Memory Based MT

Antal van den Bosch

Peter Berck

Tilburg centre for Creative Computing

Tilburg University

The Netherlands

http://ilk.uvt.nl/mbmt

Outline

- MBMT
- Evaluation
- Software
- Conclusions

What is Memory Based MT?

- Example based MT
- Take bits of source text, map to bits of target text
- Recombine the target bits into a sentence

Details

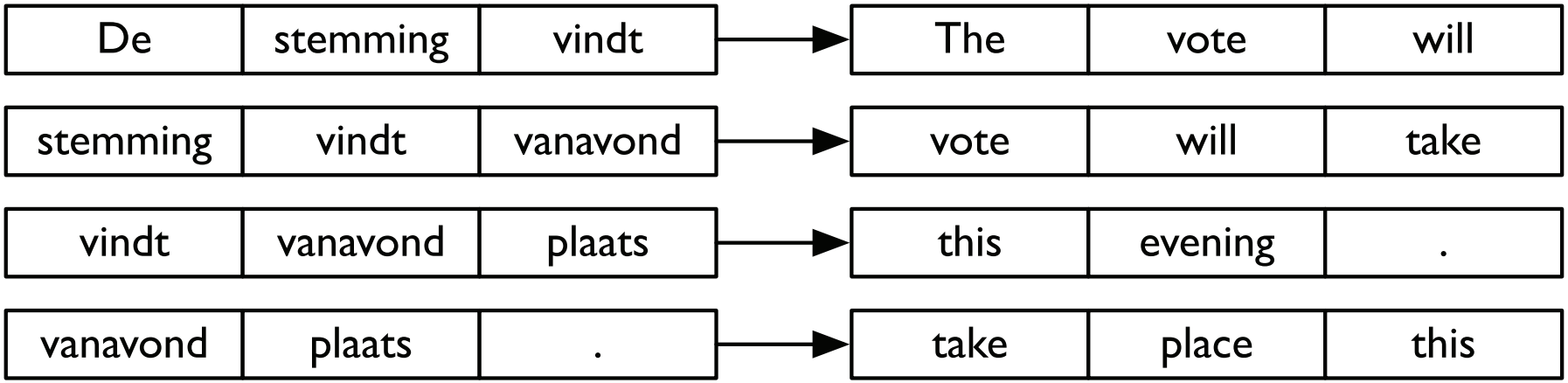
- Use (GIZA++) aligned sentences

Aligned Text

Details

- Use (GIZA++) aligned sentences
- Use trigrams

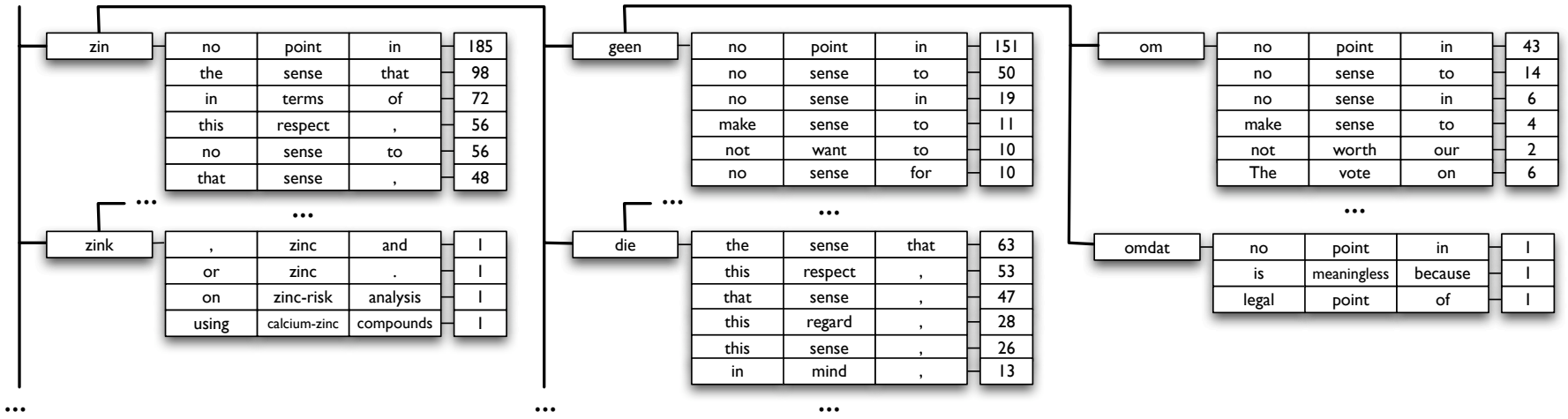
Trigram Mapping



Details

- Use (GIZA++) aligned sentences
- Use trigrams
- Decision tree based k-NN classifier

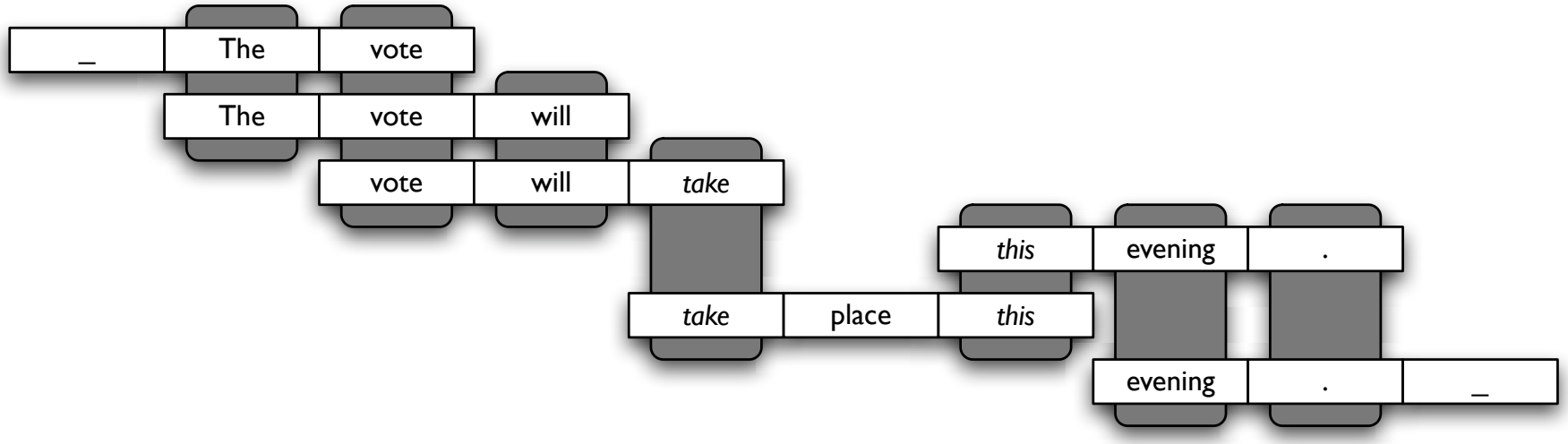
IGTree


zin
no point in 185
the sense that 98
in terms of 72
this respect , 56
no sense to 56
that sense , 48
zink
, zinc and 1
or zinc . 1
on zinc-risk analysis 1
using calcium-zinc compounds 1
geen
no point in 151
no sense to 50
no sense in 19
make sense to 11
not want to 10
no sense for 10
die
the sense that 63
this respect , 53
that sense , 47
this regard , 28
this sense , 26
in mind , 13
om
no point in 43
no sense to 14
no sense in 6
make sense to 4
not worth our 2
The vote on 6
omdat
no point in 1
is meaningless because 1
legal point of 1

Details

- Use (GIZA++) aligned sentences
- Use trigrams
- Decision tree based k-NN classifier
- Decoder

Recombine


_
The
vote
The
vote
will
vote
will
take
take
place
this
this
evening
.
evening
.
_

Details

- Use (GIZA++) aligned sentences
- Use trigrams
- Decision tree based k-NN classifier
- Decoder
- LM assigns perplexity score to sentence

LM Disambiguation

NGOs are good within the European Union .

perplexity = 198.326

NGOs are good the EU , within the European .

