sentence pairs with $\text{length}(\text{source})/\text{length}(\text{target}) > 9$ or vice versa
- all sentence pairs where either source or target length (in words) is outside given interval, e.g. [1,80]

eBay Best Practices

Our data

**Filtering phrase
tables**

Filtering data

**Experimental
results**

eBay Best Practices

Our data

Examples of eBay data which we translate

- Item descriptions
 - suitable for most lamps by Kandem, Kaisee-Idell, Midgard, Reif-Dresden, HALA, HELO, LBL, AKA and many others
 - *** Adorable face ***
 - Impressive table clock with brazed figure group of two putti playing music, framed by two handle pitchers; partially painted clock face with Roman and Arabic numerals and Fantaisies-Roskopf hands; key wind-up; with gold-plated antique relief; mechanism marked "Japy Freres & Cie Paris"; Cleaning of mechanism is recommended.
- Item titles
 - Sausage boiler broth boiler butcher's boiler boiler pot boiler insert
 - A single hand iron clock, Zappler, front commuters, Baroque
 - Rasta wig with dreadlocks Rasta Hat Rasta braids
- Search queries (lower-cased)
 - iphone 8 / iphone 8 case / battery iphone 8
 - apple
 - renault megane 1.6 16v 2000
 - shirt / shirts man / shirt real madrid 2016 / tommy shirt xxl

eBay Best Practices

Phrase table filtering

What do we do?

Depends on use case, data, etc.

Some options for **phrase table** filtering:

- If a phrase consists of **punctuation marks** only: the translation should also contain only punctuation marks
- If a phrase contains **quotation marks**: the translation should also contain quotation marks
- If a phrase **starts or ends** with a special punctuation mark: the translation should also start or end with it
- **Length ratio** between source and target phrase should be within a certain range
- For each **placeholder** type (NUM etc.): source and target phrase should contain equal number of placeholders

Note that phrase table entries are normalized, tokenized & preprocessed

eBay Best Practices

Filtering out-of-domain data –
Invitation model

Invitation Model

Hoang, C., & Sima'an, K. (2014). Latent domain translation models in mix-of-domains haystack. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers (pp. 1928-1939)

Core idea: differentiate out-of-domain sentences (and implicitly, a lot of corrupt entries) from in-domain sentences

$$P(D | \mathbf{f}, \mathbf{e}) = \frac{P(\mathbf{f}, \mathbf{e}, D)}{\sum_{D \in \{D_1, D_0\}} P(\mathbf{f}, \mathbf{e}, D)}$$
$$P(\mathbf{f}, \mathbf{e}, D) = \frac{1}{2} \times P(D) \times \{P_{lm}(\mathbf{e} | D)P_t(\mathbf{f} | \mathbf{e}, D) + P_{lm}(\mathbf{f} | D)P_t(\mathbf{e} | \mathbf{f}, D)\}$$

eBay Best Practices

Filtering out-of-domain data –
IBM-1

IBM-1-based filtering

1. Train a translation model („IBM-1“) on bilingual data
2. For each sentence pair in corpus, compute translation probabilities in both directions (source-to-target and target-to-source)
3. Sort all sentence pairs by this probability
4. Use top n% of sentence pairs

eBay Best Practices

Filtering out-of-domain data –
PER

Edit distance

(aka Levenshtein distance, WER=word error rate)

- Measure dissimilarity between 2 strings
- Minimum number of operations required to transform one string into the other: substitution, deletion, insertion
- Takes word order into account

Position-independent error rate (PER)

- Measure dissimilarity between 2 strings
- Ignore word order, compare bag of words
- Difference in the count of the words between the 2 strings

=> Both are normalized by the length of the reference string

eBay Best Practices

Filtering out-of-domain data –
PER

PER-based filtering

1. Train MT system (or use existing one)
2. Translate out-of-domain corpus
3. For each sentence pair in corpus, compute PER between MT and existing target
4. Sort all sentence pairs by this distance
5. Use top n% of sentence pairs

eBay Best Practices

Filtering out-of-domain data –
LangID

Language-ID-based filtering

1. For each sentence pair in out-of-domain corpus, automatically determine language of source and target sentence
2. Keep sentence pair if
 - a. Source language correct or “unknown”
 - b. Target language correct or “unknown”
3. Remove all other sentence pairs

Language identification using the CLD2 tool.

Note that CLD2 is designed for longer sequences (~200 characters)

➤ does not work well on short sequences, search queries, lists of named entities, etc.

<https://github.com/CLD2Owners/cld2>

Experimental Results

Phrase-based SMT

English-Spanish Description Translation

- In-domain data: item titles, item descriptions, search queries, brand lists
- Out-of-domain data: EPPS, EU bookshop, etc.
- Filtering out-of-domain based on IBM-1 translation model

| Results on description test data | BLEU | TER | PER |
|-----------------------------------|------|------|------|
| baseline | 39.1 | 43.4 | 34.8 |
| + filtering of out-of-domain data | 40.0 | 42.2 | 34.0 |

Experimental Results

Phrase-based SMT

German-English and English-German

- In-domain data: item titles, item descriptions, search queries, brand lists
- Out-of-domain data: UN, TAUS, News, etc.

English-German: Translation of eBay item titles

| | BLEU | PER |
|---|------|------|
| baseline | 38.8 | 41.9 |
| + PER-filtering of out-of-domain data | 39.4 | 41.3 |
| + language-ID-filtering of out-of-domain data | 39.3 | 41.8 |

German-English: Translation of user search queries on eBay

| | BLEU | PER |
|---|------|------|
| baseline | 56.4 | 26.5 |
| + out-of-domain data, one corpus PER-filtered | 56.1 | 26.6 |
| ++ PER-filtering all out-of-domain data | 55.0 | 26.8 |

Experimental Results

NMT

English-Chinese Title and Description Translation

- In-domain data: item titles, item descriptions, search queries
- Out-of-domain data: UN, icecat, in-house collections, ...
- Development data: in-domain data used for optimizing the MT system
- RNN-based architecture works best for sentence lengths observed in training
- Filtering:
 1. remove all sentences from training which are shorter than shortest development-data sentence
 2. 1 + remove all sentences which are longer than longest development-data sentence

| | BLEU on descriptions | BLEU on titles |
|---|----------------------|----------------|
| Train on in-domain + all out-of-domain data | 31.1 | 24.1 |
| + remove shorter | 32.1 | 25.2 |
| ++ remove longer | 32.6 | 25.7 |

