TUTORIAL

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

November 1, 2016

ModernMT

Marcello Federico
Marco Trombetti

AMTA
Association for Machine Translation in the Americas

ModernMT Tutorial

Marcello Federico - FBK, Italy
Marco Trombetti - MateCat, Italy
Davide Caroselli - Translated, Italy

AMTA - 1 November 2017, Austin, Texas

ModernMT Next Generation Machine Translation

Outline

- Introduction + (Marcello, 40 min)
- Development + (Davide, 20 min)
- *Break*
- QA and discussion (all, 20 min)
- Hands-on (Marco & Davide, 70 min)

ModernMT Introduction

Marcello Federico - FBK, Italy

AMTA - 1 November 2017, Austin, Texas

ModernMT Next Generation Machine Translation

Translated FONDAZIONE BRUNO KESSLER THE UNIVERSITY of EDINBURGH TAUS

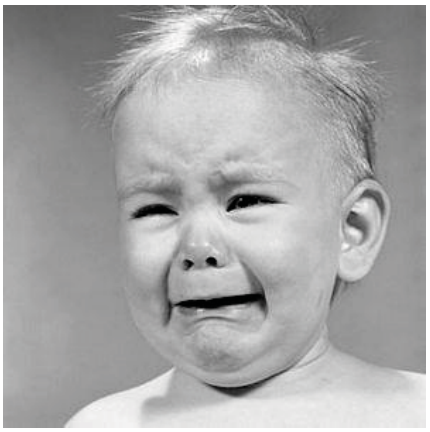
Translators' pains with MT

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

Setting up MT for CAT today

(1) select TMs
(2) collect extra data
(3) train and evaluate engine
(4) doesn't work? back to (2)
(5) apply MT on documents
(6) create fake TM with MT
(7) import TMs in CAT tool
(8) start translating
(9) adapt engine to new data
(10) new project? back to (1)

ModernMT Next Generation Machine Translation

The Modern MT way

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

ModernMT Next Generation Machine Translation

Enhance translator's experience

Adaptation from TMs
Learning from corrections

Modern MT in a nutshell

fast training
manages **context**
learns from users
scales with data and users

ModernMT Next Generation Machine Translation

Team

Business

Translated.net

TAUS

Research

FONDAZIONE BRUNO KESSLER

THE UNIVERSITY of EDINBURGH

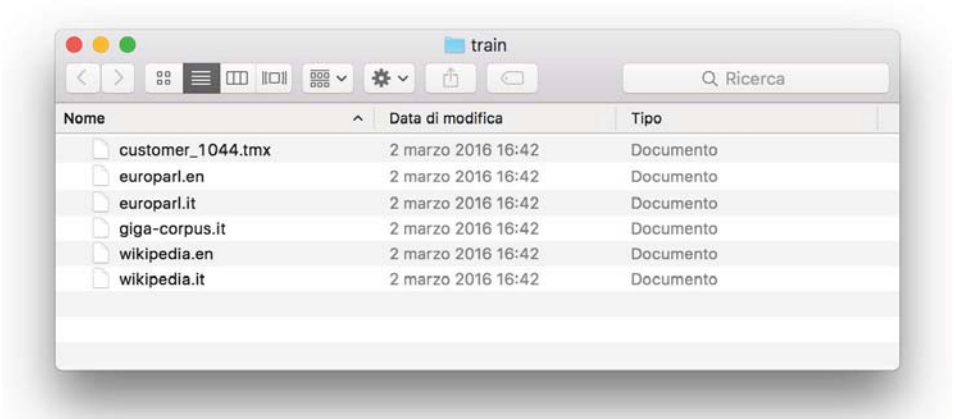
ModernMT Next Generation Machine Translation

Acknowledgment

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

ModernMT Next Generation Machine Translation

Fast and easy training




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

ModernMT Next Generation Machine Translation

Prototype - 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

REST API

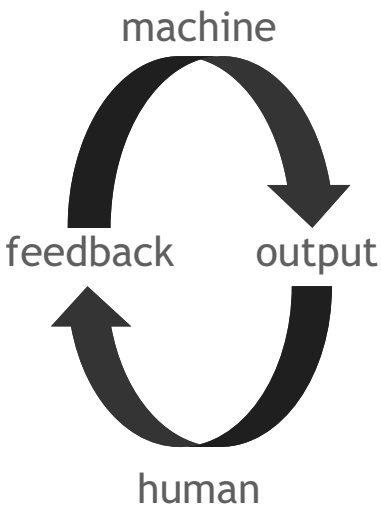
GET /translate?**q**=party&**context**=We+approved+the+law