Perplexity = 243.701

Details

- Use (GIZA++) aligned sentences
- Use trigrams
- Decision tree based k-NN classifier
- Decoder: when more than one answer, use language model
- LM assigns perplexity score to sentence
- No explicit linguistic knowledge


deze zin kan nooit vertaald worden .

Experiments

- OpenSub 2 million instances
- JRC-Aquis 14 million instances
- EMEA 9 million instances

- LM: Reuters 37 million instances

Results: OpenSub

	WER	PER	BLEU	METEOR	NIST
Moses	53.2878	46.9615	0.3289	0.5407899	5.9035
MBMT	**68.3948**	**61.3335**	**0.1631**	**0.4015985**	**4.2428**
Google	50.0984	45.0847	0.3056	0.5223539	5.7893
Systran	60.7691	54.6135	0.1749	0.4500350	4.5828

Results: JRC-Aquis

	WER	PER	BLEU	METEOR	NIST
MBMT	**58.5586**	**36.7447**	**0.4513**	**0.6336529**	**7.8306**
Google	48.4244	32.8729	0.4713	0.6511708	8.2668
Systran	60.8488	43.0711	0.3321	0.5549924	6.7365

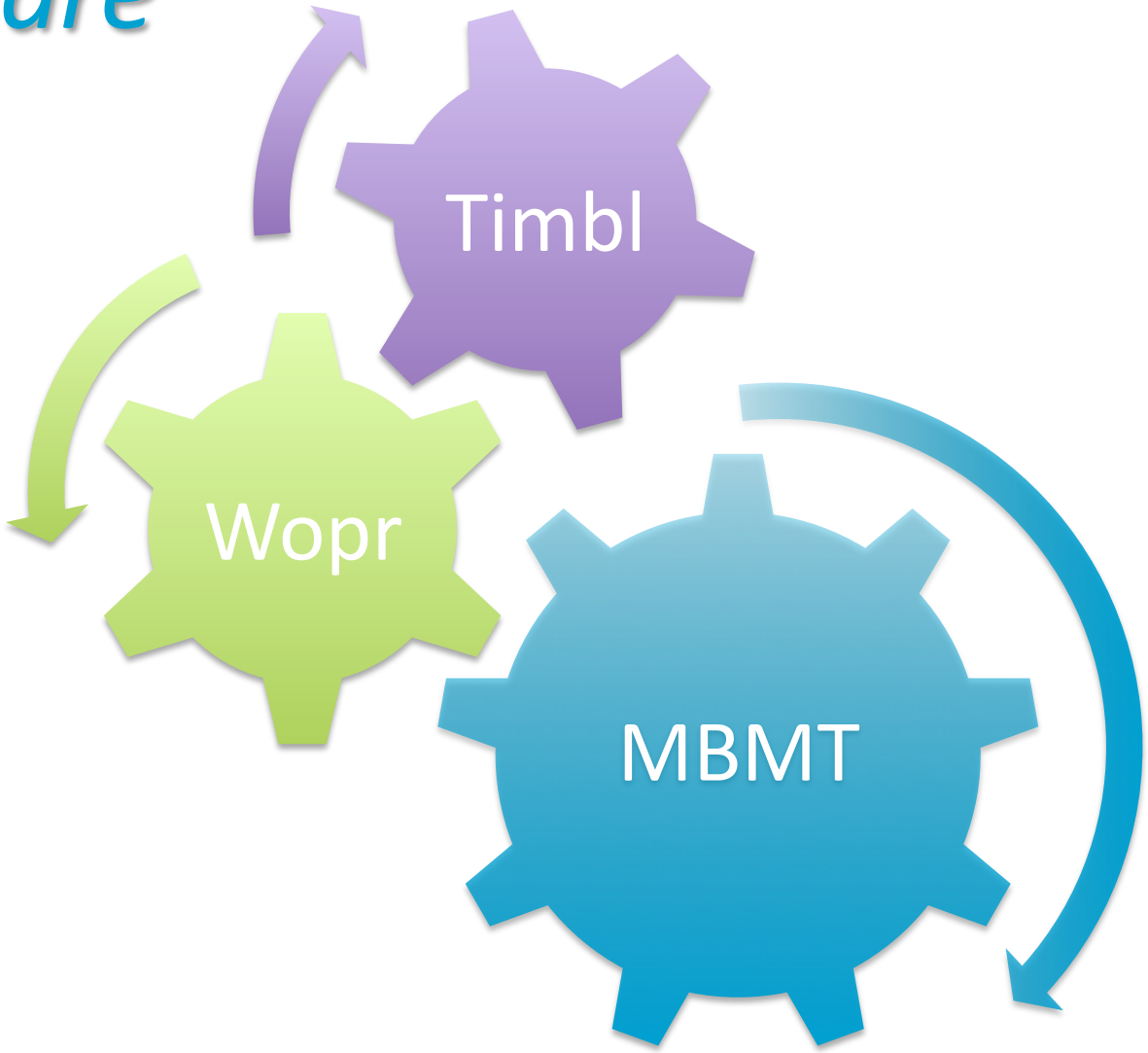
Results: EMEA

	WER	PER	BLEU	METEOR	NIST
Moses	46.5543	39.3553	0.4701	0.6501440	7.0593
MBMT	**72.7873**	**63.6633**	**0.2633**	**0.4801131**	**5.1145**
Google	57.5692	50.4385	0.3918	0.5829913	6.3772
Systran	63.2418	55.1430	0.2895	0.5366058	5.4716

Results: Performance & Speed

	WER	PER	BLEU	METEOR	NIST	Train	Test
MBMT	72.7	63.6	0.238	0.460	4.97	20:17	0:08
Moses	46.6	39.4	0.470	0.650	7.06	3:10:06	2:51

Software



Firefox File Edit View History Bookmarks Tools Window Help (Charged) Fri 17:04

MBMT: Memory-based machine translation

http://ilk.uvt.nl/mbmt/

Most Visited Getting Started Latest Headlines Apple News nu 101 Google Groups Cam Nij Tw menu GF /. WOPR MF NRC NS start [Peter] ResearchWiki mm sms NS