Experimental Results

SMT vs. NMT

From literature / experience:

- SMT can handle up to 10% noise in training data
 - S. Khadivi, H. Ney, "[Automatic Filtering of Bilingual Corpora for Statistical Machine Translation](#)", 10th International Conference on Application of Natural Language to Information Systems, 2005
- NMT more sensitive to noise / domain / amount of data
 - P. Koehn, R. Knowles, "[Six challenges for NMT](#)", First Workshop on NMT, 2017
 - B. Chen, R. Kuhn, G. Foster, C. Cherry and F. Huang, "[Bilingual Methods for Adaptive Training Data Selection for Machine Translation](#)", AMTA, 2016
 - ongoing research

**Thank you
to our colleagues
Shahram Khadivi,
Michael Kozielski,
Gregor Leusch, Shen
Yan**

ebay

Thank You

ebay

Quality of Data: a Localization viewpoint



eBay Localization Practices

Things to keep in mind

- What can you do that:
 - Requires knowledge of a language
 - Does not require that knowledge?
- What can you do without programming?
 - What should you know on data and tools?

Finding issues

Localization finds issues

- Usually on the content being translated
 - Use computer-assisted QA tools
- What about the TMs?
- And what about the corpora for MT?
- What if we find issues with tools on those?

1. Linguistic checks

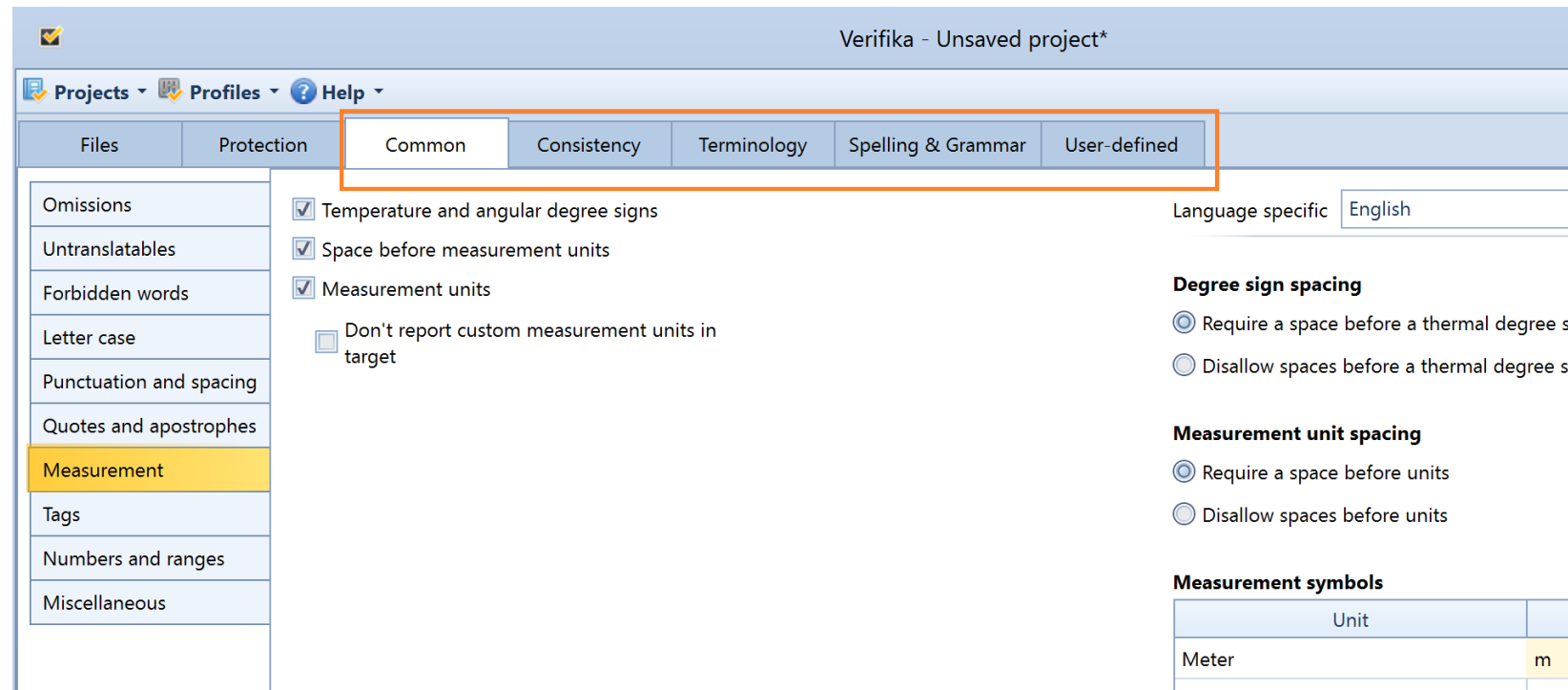
Finding linguistic issues in corpora/TM

- Using a QA tool to:
 - Find issues
 - Explore corpora
- Terminology
- Inconsistency
- Spelling
- Other issues
- Bilingual search

1. Linguistic checks

Finding linguistic issues in corpora/TM

- This is a QA tool



1. Linguistic checks

Finding linguistic issues in corpora/TM

- This is a report

| Terminology Errors (768) | | | | | | | |
|--------------------------|-----------------------------|------------------------------|--------------|------------------|---|--|----|
| Language | Portuguese (Brazil) | | | | | | |
| Error # | File | Glossary | Source term | Target term | Source | Target | Pr |
| 208 | bilino-for-vfk.xlsx (10402) | Glossary-without-debate.xlsx | human rights | direitos humanos | According to these conclusions, the Charter must contain three categories of rights. The first category is rights of freedom and equality and procedural rights, as guaranteed by the European Convention for the Protection of Human Rights and Fundamental Freedoms. | Os Chefes de Estado e de Governo estabeleceram, nas conclusões do Conselho Europeu de Colónia, as grandes linhas do conteúdo da Carta. | |
| 221 | | | human rights | direitos humanos | This House therefore proposes the adoption of urgent measures encompassing all the relevant factors and to be applied immediately. We must ensure respect for human rights, tolerance and multiculturalism and we must obtain an increase in budgets and action lines for social and health issues. | Não esqueçamos, a propósito, que o agressor da vítima era um doente mental não devidamente tratado. | |
| 290 | bilino-for-vfk.xlsx (15579) | Glossary-without-debate.xlsx | | | | | |