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

```
> mmt start
```

ModernMT Next Generation Machine Translation

learning from users



machine

-> **supports** the user

-> **improves** over time

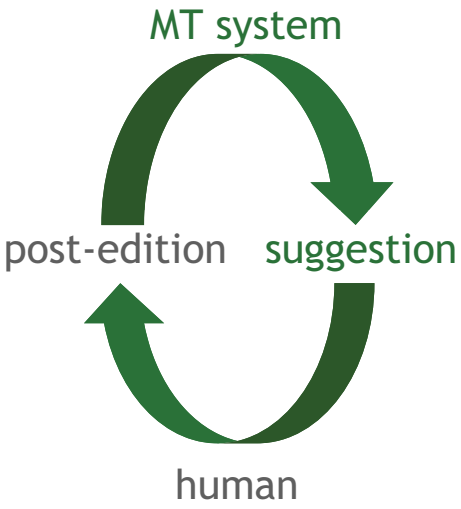
human

-> keeps her workflow

-> is an **unaware trainer**

ModernMT Next Generation Machine Translation

incremental learning



goal

-> **correct** recurring MT errors

-> keep **system** up-to-date

features

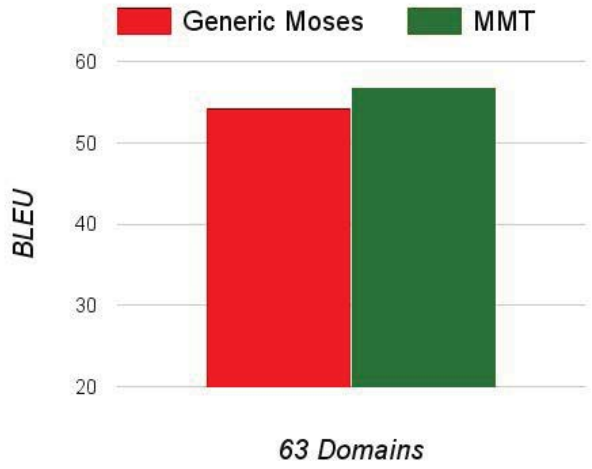
-> **adapt** to user/domain

-> **real-time** processing

Enhanced text processing

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

Prototype (March 2016)