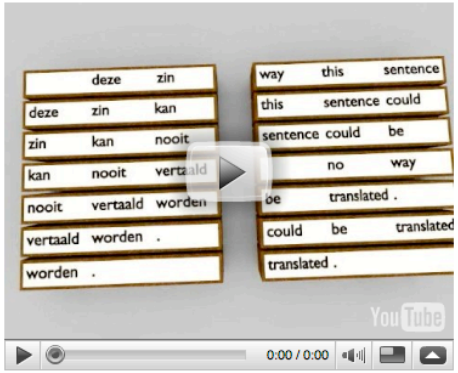
message masala | MT Marathon 2009 | MIT Press Journals – Computati... | Google Maps | MBMT: Memory-based machine...

MBMT

Memory-based machine translation

MBMT: Memory-based machine translation

MBMT is a software package for training and running a machine translation system. It is based on the *k*-nearest neighbor classifier as implemented in **TiMBL**, and also features a memory-based target language model based on **WOPR**, a memory-based language model. It assumes a word-aligned bilingual parallel training corpus, such as produced by **GIZA++**.



Features

- Generates a machine translation model from a word-aligned parallel corpus
- Fast training and translation

MBMT is free software; you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation.

Written by

MBMT is written by Antal van den Bosch, with contributions from Peter Berck and Ko van der Sloot.

Archived versions

version 0.1

Download and installation

mbmt-0.1.tar.gz (139 Kb)

To install, please follow these basic instructions:

- **MBMT** relies on an installation and availability in $PATH of the following two packages:
  - **Timbl** version 6.1.4 (or higher);
  - **WOPR** version 1.4.6 (or higher)
- The tarball will unpack ('tar zxvf mbmt-0.1.tar.gz') in a directory called 'mbmt-0.1'.
- In the 'mbmt-0.1' directory, issue a './configure' command, followed by 'make'.
- If you want to install the software elsewhere, issue a './configure --prefix <install-dir>', followed by 'make' and 'make install'.