Common Errors Consistency errors **Terminology Errors** Spelling Errors-PTBR Spelling Errors

1. Linguistic checks

Finding linguistic issues in corpora/TM

- Using a QA tool
- 1. Terminology
 - How to find terminology?
 - QA findings

1. Linguistic checks

Finding linguistic issues in corpora/TM

- 1. Terminology
 - How to find terminology?
 - By Frequency
 - Use a tool (AntConc, Okapi)

1. Linguistic checks

- Find words (1-grams)

AntConc 3.4.4w (Windows) 2014

File Global Settings Tool Preferences Help

Concordance Concordance Plot File View Clusters/N-Grams Collocates Word List Keyword List

Word types: 16288 Word Tokens: 357129 Search Hits: 0

| Rank | Freq | Word | Lemma | Word Form(s) |
|------|------|------------|-------|--------------|
| 1 | 3941 | european | | |
| 2 | 3196 | commission | | |
| 3 | 2168 | president | | |
| 4 | 2044 | union | | |
| 5 | 1853 | parliament | | |
| 6 | 1579 | rights | | |
| 7 | 1559 | states | | |
| 8 | 1546 | member | | |
| 9 | 1448 | council | | |
| 10 | 1410 | europe | | |
| 11 | 1256 | policy | | |
| 12 | 1188 | countries | | |
| 13 | 1078 | people | | |
| 14 | 1041 | social | | |
| 15 | 915 | human | | |

Search Term Words Case Regex Advanced

Hit Location Search Only 0

Lemma List Loaded

Sort by Invert Order Sort by Freq

Total No. 1

Files Processed

Clone Results

1. Linguistic checks

- Find multiple words (2-3 grams)

| | |
|------|------------------------|
| 1368 | european union |
| 1165 | member states |
| 724 | human rights |
| 618 | european parliament |
| 312 | madam president |
| 257 | fundamental rights |
| 213 | member state |
| 212 | ladies and gentlemen |
| 185 | president commissioner |
| 167 | common position |

1. Linguistic checks

- Find the translations (bilingual search)

Search

■ Source and target texts are found

european union

União Europeia

Replace with

| Source segment | Target segment |
|--|---|
| as a number of Members have requested, on behalf of all the victims concerned, particularly those of the terrible storms, in the various countries of the European Union . | número de colegas - que observássemos um minuto de silêncio por todas as vítimas, nomeadamente das tempestades, nos diferentes países da União Europeia que foram afectados. |
| Thirdly, this directive, as it currently stands in the common position, guarantees - in particular because it confines itself exclusively to minimum standards - a high degree of flexibility and modest regulation by the European Union ; by adopting it we contribute to the Member States' bearing a high level of individual responsibility. | Terceiro: com a Directiva, tal como ela hoje se apresenta enquanto posição comum, em especial por se restringir exclusivamente às exigências mínimas, garantimos um elevado grau de flexibilidade e uma regulamentação reduzida por parte da União Europeia , contribuindo para um elevado grau de auto-responsabilização dos Estados-Membros. |
| My Group will therefore support the common position and looks forward to the enactment of the legislation which will provide us with yet another tool in our fight to make transport in the European Union as safe as possible. | Assim, o Grupo PSE dará o seu apoio à posição comum, ficando a aguardar com expectativa a aprovação desta legislação, que nos irá dotar de um novo instrumento na nossa luta para tornar tão seguros quanto possível os meios de transporte na União Europeia . |

Finding linguistic issues in corpora/TM

1. Linguistic checks

- 1. Terminology
 - QA findings
 - Terminology inconsistency
 - Human rights = “direitos humanos” 60%
 - Human rights = “direitos do Homem” 40%

| Source term | Target term | Source | Target |
|--------------|-------------------|---|---|
| human rights | direitos do homem | Subject: Implications for human rights of the construction of the Ilisu Dam in Turkey Given that Turkey has recently been granted applicant country status, what is the Commission's view of the implications for human rights of the construction of the Ilisu Dam, given the massive displacement of the Kurdish and the other people of the region that will result? | Objecto: Implicações para os direitos humanos da construção da barragem de Ilisu, na Turquia Visto que foi recentemente concedido à Turquia o estatuto de país candidato, o que pensa a Comissão das implicações para os direitos humanos da construção da barragem de Ilisu, tendo em conta o deslocação maciça de Curdos e de outras populações da região que tal irá provocar? |

| Source term | Target term | Source | Target |
|--------------|------------------|---|--|
| human rights | direitos humanos | Are you not afraid that such penalties might get out of hand and serve to punish not violations of human rights but simple differences of opinion, unpopular opinions or deviations from the dominant European thinking? | Não teme que possa haver desvios neste tipo de sanções e que estas sirvam para punir, não as violações dos direitos do Homem , mas sim simples divergências, delitos de opinião ou divergências em relação ao pensamento dominante europeu? |

Finding linguistic issues in corpora/TM

1. Linguistic checks

- 1. Terminology
 - QA findings
 - Terminology errors
 - Human rights untranslated

| | | | |
|--------------|------------------|--|---|
| human rights | direitos humanos | Your approach, which I am very much in favour of, since it is a coherent one, takes its inspiration from the issue of human rights ; I wholeheartedly agree that we must embrace the cause of human rights . | O princípio que o senhor Comissário refere é coerente e inspira-se no tema dos direitos humanos ; dou todo o meu apoio a que se aprofunde o tema human rights . |
|--------------|------------------|--|---|

1. Linguistic checks

Finding linguistic issues in corpora/TM

- 1. Terminology
 - QA findings
 - Misalignments

These are not translations of the source.

| Source term | Target term | Source | Target |
|--------------|------------------|---|---|
| human rights | direitos humanos | According to these conclusions, the Charter must contain three categories of rights. The first category is rights of freedom and equality and procedural rights, as guaranteed by the European Convention for the Protection of Human Rights and Fundamental Freedoms. | Os Chefes de Estado e de Governo estabeleceram, nas conclusões do Conselho Europeu de Colónia, as grandes linhas do conteúdo da Carta. |
| human rights | direitos humanos | This House therefore proposes the adoption of urgent measures encompassing all the relevant factors and to be applied immediately. We must ensure respect for human rights , tolerance and multiculturalism and we must obtain an increase in budgets and action lines for social and health issues. | Não esqueçamos, a propósito, que o agressor da vítima era um doente mental não devidamente tratado. |
| human rights | direitos humanos | They fly in the face of the very principles on which - as Mr Havel said yesterday - the European Union was founded: principles of liberty, democracy and respect for human rights . | Claro que nestes últimos anos se fizeram progressos, mas temos de continuar a esforçar-nos juntos por criar um clima de tolerância, em que o racismo e a xenofobia sejam considerados totalmente reprováveis e inaceitáveis, ao mesmo tempo que tratamos com severidade incidentes como aqueles de que estamos a falar aqui esta tarde. |

1. Linguistic checks

Finding linguistic issues in corpora/TM

- Using a QA tool
- 2. Inconsistency
 - Does it matter for MT?