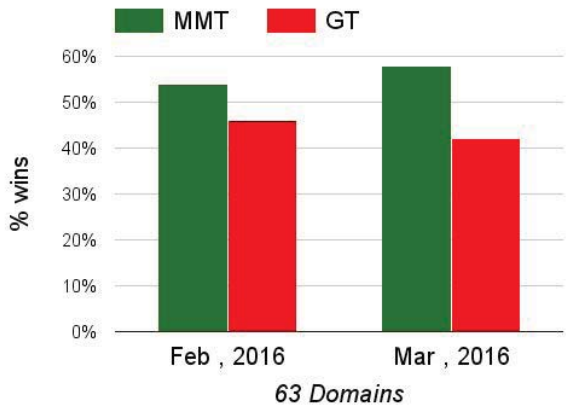
Benchmark 1.1
- **No XML tags**

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

Core Technology

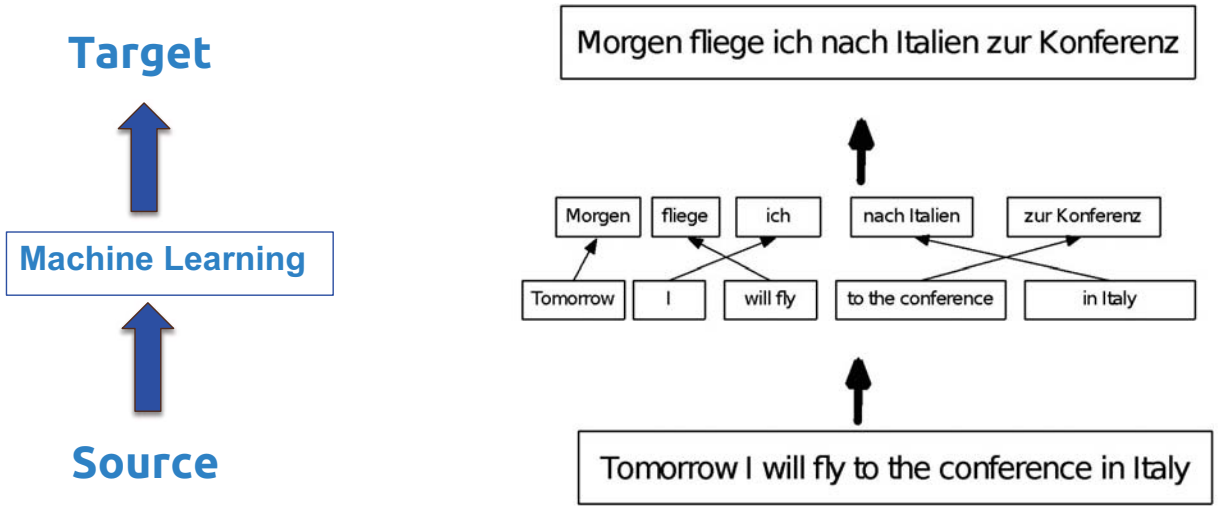
English

German

Source → Machine Learning → Target

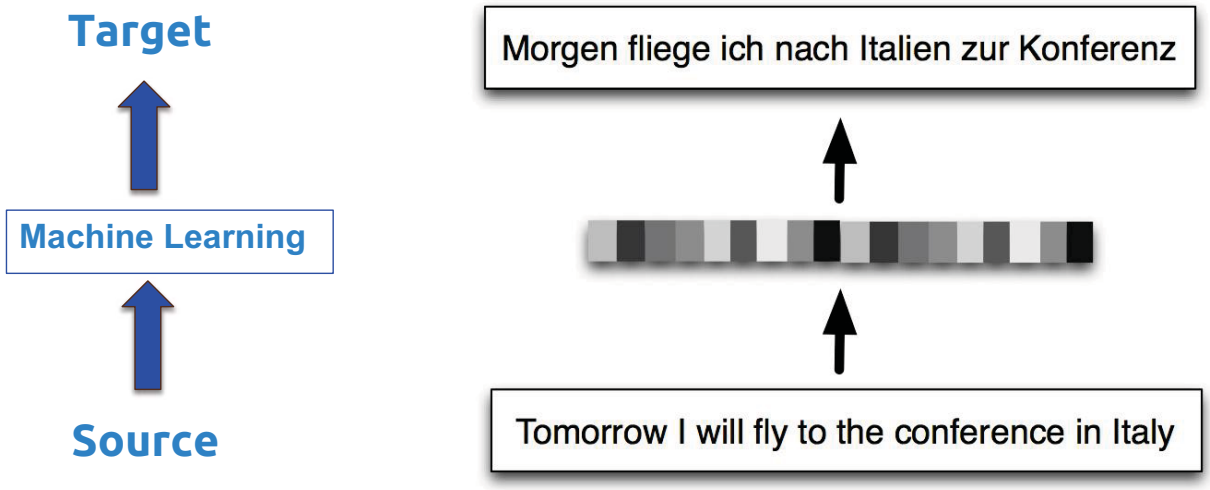
ModernMT Next Generation Machine Translation

Today: Phrase-based MT



ModernMT Next Generation Machine Translation

Tomorrow: Neural MT (2017)