Quick start:

- The 'mbmt.sh' script included in the package runs a full training and translation process, based on a **GIZA++**-aligned 'A3.final' file, and a target-language text (one sentence per line);
- For example, run the script by issuing 'mbmt.sh JRC-Acquis.sample.A3.final JRC-Acquis.source.test.txt' (Dutch-English files extracted from the **JRC-Acquis** multilingual parallel corpus).

MBMT has been compiled successfully with gcc (4.0 - 4.2), on Intel platforms running several versions of Linux and the Mac OS X platform.

References

For more information and background on MBMT, see

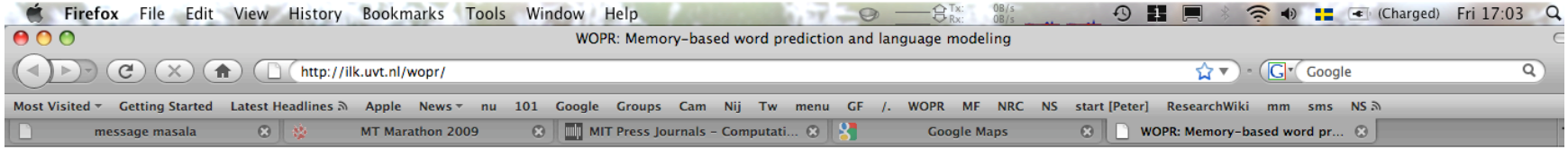
- MT Marathon paper
- Van den Bosch, A., Stroppa, N., and Way, A. (2007). A memory-based classification approach to marker-based EBMT. In F. Van Eynde, V. Vandeghinste, and I. Schuurman (Eds.), *Proceedings of the METIS-II Workshop on New Approaches to Machine Translation*, pp. 63-72. Leuven, Belgium.
- Stroppa, N., Van den Bosch, A., and Way, A. (2007). Exploiting source similarity for SMT using context-informed features. In A. Way and B. Gawronska (Eds.), *Proceedings of the 11th International Conference on Theoretical Issues in Machine Translation* (TMI 2007), *Skövde University Studies in Informatics* **2007:1**, pp. 231-240.

Sponsor

MBMT is developed as part of the Implicit Linguistics project, funded by NWO, the Netherlands Organisation for Scientific Research.

Antal.vdnBosch@uvt.nl | Last update: Sun Jan 18 2009

http://ilk.uvt.nl/mbmt



WOPR

Memory-based word prediction and language modeling

WOPR: Memory-based word prediction and language modeling

WOPR is a wrapper around the *k*-nearest neighbor classifier in **TiMBL**, offering word prediction and language modeling functionalities. Trained on a text corpus, **WOPR** can predict missing words, report perplexities at the word level and the text level, and generate spelling correction hypotheses.

*The **WOPR** name is obviously a blatant cultural reference to the mainframe computer WOPR, "War Operation Plan Response", a key role player in the 1983 US movie War Games. Through a hacked phone dailup connection, WOPR enjoys playing games with a teenager played by a young Matthew Broderick, almost causing a full nuclear war. Image from Wikipedia.*

Features

- Generates language models
- Tests language models on new text, reporting perplexities, prediction distributions, word-level entropies and perplexities
- Optionally exports ARPA-formatted language model files
- Optionally filters its output for spelling correction candidates

WOPR is free software; you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation.

Written by

WOPR is written by Peter Berck, with input from Antal van den Bosch and Ko van der Sloot.

Archived versions

version 1.4.6

Download and installation

wopr-1.4.7.tar.gz (120 Kb)

To install, please follow these basic instructions:

- Please make sure you have installed **Timbl** version 6.1.4 (or higher).
- Follow the HOWTO
- For general information on Wopr, see here.

Wopr has been tested on

- Intel platform running several versions of Linux
- AMD64 platform running gentoo linux
- Mac OS X platform

WOPR incorporated

WOPR is used in **MBMT**, our memory-based machine translation software.