1. Linguistic checks

Finding linguistic issues in corpora/TM

- 2. Inconsistency
 - Does it matter for MT?

| Source | Target |
|-------------------------|---|
| Are there any comments? | Há alguma observação? |
| Are there any comments? | Há alguma observação sobre as actas? |

- Also finds misalignments

| Source | Target |
|--------------|---|
| Human rights | Senhor Presidente, antes de passarmos ao debate sobre o Kosovo ou sobre Mitrovica, devo confessar que se apodera de mim um sentimento de amargura. |
| Human rights | Direitos do Homem |

1. Linguistic checks

Finding linguistic issues in corpora/TM

- 3. Spelling
- Maybe find different locales in the corpora? PTBR vs. PTPT?
 - Ação vs. Acção
 - (Also, MT fixes itself)

1. Linguistic checks

Finding linguistic issues in corpora/TM

- 4. Others
 - Length difference – finds misalignments

| Source | Target |
|---|--|
| EXPLANATIONS OF VOTE | Senhora Presidente, votei a favor do relatório Schmidt sobre as disposições respeitantes ao investimento em valores mobiliários porque, como declarei esta manhã no hemiciclo, considero muito importante que seja bem utilizado o dinheiro dos cidadãos da União Europeia, que são, em grande parte, cidadãos idosos e reformados, que desejam viver serenamente a sua reforma e que, após as dificuldades encontradas ao longo da sua vida de trabalho, conseguiram, |
| Brok report (A5-0029/2000) | Senhora Presidente, nos últimos tempos falou-se muito da importância do alargamento da União Europeia. |
| Both countries also qualify for financial aid under the MEDA programme (B7-4012). | Durante as primeira e segunda leituras do orçamento para 2000, o Parlamento Europeu salientou que o novo título "pré-adesão" (B7-0) proposto pela Comissão no anteprojecto de orçamento não deveria restringir-se aos países associados da Europa Central e Oriental, deveria igualmente ser alargado a Malta e a Chipre. |
| In Zambia 25% of teachers have died of AIDS. | Nalguns países, há vinte anos, as crianças tinham mais probabilidades de acesso ao ensino básico e aos cuidados de saúde do que as que têm hoje. |
| What should this global actor' s objective be? | O cidadão esperaria sem dúvida uma resposta do tipo: "defender melhor os países da Europa" , uma vez que essa é a missão tradicional e primordial da associação política. |

1. Linguistic checks

Finding linguistic issues in corpora/TM

- 4. Others
 - Number mismatches
 - Can be better than digits to digits

Language specific English

Digit to text

| Digit | |
|-------|-------|
| 4 | four |
| 5 | five |
| 6 | six |
| 7 | seven |
| 8 | eight |
| 9 | nine |
| 10 | ten |

1. Linguistic checks

Finding linguistic issues in corpora/TM

- 4. Others
 - Number mismatches
 - Finds misalignments

This is why I personally have an idea to supplement Article 7 of the EU Treaty with a new paragraph which would give the democratic warning or monitoring procedure a legal basis. Finally, this is why we have public debate.

A única maneira de fazer recuar as más ideias é exprimirmos outras.

- Can find errors (few)

This is also expressed in Recital 14.

Esse aspecto também é referido no considerando 41.

1. Linguistic checks

Misalignment

Translating the corpora and Comparing to itself
(an idea from the Science team)

- Several of the previous linguistic checks find misalignments
- They are based on knowing the language
- They find individual segments, not blocks of them

Is there another way?

1. Linguistic checks

Translating the corpora and Comparing to itself

EuroParl data sample

- 28,874 segments – 746,866 words
- 1,258 segments misaligned – 32,519
 - Found manually, looking through the data
 - With language knowledge
 - 4.4% of words

1. Linguistic checks

Translating the corpora and Comparing to itself

Last aligned and first misaligned segments

| | | |
|--|---|------------------|
| I see that he has left the House. We shall therefore go on to the explanations of vote. | Constato que abandonou o hemiciclo, pelo que passamos de imediato às declarações de voto. | last aligned |
| EXPLANATIONS OF VOTE | Senhora Presidente, votei a favor do relatório Schmidt sobre as disposições respeitantes ao investimento em valores mobiliários porque, como declarei esta manhã no hemiciclo, considero muito importante que seja bem utilizado o dinheiro dos cidadãos da União Europeia, que são, em grande parte, cidadãos idosos e reformados, que desejam viver serenamente a sua reforma e que, após as dificuldades encontradas ao longo da sua vida de trabalho, conseguiram, finalmente, a tranquilidade económica. | first misaligned |
| Madam President, I voted in favour of the Schmidt report on regulation of investments in transferable securities because, as I stated in this Chamber this morning, I consider it very important that the monies of the citizens of the European Union be put to good use. The vast majority of these citizens are elderly people and pensioners, who want a peaceful retirement and who, after the problems encountered during the course of their working lives, have at last found financial peace of mind. | As disposições em matéria de investimento mobiliário, em grande parte efectuado por pessoas idosas reformadas, são, por conseguinte, um grande progresso para a União Europeia. | |
| The regulation of investments in transferable securities, the majority of which are made by elderly pensioners, is thus a major step forward for the European Union. | Relatório Olle Schmidt (A5-0025/2000) | |
| Report by Olle Schmidt (A5-0025/2000) | . | |

1. Linguistic checks

Edit Distance

Amount of change to get from one sentence to another

- It can be counted in characters or words
- A case for words:
 - if you change a word out of 10, is it 10%? Yes
 - if you change a longer word, is it more change? No
- A case for characters
 - if you add an "s" to make a plural, did you change a whole word? No

1. Linguistic checks

Edit Distance

Amount of change between two texts

- Edit Distance is not normalized/proportional
 - it is the number of words or characters changed
- Dividing by the total number of words will make it proportional, a % of the total "amount" (words or characters)
- This creates a "% of change"
- This takes into account the word order.

