MT adaptation from TMs in ModernMT

Marcello Federico - FBK, Italy

AMTA, Oct 29 2016 - Austin, Texas

ModernMT Next Generation Machine Translation

Translated FONDAZIONE BRUNO KESSLER THE UNIVERSITY of EDINBURGH TAUS

Translators' pains with MT

output is often **poor** or **contextually wrong**

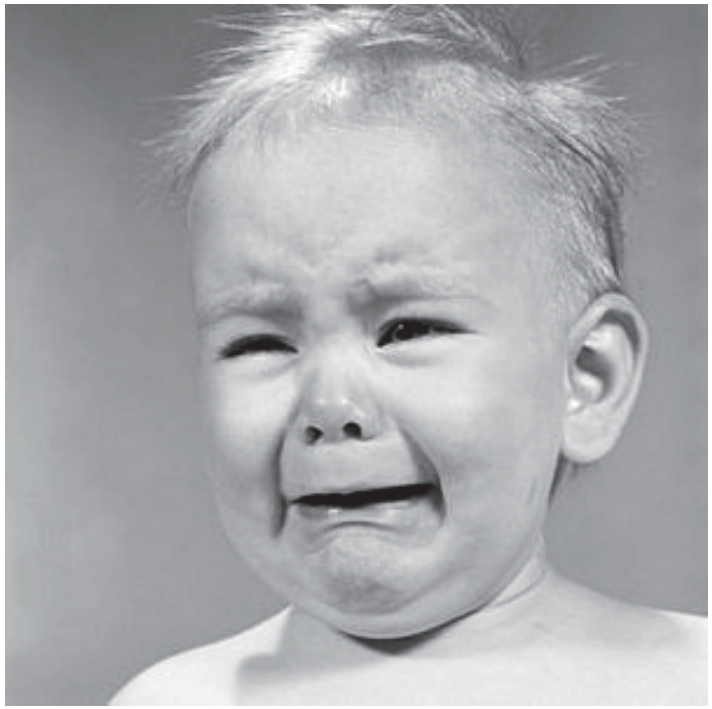
ModernMT Next Generation Machine Translation

LSP engineers don’t laugh either

cumbersome setup of MT

lack of training data

online MT is **too generic**



ModernMT Next Generation Machine Translation

The Modern MT way

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

ModernMT Next Generation Machine Translation

Modern MT in a nutshell

zero training **time**

manages **context**

learns from users

scales with data and users

ModernMT Next Generation Machine Translation

Team

Business

Research

ModernMT Next Generation Machine Translation

Roadmap

development started

first **alpha release**.

10 langs,
fast training,
context aware,
distributed

first **beta release**

45 langs,
Incremental learning

final release

enterprise ready

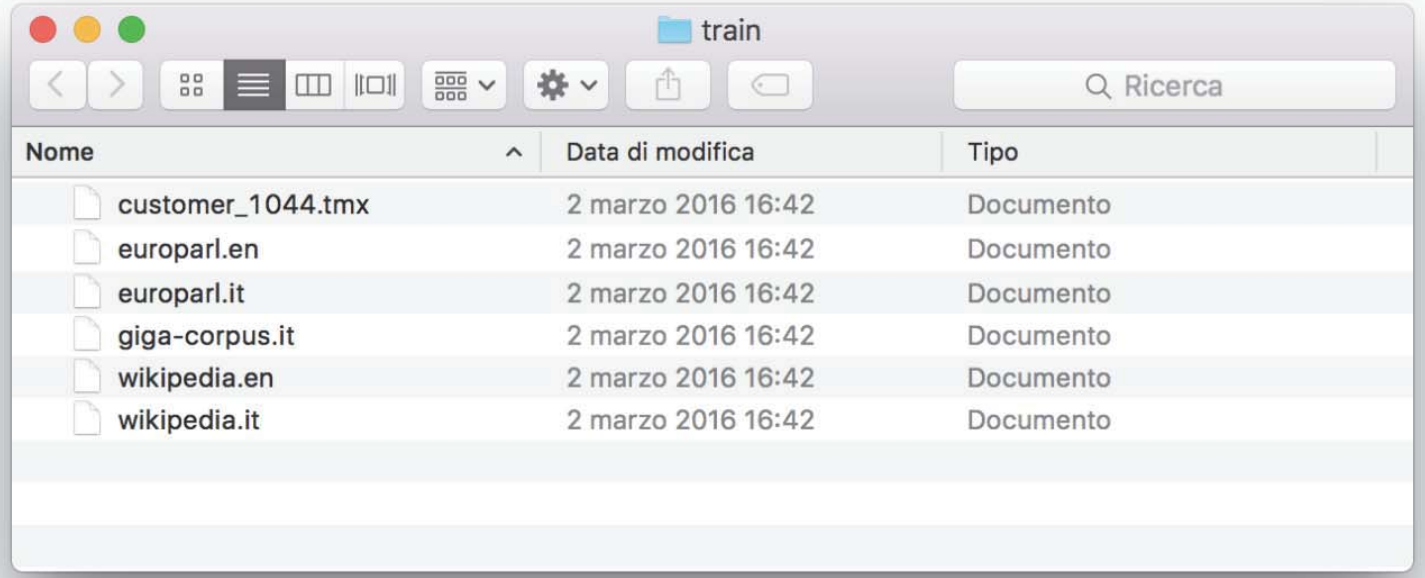
ModernMT Next Generation Machine Translation

Roadmap

2015 Q1	2016 Q2	2016 Q4	2017 Q4
development started	first **alpha release**. 2 langs, fast training, context aware, distributed	first **beta release** **plug-in** 10 langs, **incremental learning**	**final release** enterprise ready

ModernMT Next Generation Machine Translation

Prototype - Easy training




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

ModernMT Next Generation Machine Translation

Prototype (April 2016) - Fast training

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

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

ModernMT Next Generation Machine Translation

Context aware translation

TEXT 1

We're going out.
party

TRANSLATION

Nous sortons.
fête

ModernMT Next Generation Machine Translation

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

ModernMT Next Generation Machine Translation

Context aware translation

SENTENCE

party

CONTEXT

We are going out.

