Improving Coverage of Translation Memories with Language Modelling

Vít Baisa, Josef Bušta, Aleš Horák

Natural Language Processing Centre
Faculty of Informatics
Masaryk University

RASLAN 2014

Introduction

- Translation memories:
  - used in computer-aided translation systems,
  - manually built,
  - relatively small and focused,
  - usually in-house and not for (even academical) use.
- Our goal is to expand a TM to increase its coverage.
- We work with En$\leftrightarrow$Cs language pair.

Related work

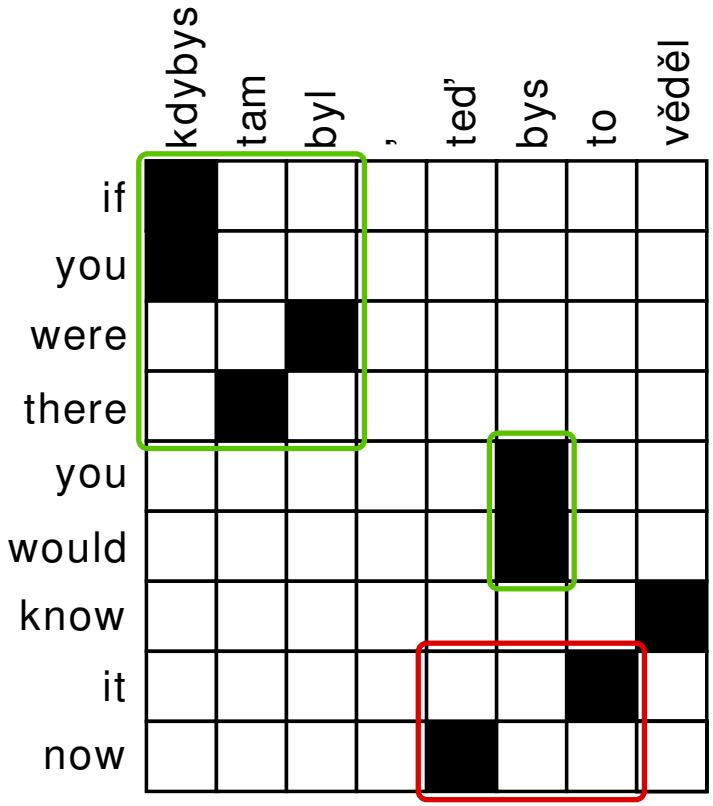
- TM are understudied resources,
- related topic: machine translation, example-based machine translation (EBMT),
- papers focused on searching and matching algorithms for CAT systems,
- *WeBiText*: TM from bilingual Canadian websites.
- *TransSearch*: EBMT system, Hansard corpus; linguistically motivated segments.

Methods for expanding TM

- Subsegment generation,
- subsegment combination,
- subsegment lexicalization,
- machine translation of subsegments.

Subsegment generating

- Use TM to train translational model (MGIZA),
- build word alignment for pairs from TM, it can be displayed in a word matrix,
- generate new subsegments and score them,
- resulting pairs can be added directly to TM or
- can be combined together.



Join & substitute

- we can make new segments by two methods
- **join** – we can concatenate several subsegments and check the result against a language model trained on large monolingual corpus
- **substitute** – sometimes a segment to be translated and a segment in a TM differ only in one or two words in the middle, in that case we can translate it using another subsegment but at the same time locate position of this in-the-middle words using the word matrix; these words may be translated automatically (using a dictionary, another subsegment from a TM) or simply left out for manual translation
- overlapping and non-overlapping **join** and **substitute** can be distinguished

Overlapping substitute example

new subsegment	Provozovatelé musí dodržovat zvláštní pravidla pro výzkumné
its translation	Operators shall comply with the special rules on research
from subsegments	Provozovatelé **musí** vytvářet **zvláštní** pravidla pro výzkumné | **musí** dodržovat **zvláštní**
their translations	Operators **shall** create **the special** rules on research | **shall** comply with **the special**

Subsegment lexicalizatoin

- generalization of the previous method
- all segments are tokenized and lemmatized
- searching and matching operations work on lemmata
- **ljoin** – concatenation of two different segments from *TM* and TM^{sub} on lemmata; when concatenating into new resulting segments, appropriate word form (case, gender and number) is generated
- **lsubstitute** – substitution of a part of target segment with another segment using lemmata

With this method we expect increasing the recall (coverage) but at the same time not decreasing the translation accuracy of original segments from *TM*. So it is partially rule-based method.

Machine translation of subsegments, example

A sentence from TM:
Návod na použití desinfekčního přípravku najdete na konci této brožury

A manual translation:
You can find instructions for use of disinfectant at the end of this brochure

A sentence for translation:
Návod na použití kartáče na vlasy najdete na konci této brožury

Not in TM: *kartáče na vlasy*
Google Translate returns: *hairbrush* (after lemmatization).

$\rightarrow$ Substitute the translation in the existing segment from TM.

Evaluation: subsegments generation & combination

We used a sample of TM and a testing document provided by a Czech translation services provider; as evaluation metrics we used the one used by MemoQ (CAT system).

	TM		TM^{sub}		TM^{NS}	
	Seg	**%**	**Seg**	**%**	**Seg**	**%**
matches	23	0.4	165	0.51	0	0
	TM^{OJ}		TM^{NJ}		TM^{all}	
	Seg	**%**	**Seg**	**%**	**Seg**	**%**
matches	6	0.07	4	0.05	182	**0.86**

TM^{OJ} – overlapping join

TM^{NJ} – non-overlapping join

TM^{sub} – generated subsegments

TM^{NS} – substitute

TM – translation memory

$TM^{all} = TM + TM^{sub} + TM^{NS} + TM^{OJ} + TM^{NJ}$

Evaluation: subsegments generation & combination

For an independent comparison, we also present our results for DGT translation memory; as evaluation metrics we used the one used by MemoQ (CAT system).

	TM		TM^{sub}		TM^{NS}	
	Seg	**%**	**Seg**	**%**	**Seg**	**%**
matches	31	0.03	276	0.25	58	0.45
	TM^{OJ}		TM^{NJ}		TM^{all}	
	Seg	**%**	**Seg**	**%**	**Seg**	**%**
matches	38	0.1	29	0.09	358	**0.46**

TM^{OJ} – overlapping join

TM^{NJ} – non-overlapping join

TM^{sub} – generated subsegments

TM^{NS} – substitute

TM – translation memory

$TM^{all} = TM + TM^{sub} + TM^{NS} + TM^{OJ} + TM^{NJ}$

Evaluation: METEOR score

	TM^{s}				
feature	TM^{sub}	TM^{OJ}	TM^{NJ}	TM^{NS}	TM^{all}
precision	0.60	0.63	0.70	0.66	0.61
recall	0.67	0.74	0.74	0.71	0.68
f1	0.64	0.68	0.72	0.68	0.64
METEOR score	0.31	0.37	**0.38**	**0.38**	0.31

	DGT-MT				
feature	TM^{sub}	TM^{OJ}	TM^{NJ}	TM^{NS}	TM^{all}
precision	0.76	0.93	0.91	0.81	0.80
recall	0.78	0.86	0.88	0.85	0.81
f1	0.77	0.89	0.89	0.83	0.81
METEOR score	0.40	0.50	**0.51**	0.45	0.43

Conclusion and the future

- The preliminary results are promising,
- we are working on improvement of the first two methods,
- the rest of methods will be implemented,
- we expect a higher coverage.
- More detailed evaluation using biggger data, comparison with other techniques.