- There are several metrics, but most are based on a % of change

- TER - is the % of change in words
- PER - Position-independent error rate - independent of word order

1. Linguistic checks

Edit Distance

Amount of change between two texts

- Most importantly, any way you measure will show what is next

1. Linguistic checks

Edit Distance

Experiment

- Sample
 - A number of segments that appear in the corpora before the misaligned ones (333)
 - All the misaligned segments (1258)
 - A number of segments that appear in the corpora following the misaligned ones (200)
- Source: EN
- Reference Target: EuroParl existing translation for PT
- 2nd Target: Obtained the MT for all segments
- Calculated the Edit Distance between Reference target and 2nd target
- Calculated the % of change

1. Linguistic checks

Edit Distance

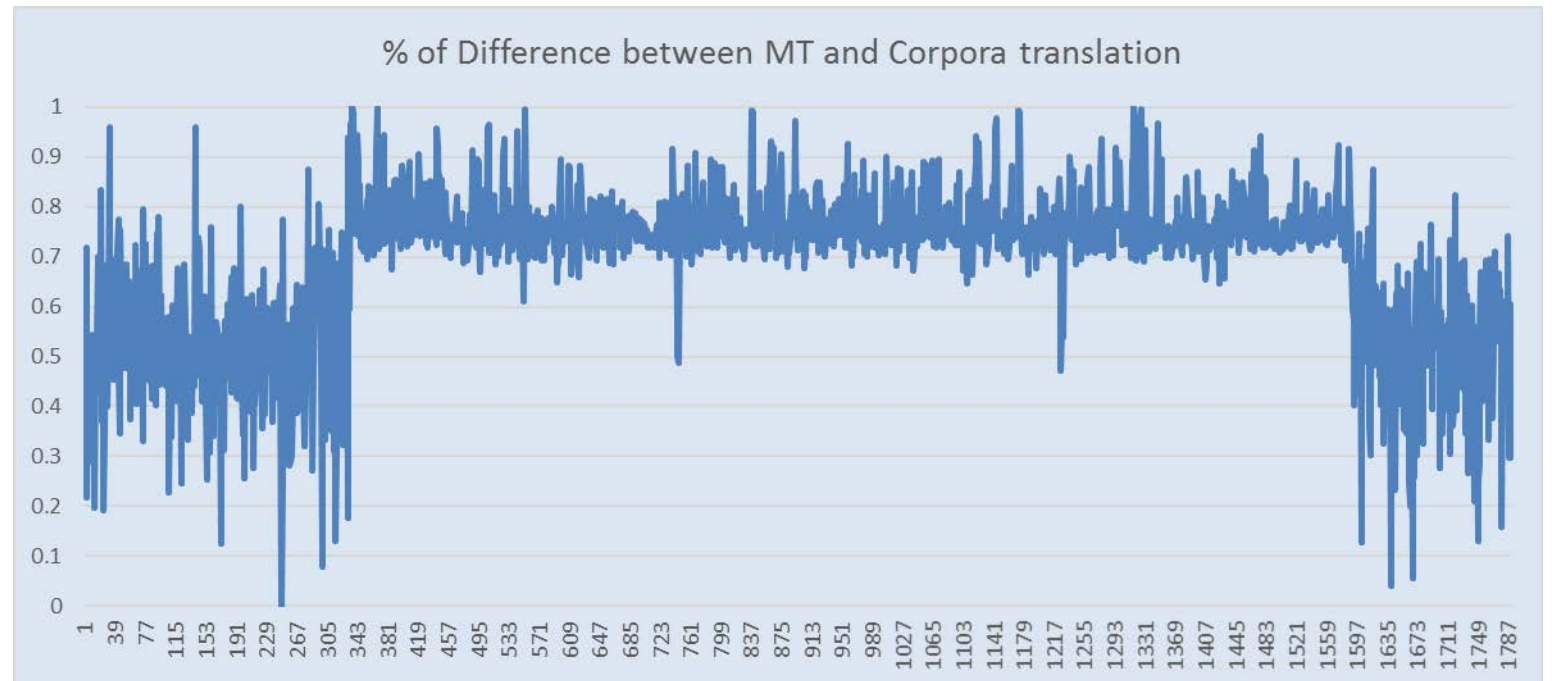
Experiment

- The expectation is that misaligned segments will not have anything to do with the machine translation, it will be very different
- The **% of change** (edit distance) will be **very high**
- Aligned segments will have an existing reference translation and a machine translation – different but similar
- So, **lower % of change** (edit distance) is expected
- The chart on next slide shows the % of change throughout the data

1. Linguistic checks

Misalignment

Translating the corpora and Comparing to itself



1. Linguistic checks

Translating the corpora and Comparing to itself

- The finding of misalignments becomes language-independent
- There is a decision to make:
 - fix the misalignment or delete the segments from corpora
 - This likely requires language knowledge
- Whatever your decision, your corpora will be better than before.

1. Linguistic checks

- Bilingual Search/Regexes (demo)

Search Files Pending changes


■ ■ Source and target texts are found

[A-Z][A-Z][A-Z]+

Search in target

Replace with

Find Add as User-defined check

|  | Source segment | |
|---|--|--------------------------|
| | The questions answered previously referred to Mrs de Palacio' s intervention, on another occasion, and not to these comments which appeared in the ABC newspaper on 18 November. | As ref det jori |
| | (Applause from the PSE Group) | (Ap |
| | (Applause from the PPE-DE Group) | (Ap |
| | The PPE-DE Group is requesting that this item be taken off the agenda. | O C que |

2. Engineering/ Preprocessing checks

- Normalize escaped characters/entities
 - Also for TMs

Escaped characters are representations of characters using only ASCII characters. For example, `€` is the escaped representation of the Euro symbol.