ModernMT Next Generation Machine Translation

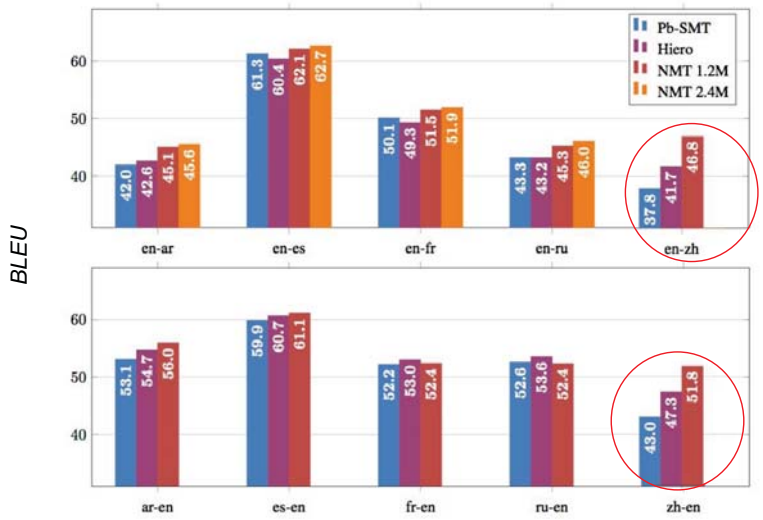
Phrase-based versus Neural MT

- ❑ PBMT has dominated for about 20 years
- ❑ NMT started to outperform PBMT in 2015
- ❑ NMT is now taking over PBMT in research

	PBMT	NMT
Computational cost	0.80$/hour (CPU-16)	0.70$/hour (GPU-1)
Training time (100Mw)	few hours	1 week
Adaptation (1Mw)	< 1 min	hours
Translation speed	455 w/s	409 w/s
Translation quality		+ 5-25% BLEU

ModernMT Next Generation Machine Translation

Phrase-based versus Neural MT



Task: United Nations Parallel Corpus v1.0.

Source:

M. Junczys-Dowmunt, T. Dwojak, H. Hoang, *Is Neural Machine Translation Ready for Deployment? A case Study on 30 Translation Direction*, arXiv. 11 Oct 2016.

ModernMT Next Generation Machine Translation

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

neural MT
enterprise ready

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

Conclusion

Modern MT

big data, context aware, enterprise

Thank You

Project website:

www.ModernMT.eu

github.com/ModernMT/MMT

ModernMT Next Generation Machine Translation

Outline

- Introduction + (Marcello, 40 min)
- Development + (Davide, 20 min)
- *Break*
- QA and discussion (all, 20 min)
- Hands-on (Marco & Davide, 70 min)

ModernMT Next Generation Machine Translation

Modern MT Development

Davide Caroselli - Translated, Italy

1 November 2016 - Austin, Texas

ModernMT Next Generation Machine Translation

Translated FONDAZIONE BRUNO KESSLER THE UNIVERSITY of EDINBURGH TAUS

The Modern MT way

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

Modern MT in a nutshell

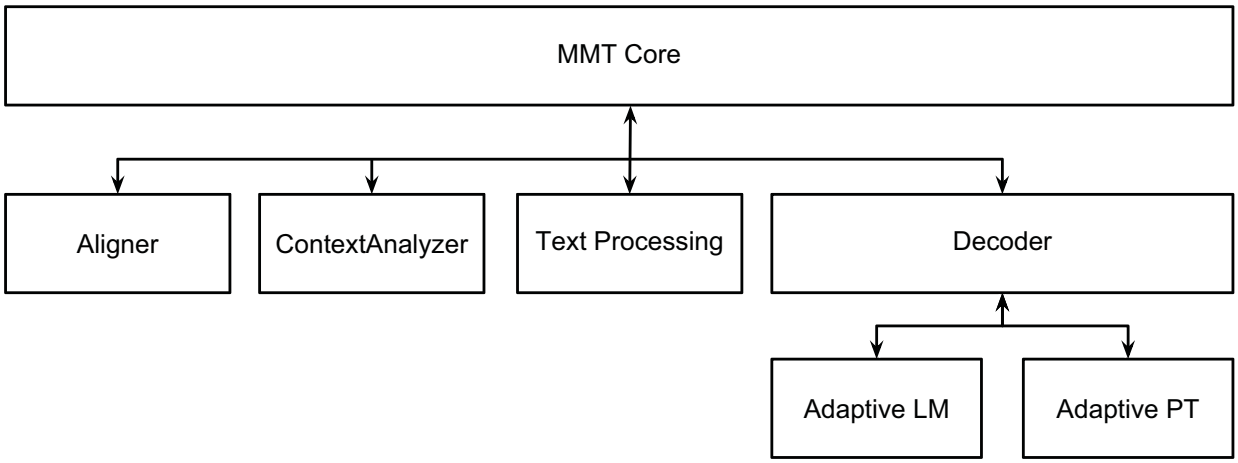
fast training
manages **context**
learns from users
scales with data and users