TRANSLATION

fête

CONTEXT

We approved the law

TRANSLATION

parti

ModernMT Next Generation Machine Translation

Context Analyzer

ModernMT Next Generation Machine Translation

REST API

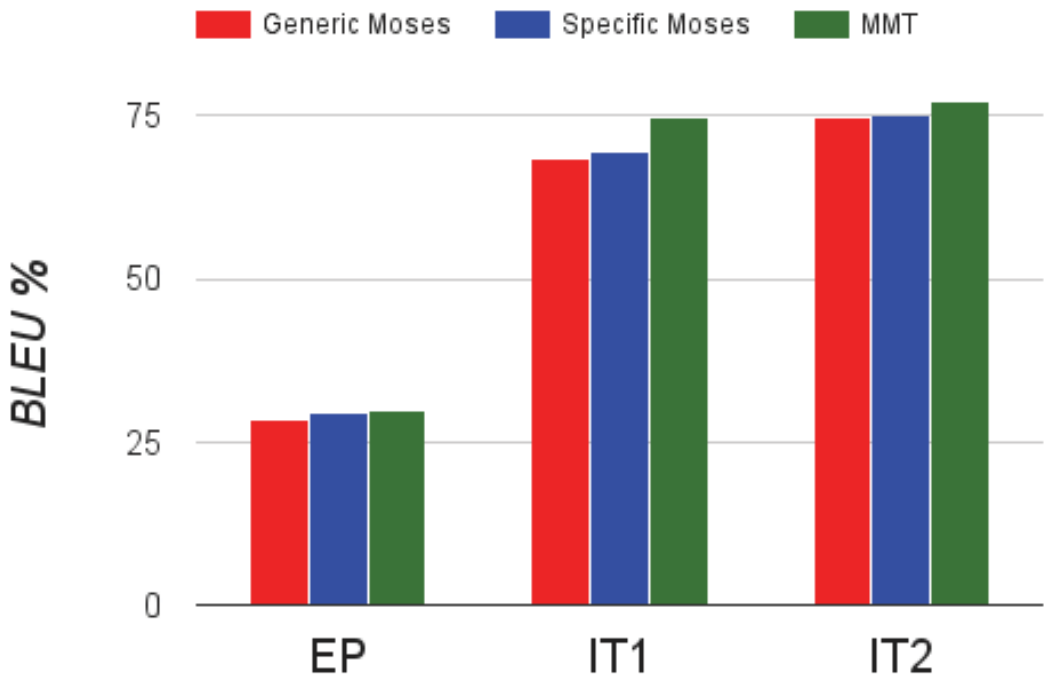
GET /translate?**q**=party&**context**=We+approved+the+law

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

```
> mmt start
```

ModernMT Next Generation Machine Translation

Minimum Viable Product (June 2015)

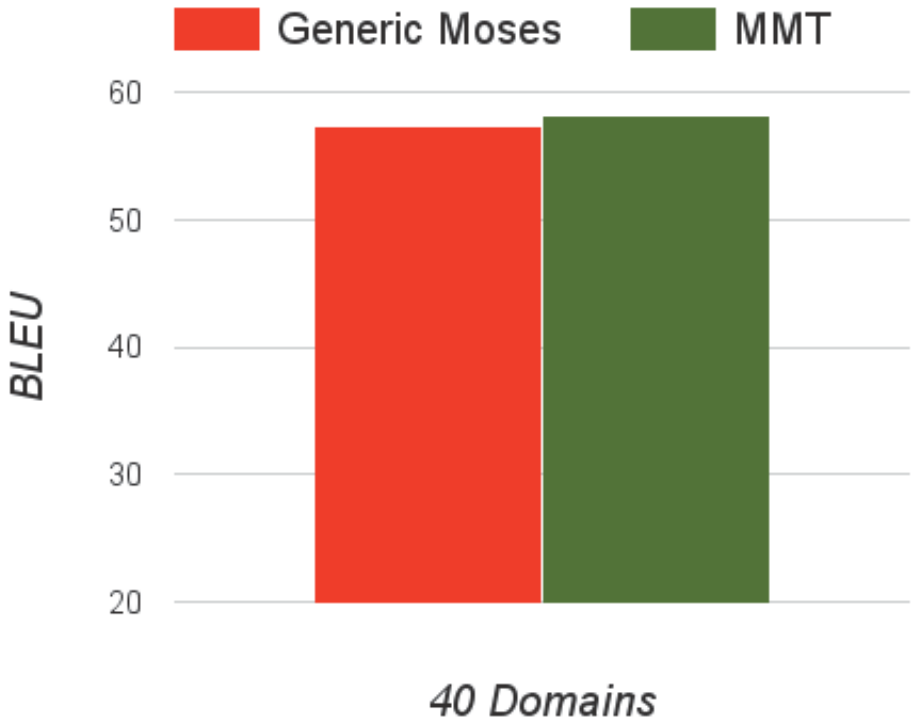


6x faster training than std Moses!