- Corpora ideally should not have escaped characters

2. Engineering/ Preprocessing checks

- Remove formatting tags for MT
 - But not for TMs
- Formatting tags
 - `<seg>`When you click `<ph x="1" type="x-LPH"/>`Call me`<ph x="2" type="x-LPH"/>`, we'll call you at a phone number you'll specify.`</seg>`
- Placeholders have meaning
 - `<seg>`You declined `<ph x="1" type="x-LPH"/>`'s counteroffer of `<ph x="2" type="x-LPH"/>`.`</seg>`
- The problem: Some systems will not distinguish between them (as above) – TM Mgmt Taskforce
- The solution: space vs. no space

Thank You

ebay

3. Creating high-quality in-domain content via post-editing

- How Science selects data
- How we post-edit
- How Bias affects the quality
- Terminology for MT

3. Creating high-quality in-domain content via post-editing

- How we post-edit

Post-edited by Vendor

Guidelines for PE

Focus on meaning over fluency and style (titles)

Some rules from Science (Book titles)

QA tool checks

Standard checks

Regular expressions

First screening to find massive quality issues

Terminology coming

Sample Review

Regular intervals

5-10% of content

Brands list

Acronyms list

eBay glossary

3. Creating high-quality in-domain content via post-editing

- How Bias affects the quality
- Are errors made by MT corrected in PE?
- 15k words studied
- Edit Distances
 - between MT and PE
 - Between PE and Gold

3. Creating high-quality in-domain content via post-editing

- How Bias affects the quality
- Are errors made by MT corrected in PE?
- Yes, Post-editors accept MT suggestions more than they should
- 18 errors per 1k words

Patterns of issues



Veteran musicians, DJs, and public speakers
are taken aback by the sensitivity and reliability of the Samson QMIC.

01

Multiple Modifiers or Words

A modifier may apply to one or several words, often adjectives and nouns.

Finding patterns systematically

Could we systematically find a pattern?

Looked into Multiple Modifiers or Words errors

- Applied POS tags
- Created a list of tokens and tags
- Applied formulas to find patterns

| Token | Tag | Simplified Tag |
|------------|--|----------------|
| the | DT,B-NP-plural | DT |
| enhanced | JJ,enhance/VBD,enhance/VBN,I-NP-plural | JJ |
| telephony | NN:U,I-NP-plural | NN |
| capability | NNS,E-NP-plural | NN |

Found 32% of the issues with two formulas, 90% of this type of issue is findable in a systematic way

Conclusion: yes, there is potential to find patterns of issues in a computer-assisted way

3. Creating high-quality in-domain content via post-editing

- Terminology for MT
- MT for localization has glossaries and content is narrow
- MT for UGC is extremely varied
- How do you harvest terms?
 - Category
 - Frequency
 - Popularity inside vs. Outside
 - Polysemous exceptions

3. Creating high-quality in-domain content via post-editing

- Terminology for MT
- Process in progress for:
 - Extracting candidates
 - Defining the translation
 - Measuring the improvement

4. Metrics

- **Experiment:**
 - **Measuring the quality of TMs**
- Ran standard checks on TMs
- Reviewed reports by types of errors:
 - Of a sample
 - For a certain amount of time
- Calculate the score:
 - Number of errors per 1k words

4. Metrics

- Results

| | | | | |
|----------|---------------------|------|-----------------------|---------|
| PTBR-TM1 | Errors per 1k words | 4.30 | Number of total words | 542847 |
| ES-TM1 | Errors per 1k words | 4.58 | Number of total words | 7800881 |
| PTBR-TM2 | Errors per 1k words | 6.94 | Number of total words | 2291248 |
| ES-TM2 | Errors per 1k words | 7.12 | Number of total words | 767397 |

- Types of errors that were efficient:
 - Terminology
 - Consistency
- **Can be done on corpora**

4. Metrics

- WIP: A new metric for MT
- Minimum human effort
- Identify challenge set
- Break down to binary bilingual task

| Is the translation of the words below correct in the Target to the right? | Answer | | |
|---|--------|--|--|
| business days | Yes | Always ship within two business days of receiving the order and make sure you specify a 0-2 day handling time. | Envía siempre el artículo dentro de los dos días laborables siguientes a la recepción del pedido y asegúrate de especificar un tiempo de manipulación de 0-2 días. |
| business days | No | To summarise, please ship the item within 10 business days or the case will be closed without a refund. | Resumiendo: envía el artículo en un plazo de 10 días hábiles o el caso se cerrará sin reembolso. |

- Offer to crowd
- Compute results without human effort (Yes/No)
- It could become automatic

4. Metrics

Where would it stand



Thank You

ebay

Discussion

What the professional looks like

- Should see “data”, beyond words
- Should see numbers: edit distance, frequency
- Should see patterns
- Should know how to use tools
 - Frequency, QA, scripts
 - Not necessarily programming
- Maybe there will be a translator professional and a “[Your language here] data analyst” professional



Training Language Professionals for a Big Data, AI Age

March, 2018

Dr. Pete Smith

Chief Analytics Officer and Professor

GILT Academic Certificate

Globalization, Internationalization, Localization, and Translation

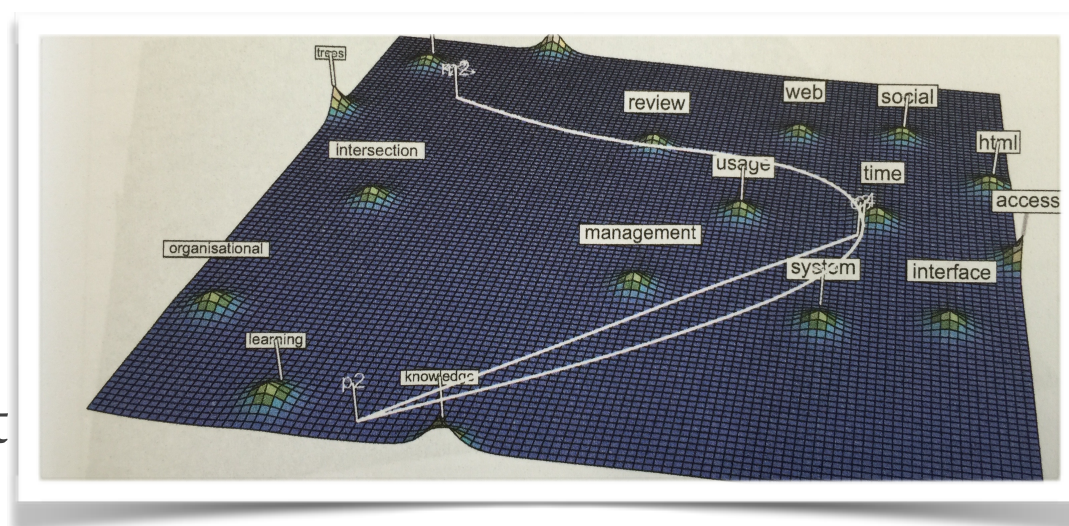
Students in UTA's nationally-known program in Localization and Translation program study localization, translation, NLP, NLU, CAT tools, and machine translation. Learners build statistical machine translation engines for world languages, including Arabic, Chinese, and Korean. Advanced research students explore natural language processing (NLP) and natural language understanding (NLU).