Sponsor

WOPR is developed as part of the Implicit Linguistics project, funded by NWO, the Netherlands Organisation for Scientific Research.

NWO

References

For more information and background on WOPR, see

- Van den Bosch, A. (2005). Scalable classification-based word prediction and confusible correction. *Traitement Automatique des Langues*, **46:2**, pp. 39-63.

Antal.vdnBosch@uvt.nl | Last update: Fri Jan 23 2009

http://ilk.uvt.nl/wopr

Specs

mbmt.sh script which uses:

- MBMT
  - C programs to make instances and recombine
- Wopr
  - C++ language model
- Timbl
  - C++ instance based learner
- Tested on Linux and OS X

Quickstart

- Install Timbl, Wopr and MBMT
- Run the mbmt.sh script which:
  - Takes an aligned file
  - Creates instances
  - Trains translation and language models
  - Runs test set

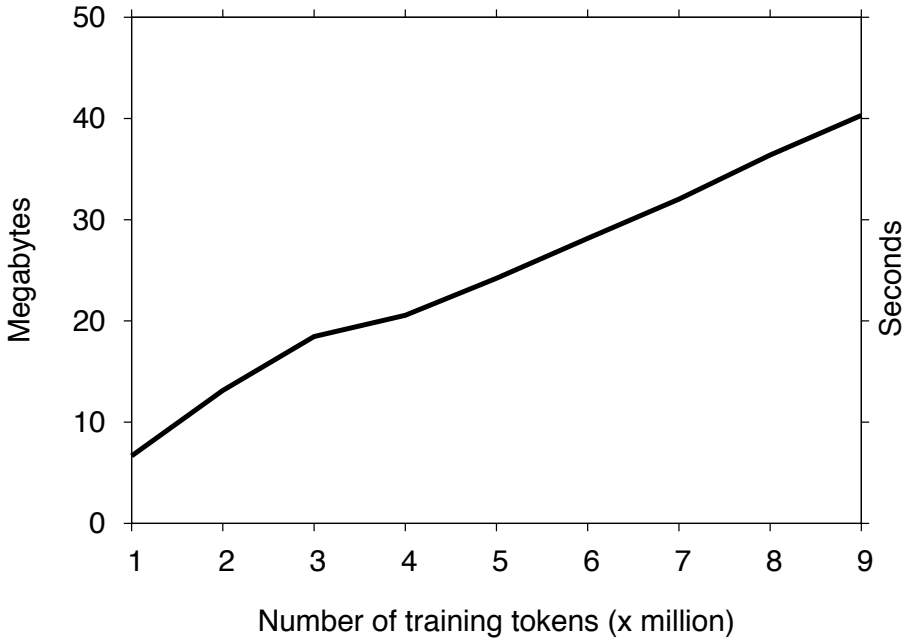
Conclusions: Cons

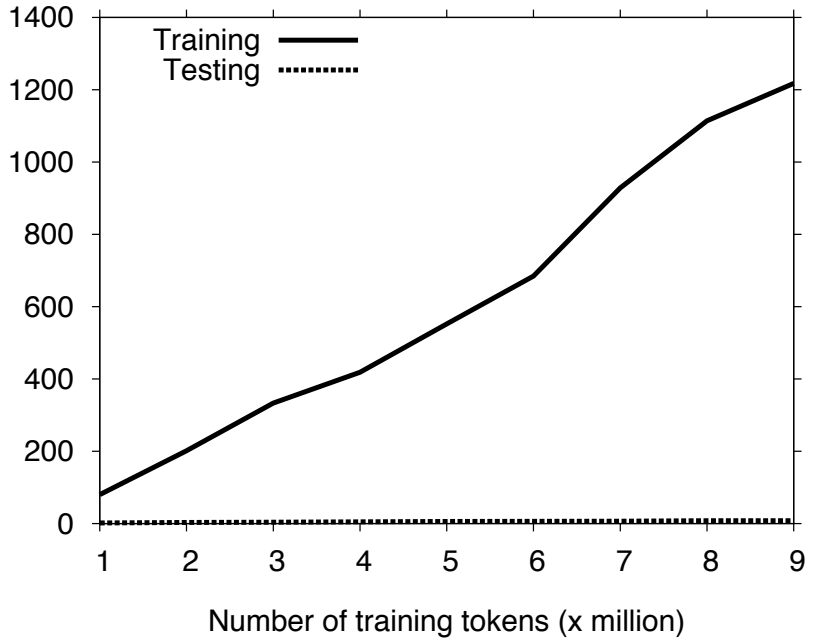
- Inferior to Moses
- Because:
  - No smoothing
  - No fertility or “null” model
  - Limited to trigrams