MMT outperformed **specific** and **generic** Moses
0.5 <= delta <= 5

ModernMT Next Generation Machine Translation

Prototype (January 2016)

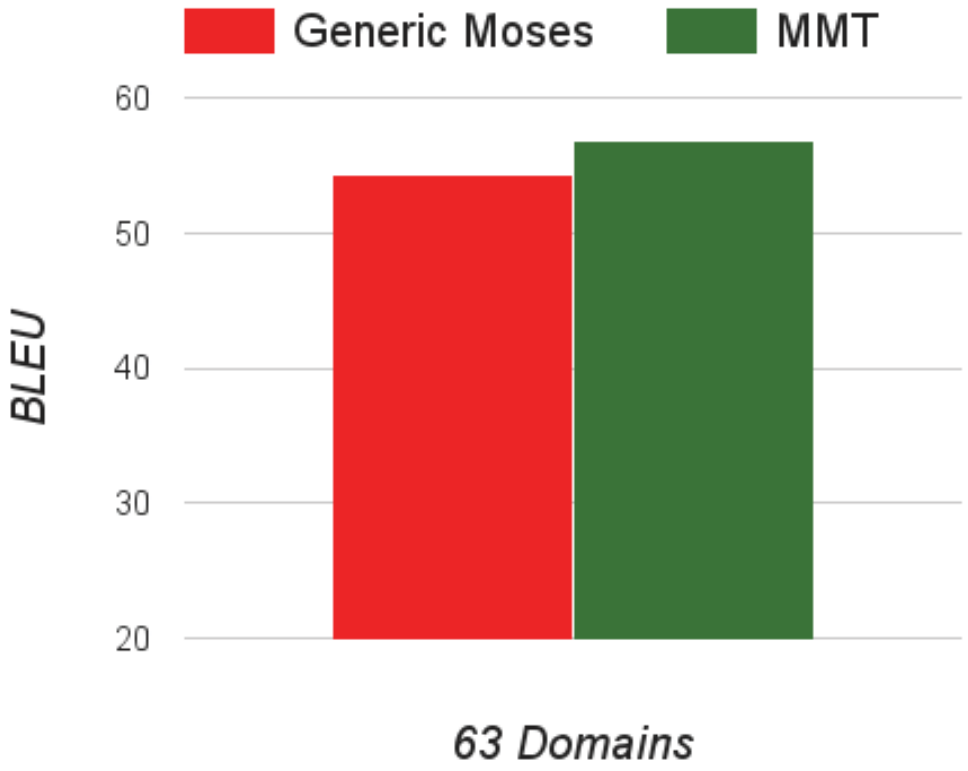


11x faster training than std Moses!

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

ModernMT Next Generation Machine Translation

Prototype (March 2016)



Benchmark 1.1

- **No tags**
- **No xml**

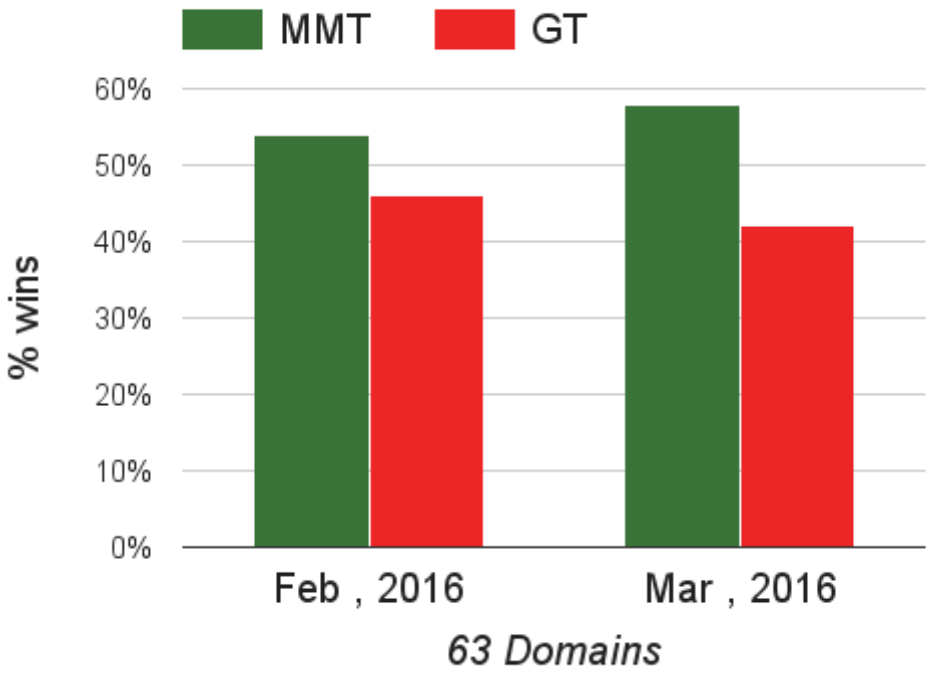
MMT outperformed **generic** Moses by >2 BLEU points

12x faster training

Speed: 1.7sec/segm

ModernMT Next Generation Machine Translation

Prototype (March 2016)