ModernMT Next Generation Machine Translation

Modern MT
core components

ModernMT Next Generation Machine Translation

Overall Architecture



ModernMT Next Generation Machine Translation

Word Alignment

Object oriented **re-implementation** of FastAlign

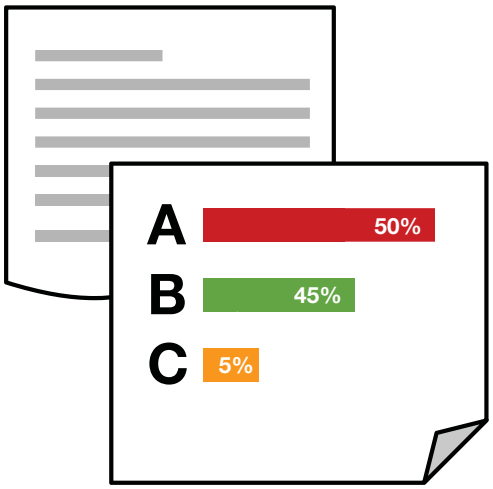
Multithreading

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

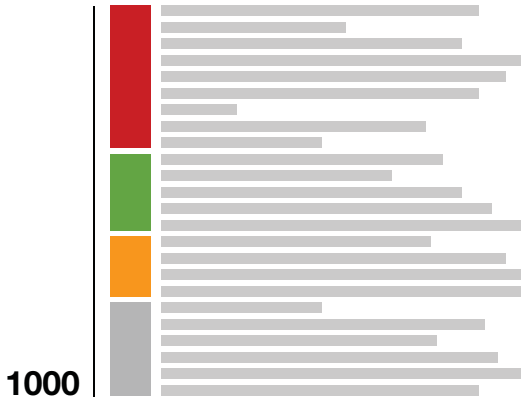
Context Analyzer



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

ModernMT Next Generation Machine Translation

Adaptive Phrase Table

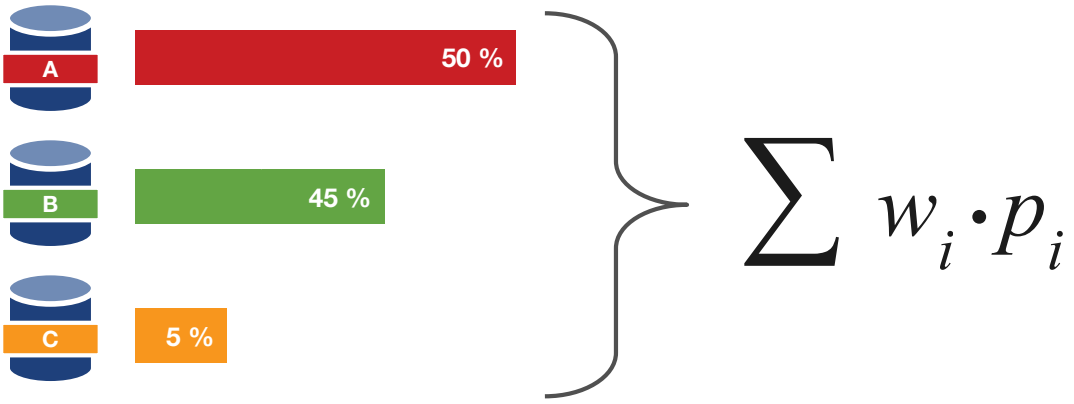


Suffix Array with Ranked Sampling

- 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

Modern MT
cool features

ModernMT Next Generation Machine Translation

Cool features

- **Fast** processing
- **Integrated** in the translation process
- **45 languages** tokenizer
- Automatic TMs **cleaning**
- Built-in **expression localizer**
- **Tag projector** for automatic placement of XML tags

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

TM Cleaning

- **Draft translations** cleaning

EN This is **just** an example → IT Questo è un esempio
IT Questo è **solo** un esempio

- **XML tags** cleaning and text normalization

Did you know that the ****Lunar distance**** is **>** 300.000km ?

