MT adaptation from TMs in ModernMT

Marcello Federico - FBK, Italy

AMTA, Oct 29 2016 - Austin, Texas
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!
zero training time
manages context
learns from users
scales with data and users
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
Roadmap

2015 Q1
development started

2016 Q2
first alpha release.
2 langs, fast training, context aware, distributed

2016 Q4
first beta release
plug-in 10 langs, incremental learning

2017 Q4
final release
enterprise ready
Prototype - Easy training

> mmt create en it path/to/data
Training takes **30s** for a **1M word TM**

**MMT** is **12x** time faster than std **Moses**
We’re going out. party

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

15%  60%  25%
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 <= delta <= 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

Speed: 1.7sec/segm
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

![Graph](image-url)

**English to Italian**

- MS Translator Hub
- WA.1
- WA.2
- WA.3
- WB.1

**English to French**

- MS Translator Hub
- WA.1
- WA.2
- WA.3
MS Translator Hub vs Modern MT

English to Italian

English to French

ModernMT
Next Generation
Machine Translation
Modern MT core technology

context adaptive
incremental learning
Word Alignment

Object oriented **re-implementation** of FastAlign

Multithreading

Incremental training

<table>
<thead>
<tr>
<th>Giza++</th>
<th>FastAlign++</th>
</tr>
</thead>
<tbody>
<tr>
<td>48,000 sec</td>
<td>2,800 sec <em>(17x speed up)</em></td>
</tr>
<tr>
<td>19.3 BLEU</td>
<td>18.9 BLEU <em>(loss -0.4)</em></td>
</tr>
</tbody>
</table>
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

\[ \sum w \cdot p \]
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

<table>
<thead>
<tr>
<th></th>
<th>Moses Perl Tokenizer</th>
<th>MMT Tokenizer</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Languages</strong></td>
<td>21</td>
<td>45 (+24)</td>
</tr>
<tr>
<td><strong>Speed</strong></td>
<td>17k w/s</td>
<td>340k w/s (x20)</td>
</tr>
</tbody>
</table>

* 4 CPU, 83M word English corpus
Numeric Expressions

Convert digits into placeholders

Translate with placeholders

Apply transformation and heuristics (for unaligned expr)

\[
\text{from 3.4\% to 5.1\% of $1,300} \quad \rightarrow \quad \text{dal 3,4\% al 5,1\% di 1.300\$}
\]

\[
\text{pre} \quad \rightarrow \quad \text{MT} \quad \rightarrow \quad \text{post}
\]

\[
\text{from 0.0\%[1] to 1.1\%[2] of $0,000[3]} \quad \rightarrow \quad \text{dal 0,0\%[1] al 1,1\%[2] di 0.000\$[3]}
\]
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

Who is the `<i>` `<b>` MMT `<b>` team `<i>`?  
Chi è il `<i>` gruppo `<b>` MMT `<b>`? 

Tag Manager

Tag projection error < 20%

Spacing errors around tags <= 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
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

ModernMT  Next Generation Machine Translation