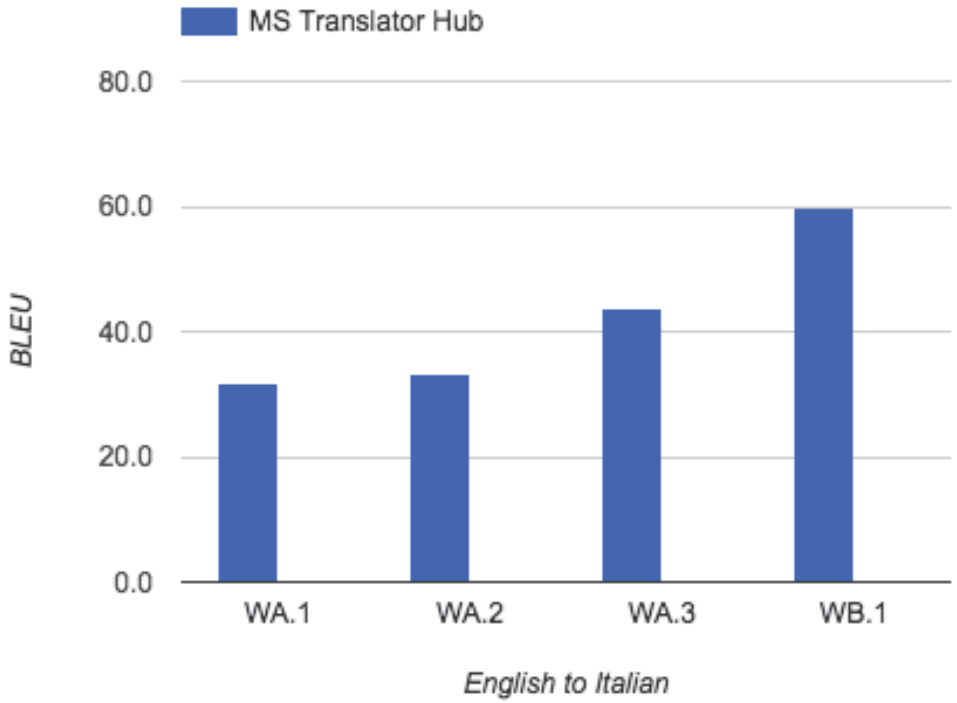
Benchmark 1.1

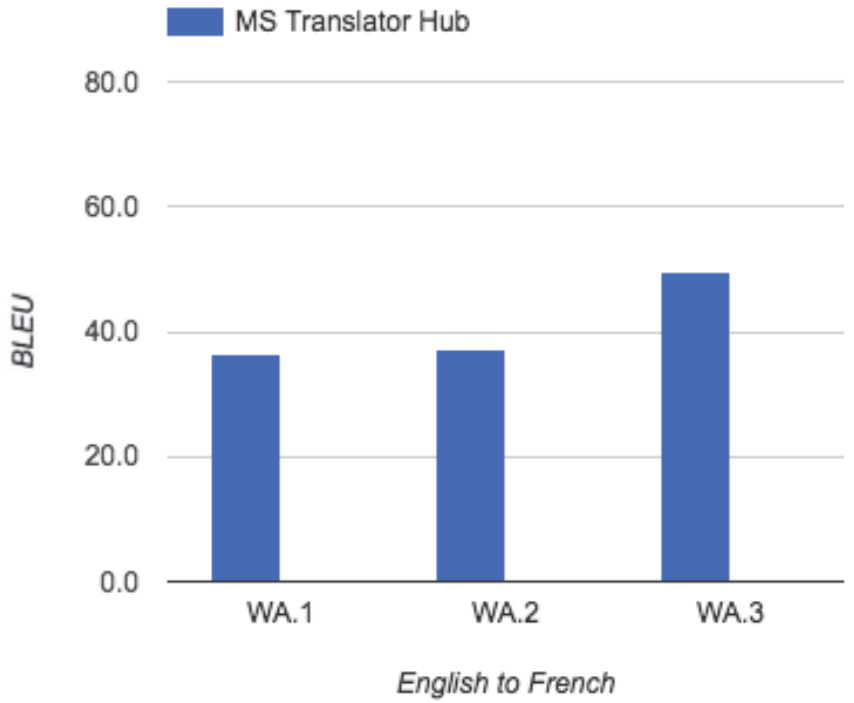
A/B testing vs GT:
~ 300 rnd segments
~ 3 judges

Distance doubled, from 8% to 16%!