↓

Did you know that the Lunar distance is **>** 300.000km ?

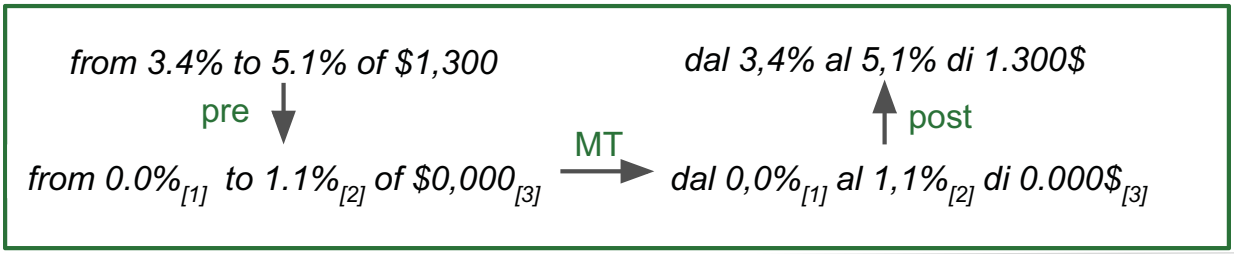
ModernMT Next Generation Machine Translation

Numeric Expressions

Convert digits into placeholders

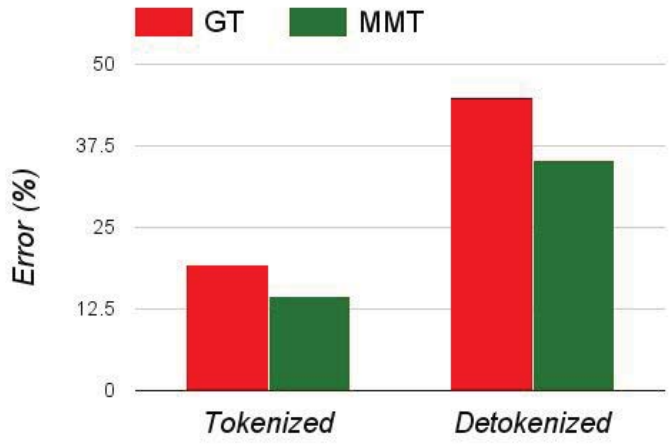
Translate with placeholders

Apply transformation and heuristics (for unaligned expr)



ModernMT Next Generation Machine Translation

Numeric Expressions



Test on 65 segments

135 numerical expressions

MMT better than GT

(rel. delta 25%-21%)

ModernMT Next Generation Machine Translation

Tag Manager

Identify, classify, and remove tags

Translate w/o tags

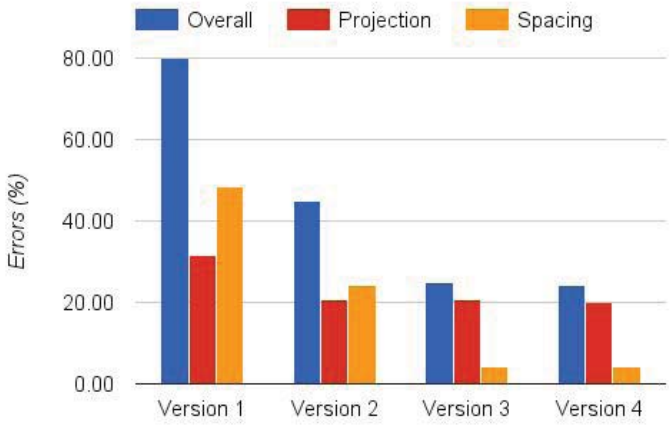
Search insertion points using alignments and heuristics

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

Who is the MMT team? → pre → *Who*$_{[1]}$ *is*$_{[3]}$ *the*$_{[3]}$ *MMT*$_{[4]}$ *team*$_{[5]}$*?*$_{[6]}$ → MT → *Chi*$_{[1]}$ *è*$_{[2]}$ *il*$_{[3]}$ *gruppo*$_{[5]}$ *MMT*$_{[4]}$*?*$_{[6]}$ → post → *Chi è il gruppo MMT?*