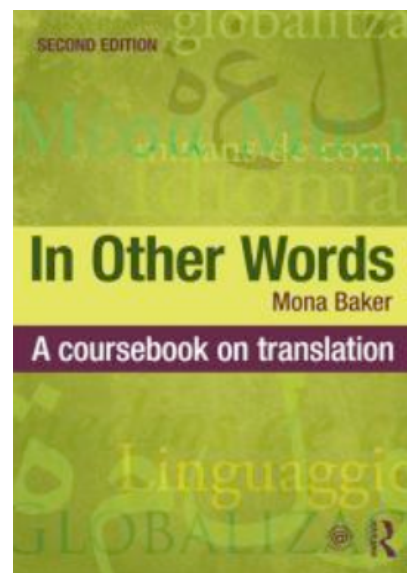
GILT Academic Certificate

Largest Undergraduate Program in GILT Nationally

Since 2007, more than 100 undergraduate learners have graduated with the “GILT Certificate.” Many of these students are pursuing “traditional” localization careers on the buyer or seller side, in roles such as project managers, localization sales, or linguistic talent as part of localization teams.



Got Global?



**GERM/RUSS 3310/4334, FREN 3320
 ARAB 3310, CHIN 3310, KORE 3310, PORT 3310
 Globalization, Internationalization, Localization,
 and Translation (GILT)
 Fall, 2017
 MWF 10 a.m.**

University of Texas—Arlington
 Dr. P. Smith, Dr. Cynthia Laborde (& MODL Teaching Faculty
 in Critical Languages)

In this course, you will systematically explore the GILT (Globalization, Internationalization, Localization, and Translation) field, such topics as: translation theory and practice; multilingual computing, localization and globalization of software and e-commerce, machine-aided translation, machine translation, international integration of content, global workforce management, global customer service, intercultural communications through 21st Century communications tools, and global virtual teams.





Teaching MT Engine Creation

KantanMT at UTA

Students in UTA's Localization and Translation seminar series utilize KantanMT to build statistical MT (SMT) engines, while exploring with Neural MT. Learners build engines in eight languages, exploring concepts of machine translation and the challenges of particular language or culture in the MT process.



Research Teams for Emerging Issues

Undergraduate students complete GILT research projects on topics such as NMT engine creation, sentiment analysis across languages and cultures, modeling language data, and issues of large-scale content archives in a multi-national world. UTA as a Tier I university encourages undergraduate research, which often disposes learners to further graduate study.



Unit 1: Custom MT, MT Engines (8 weeks)

Topic Breakdown:

- Overview of MT Theory and Practice
- Train MT Engines Utilizing *KantanMT*

Outcome Product: Five (5) Assignments and Developed MT Engine Products

Unit 2: Natural Language Processing (4 weeks)

Topic Breakdown:

- Overview of NLP Theory and Practice
- Data Collection for NLP
- Data Preprocessing
- Sentiment Analysis
- Optional: Python workshop for NLP

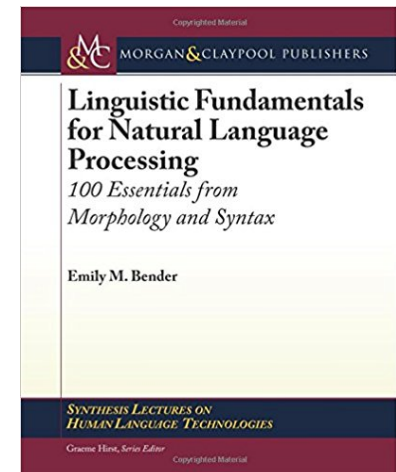
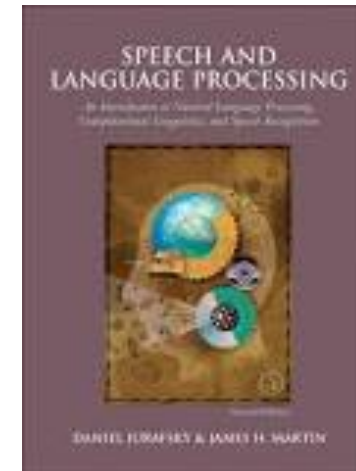
Outcome Product: Three (3) NLP Assignments

Unit 3: Natural Language Understanding (2 weeks)

Topic Breakdown:

- Overview of NLU Theory and Practice
- Semantic Resources
- Topic Modeling
- Deep Learning Applications (custom search, content recommenders, summary engines, *chatbots*, natural language generation, etc.)

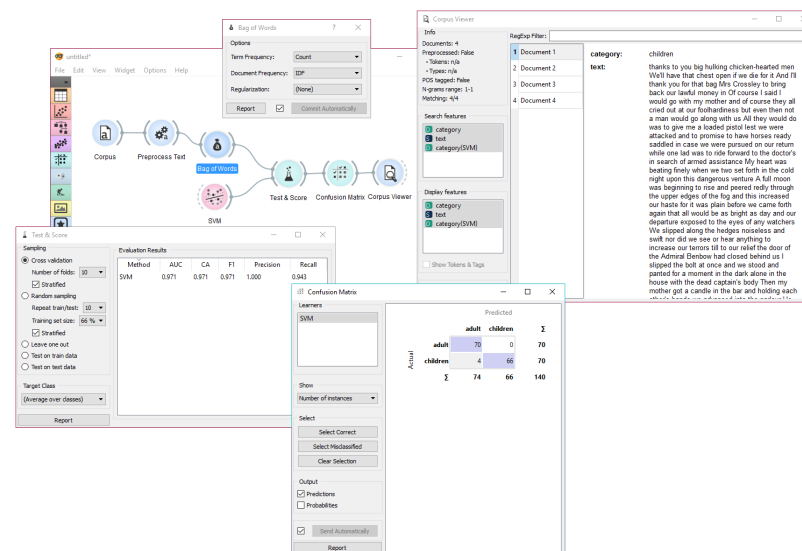
Outcome Product: Two (2) NLU Assignments



NLP

On the joys of unstructured data....

Natural language processing is a required component in the languages of study. In their focus languages, GILT students learn preprocessing, sentiment analysis, topic mining, and other NLP tasks.



NLP (con't.)

On the joys of unstructured data....

Students select to complete NLP tasks in a GUI interface (Orange) or directly in Python. Python workshops are lead by an unstructured data analyst at UTA, to foster the importance of programming in the arsenal of the future language professional.