ModernMT Next Generation Machine Translation

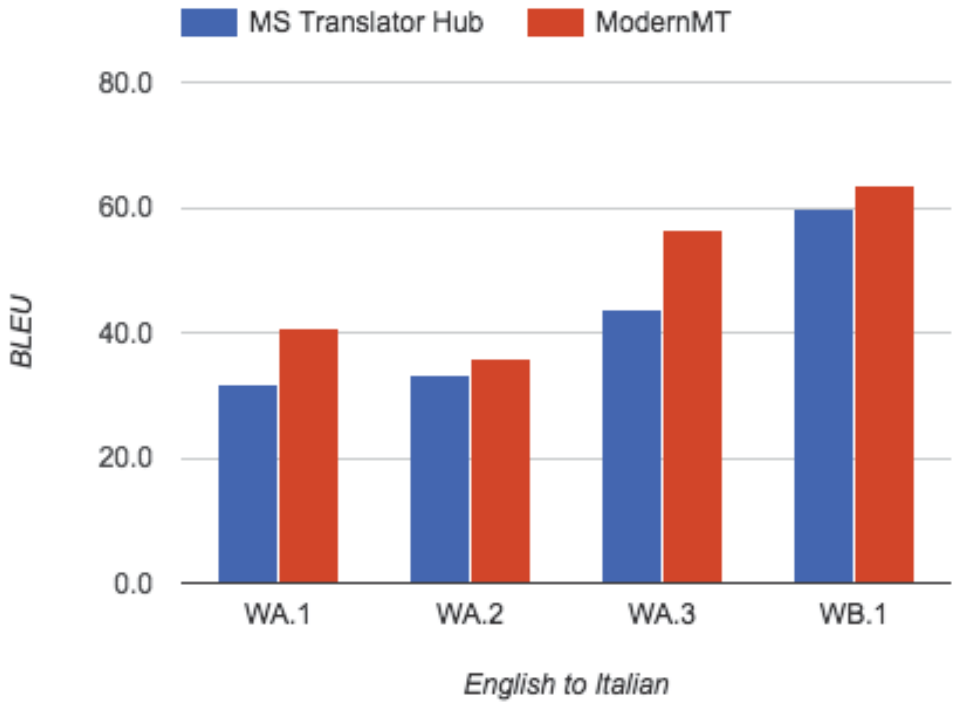
MS Translator Hub vs Modern MT

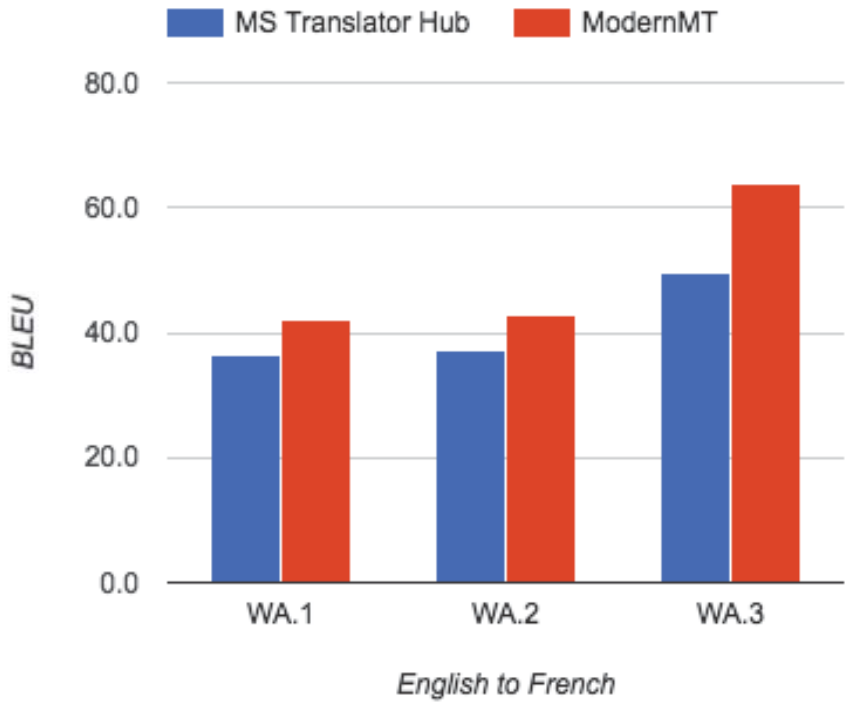




ModernMT Next Generation Machine Translation

MS Translator Hub vs Modern MT





ModernMT Next Generation Machine Translation

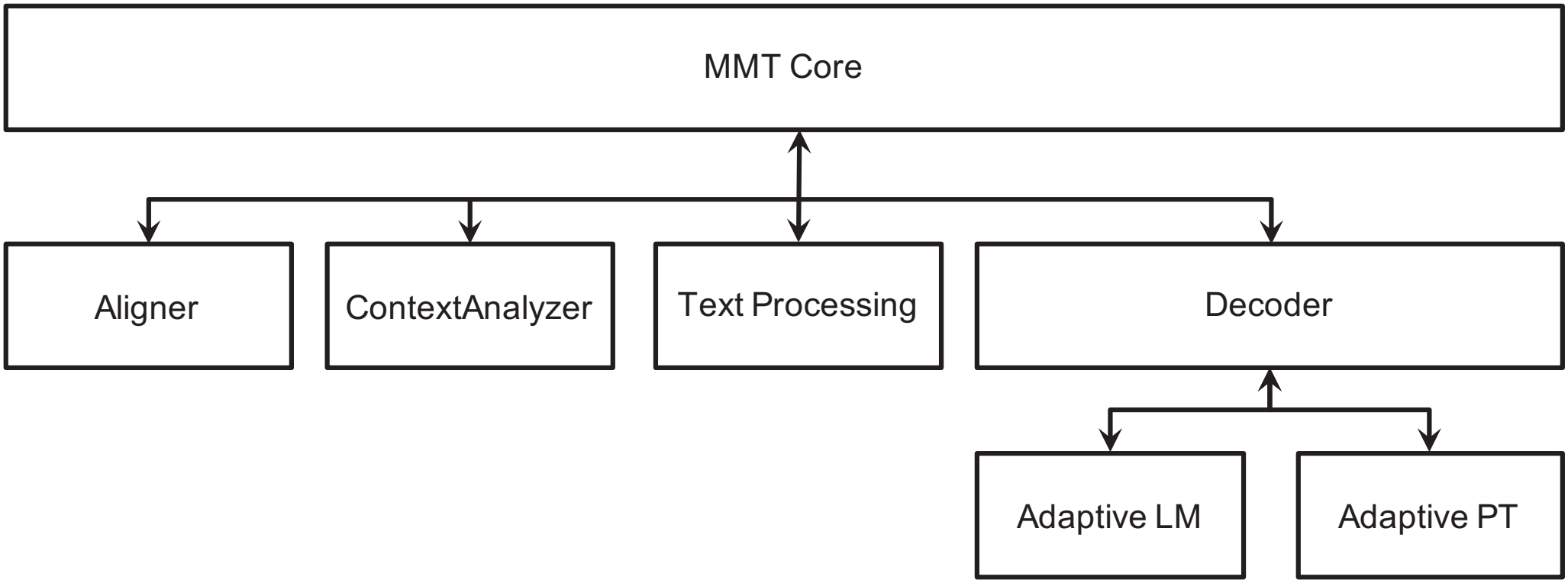
Modern MT core technology

context adaptive

incremental learning

ModernMT Next Generation Machine Translation

Overall Architecture



ModernMT Next Generation Machine Translation

Word Alignment

Object oriented **re-implementation** of FastAlign