Conclusions: Pro

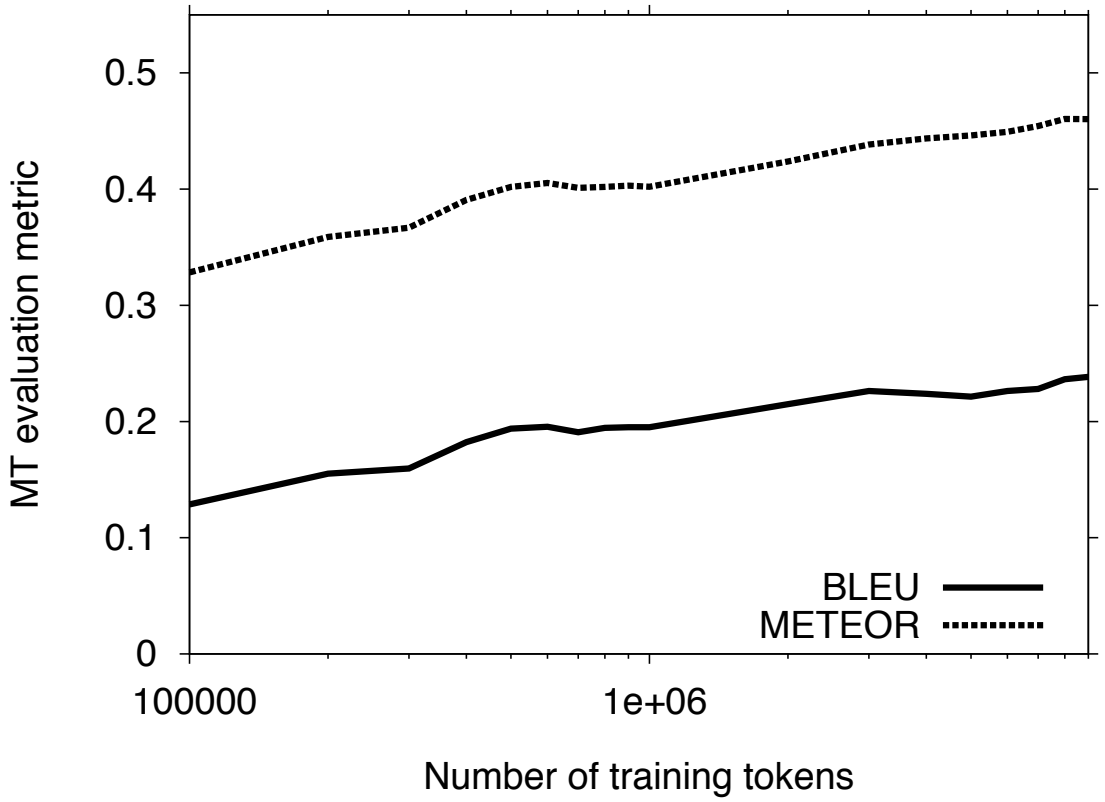
- Speed
  - training and decoding
- Memory footprint low

Memory & Speed





Performance



Future work

- Constraint Satisfaction Inference (CSI)
  - Integration of fertility model and “null model”
- Parameters
  - Beam in search
  - Classifier parameters (speed-accuracy trade-off)
  - Higher n in n-grams
- Reliance on word aligner
  - Test BerkeleyAligner

Thank you

http://ilk.uvt.nl/mbmt