ModernMT Next Generation Machine Translation

Tag Manager



Tag projection error < 20%

Spacing errors around tags <= 4.2%

ModernMT Next Generation Machine Translation

Modern MT
field testing

ModernMT Next Generation Machine Translation

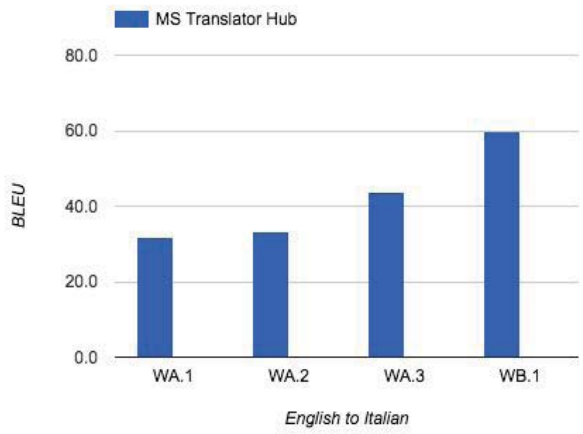
Field testing

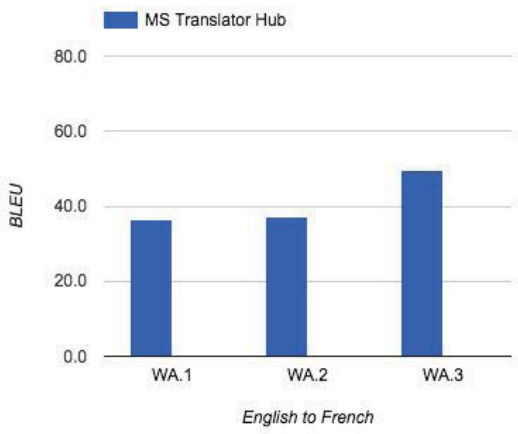
Real business cases:

- Two world-class web companies
- Private TMs and real documents as test data
- Compared with current state-of-the-art **adaptive machine translation**.
- Two language pairs: **EN→IT** and **EN→FR**

ModernMT Next Generation Machine Translation

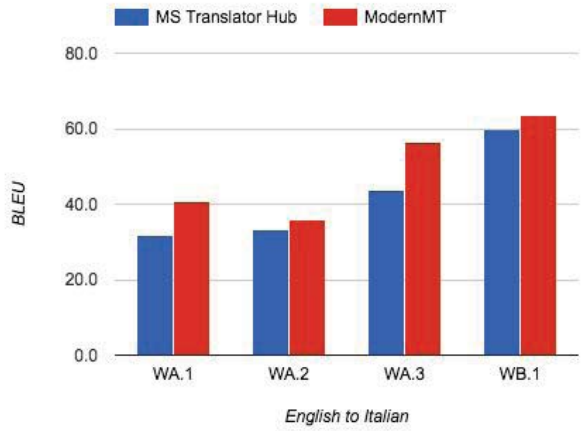
MS Translator Hub vs Modern MT

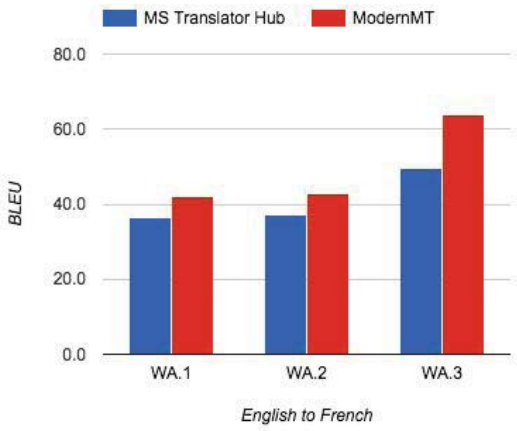




ModernMT Next Generation Machine Translation

MS Translator Hub vs Modern MT





ModernMT Next Generation Machine Translation

Thank you

easy to use product
context aware translation
incremental system
open source, **no licensing**

GitHub

github.com/ModernMT/MMT

Translated