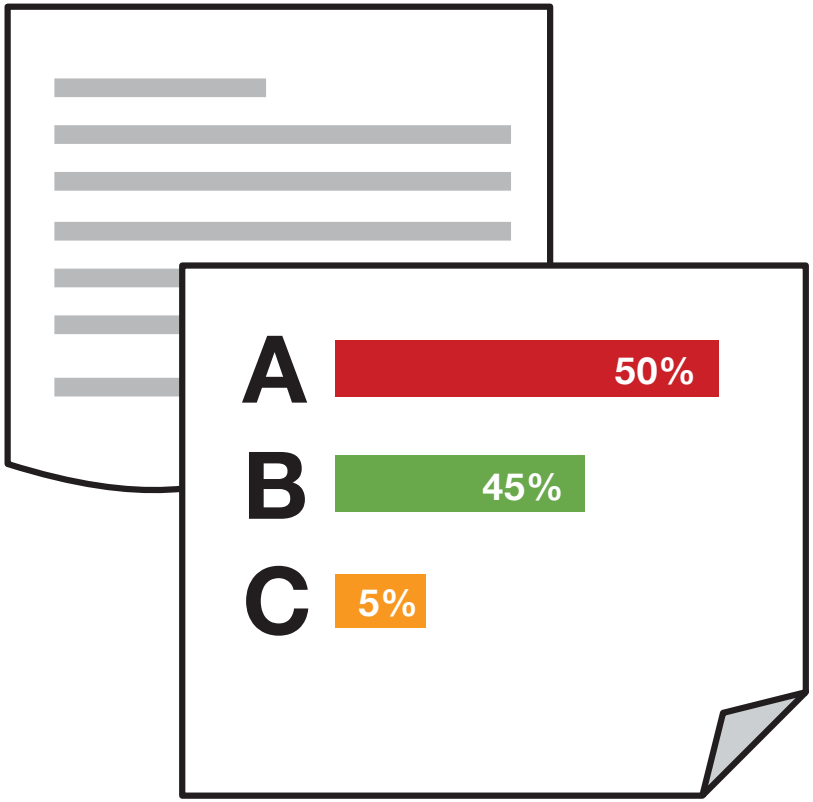
Multithreading

Incremental training

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

ModernMT Next Generation Machine Translation

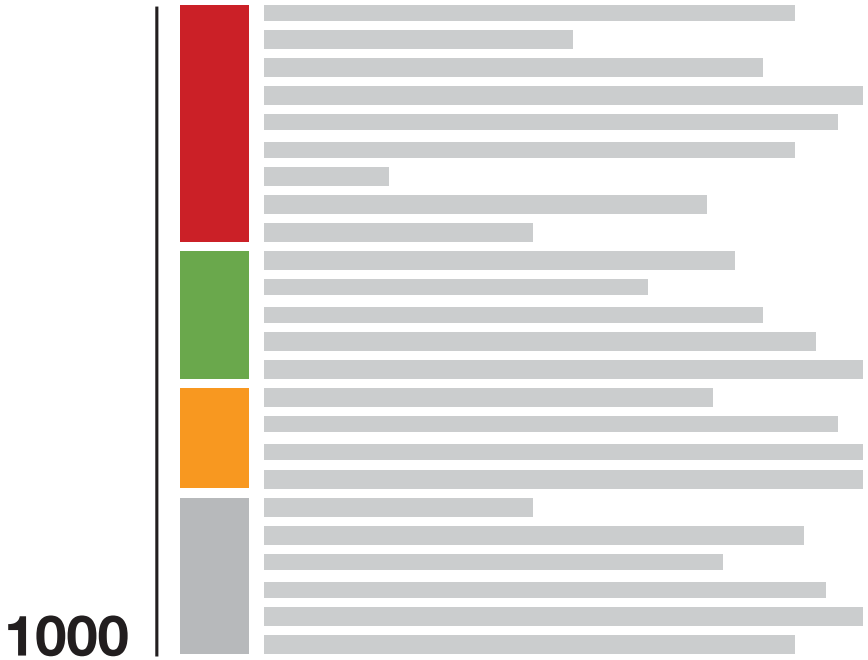
Context Analyzer



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

ModernMT Next Generation Machine Translation

Adaptive Phrase Table

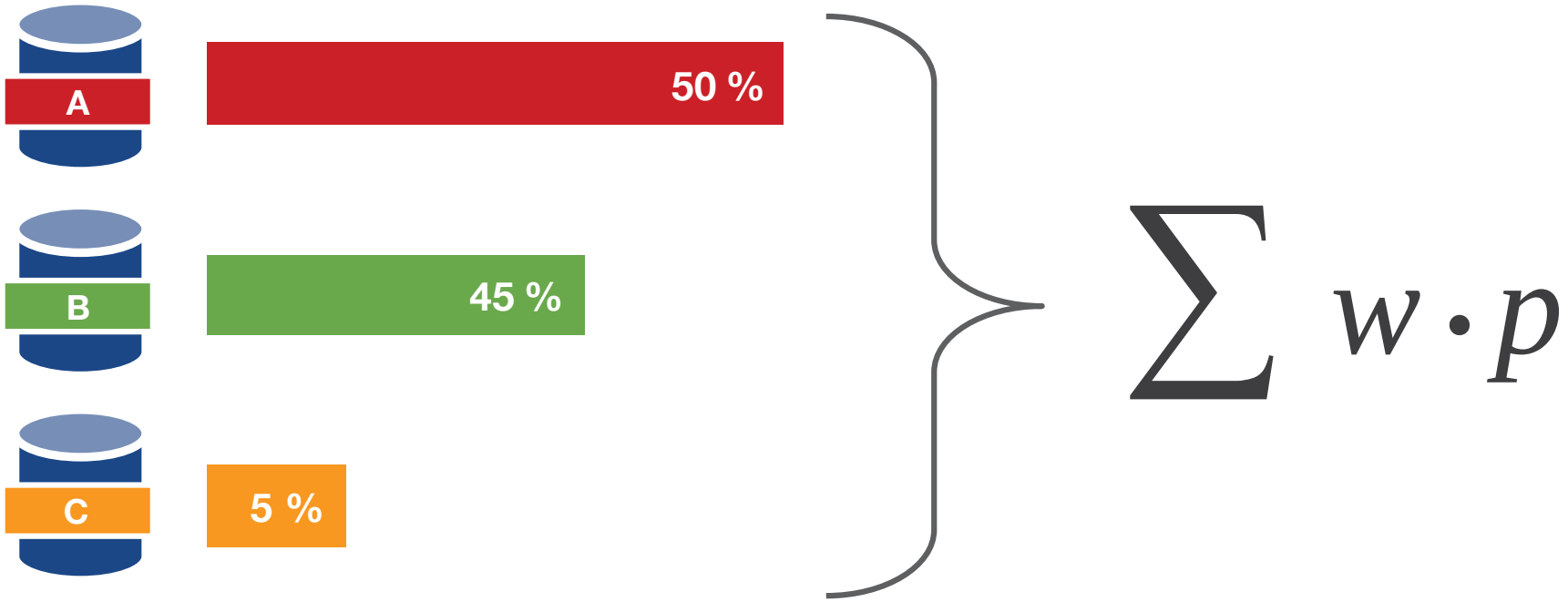


Suffix Array with
Ranked Sampling

- 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

ModernMT Next Generation Machine Translation

Adaptive Language Model