“Language as Data” Mindset

Traditional language students prepare for MT/ML/AI markets.

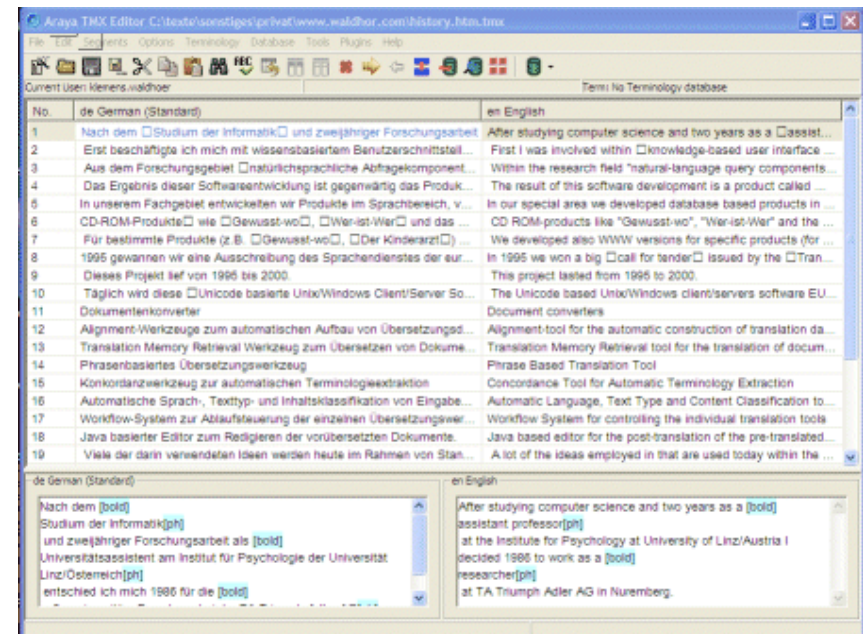
Critical to preparing students for emerging MT and ML/AI markets is a “language as data” mindset. Students in the GILT track move away from traditional academic orientations in literature and linguistics to language data collection, preparation, analysis and application.



Industry Partnerships

eBay and Bilingual Data

Colleagues at eBay, users of *KantanMT* and MT engineers, volunteered to teach data quality and data cleaning to GILT students. Via *Zoom*, MT specialists and engineers from eBay sites in California and Berlin joined the U.S.-based learners to share common problem and tools sets, as well as recent applied research.





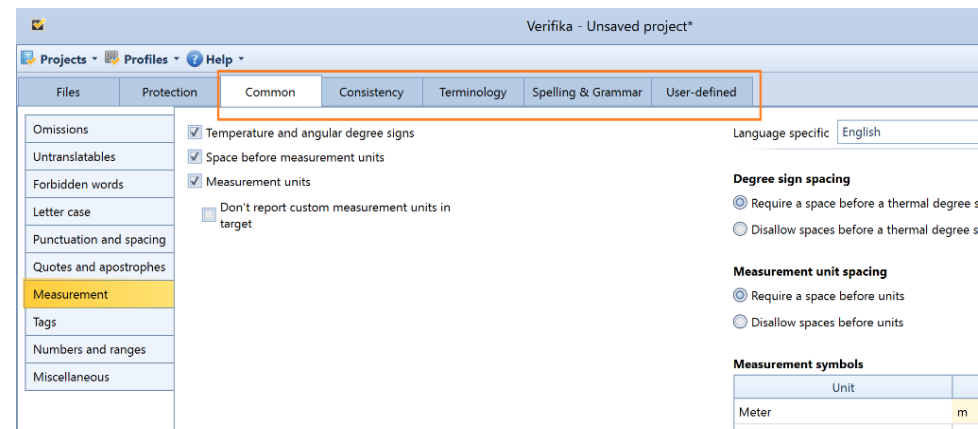




Industry Partnerships (con't.)

eBay and Bilingual Data

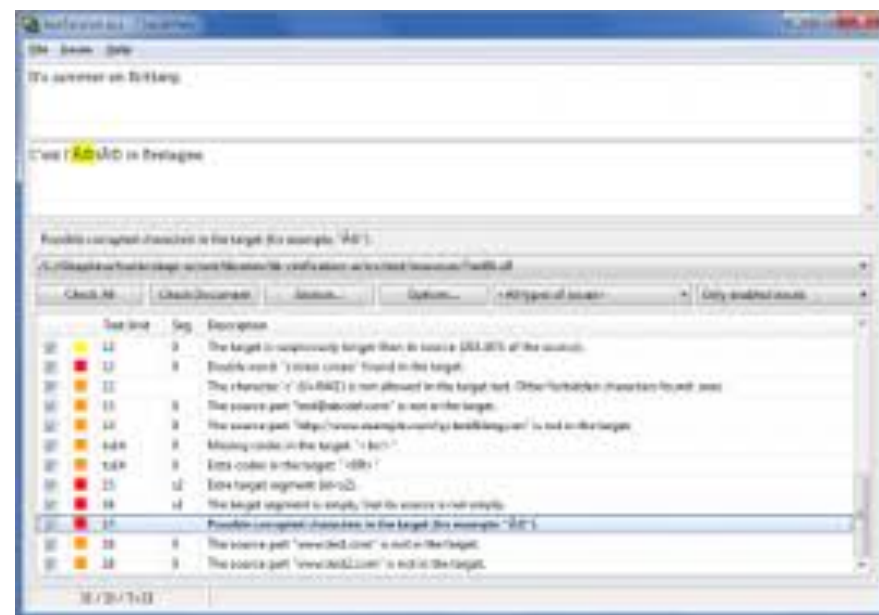
After demonstrations from eBay colleagues, UTA students are utilizing Okapi/Checkmate and Heartsome to clean bilingual data from Opus, then rebuilding *KantanMT* engines. Learners are considering QA topics: linguistic and terminology themes, inconsistency, alignment, and other issues.



Industry Partnerships (con't.)

eBay and Bilingual Data

Then UTA learners are rebuilding their *KantanMT* engines following data QA. On average, learners are seeing BLEU score improvements of XX as a result of their newfound skill and tool sets.



Most importantly, our future GILT students are also

- tied to industry colleagues in collaborative ways, sharing problem and data sets, tools and skills
- that much more prepared as future practitioners and researchers in business and industry settings.





Training Language Professionals for a Big Data, AI Age

Dr. Pete Smith

psmith@uta.edu