ModernMT Next Generation Machine Translation

ModernMT vs. Moses text processing

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

ModernMT Next Generation Machine Translation

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

ModernMT Next Generation Machine Translation

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

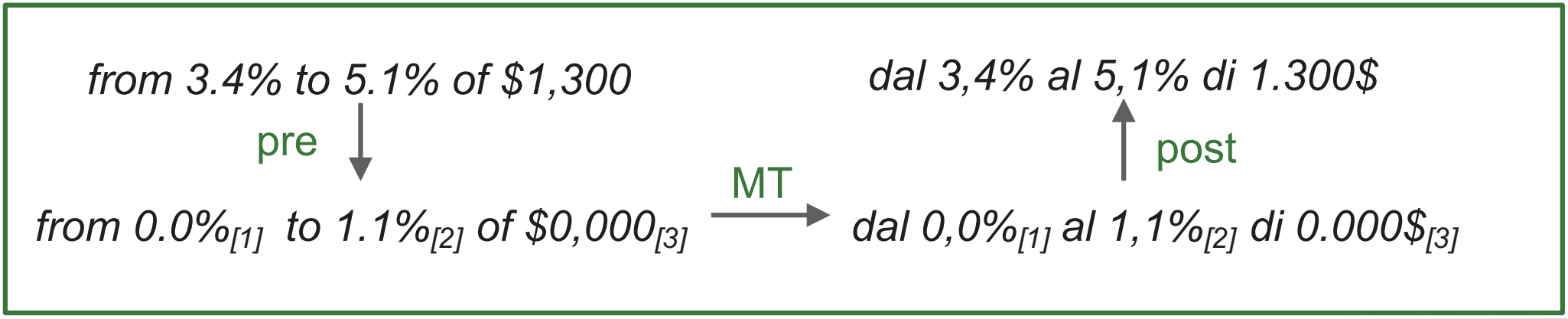
ModernMT Next Generation Machine Translation

Numeric Expressions

Convert digits into placeholders

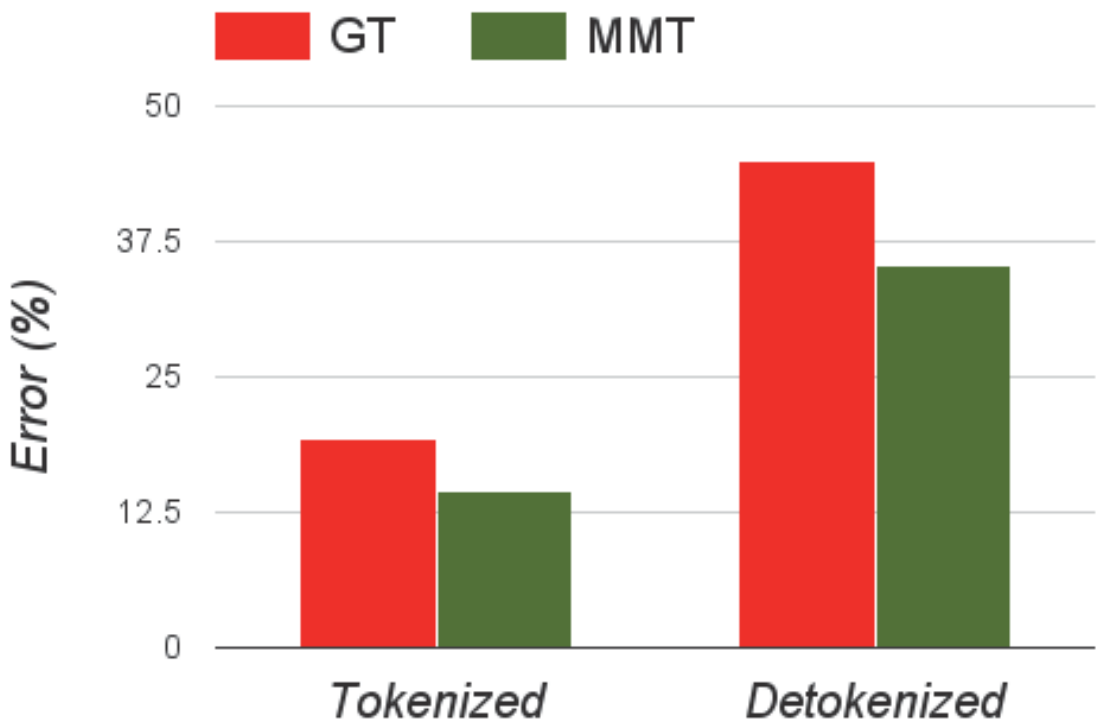
Translate with placeholders

Apply transformation and heuristics (for unaligned expr)



ModernMT Next Generation Machine Translation

Numeric Expressions



Subset of Benchmark 1.1

65 segments

135/134 expressions

MMT better than GT

(rel. delta 25%-21%)

detoken. is problematic

ModernMT Next Generation Machine Translation

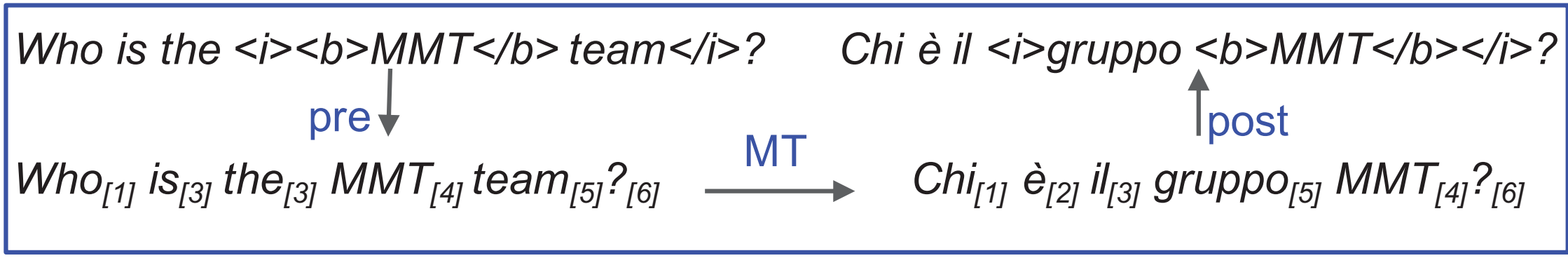
Tag Manager

Identify, classify, and remove tags

Translate w/o tags

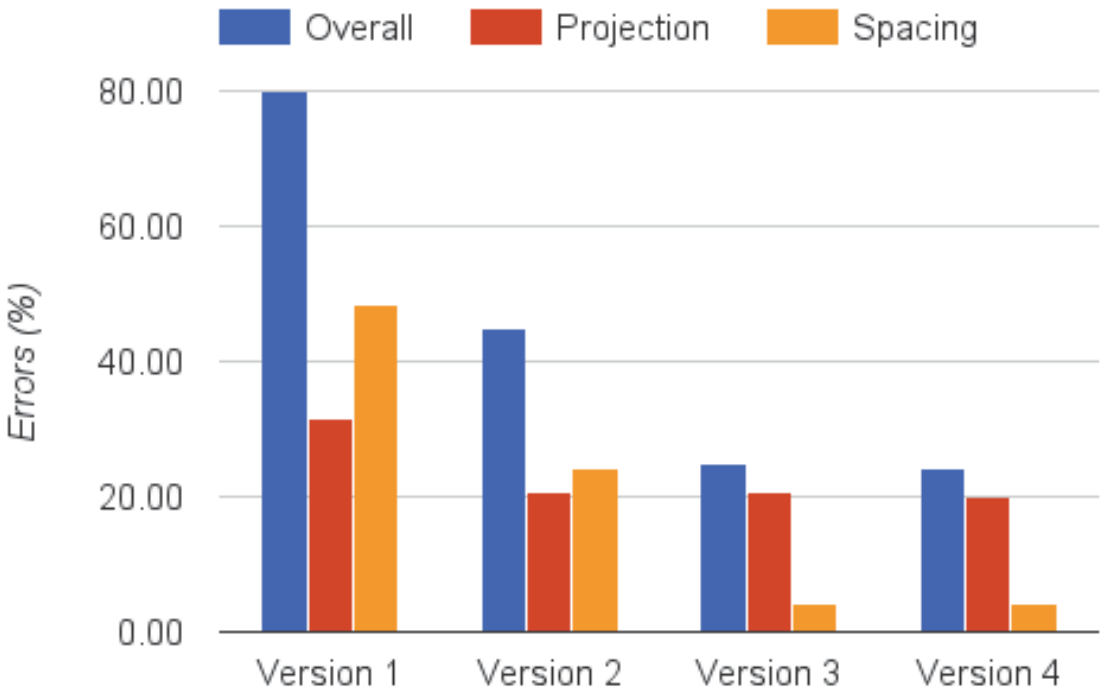
Search insertion points using alignments and heuristics

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



ModernMT Next Generation Machine Translation

Tag Manager



Tag projection error < 20%

Spacing errors around tags <= 4.2%

ModernMT Next Generation Machine Translation

Conclusion

Modern MT

big data, context aware, enterprise

ModernMT Next Generation Machine Translation

Did I mention that MMT will be free?

LGPL/Apache licences

new core technology

no licensing

github.com/ModernMT/MMT

ModernMT Next Generation Machine Translation

Thank You

Project website:

www.ModernMT.eu

github.com/ModernMT/MMT

ModernMT Next Generation Machine Translation

Acknowledgment

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

ModernMT Next Generation Machine Translation

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

ModernMT Next Generation Machine Translation