Expanding Translation Memories: Proposal and Evaluation of Several Methods

Vít Baisa, Josef Bušta, Aleš Horák

Natural Language Processing Centre
Faculty of Informatics
Masaryk University

RASLAN 2013

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

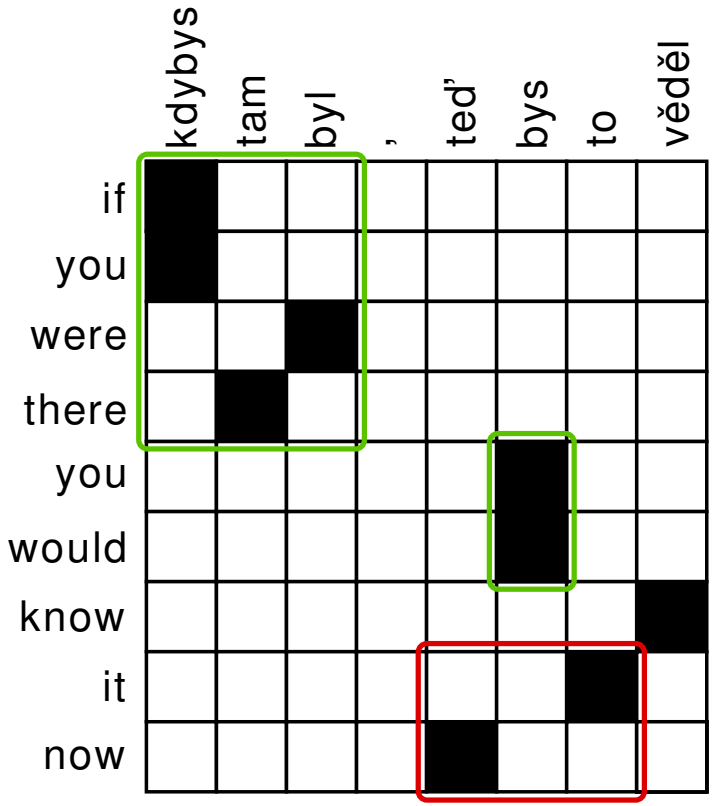
- TM are understudied resources,
- related topic: example-based machine translation (EBMT),
- papers focused on searching and matching algorithms for CAT systems,
- *WeBiText*: TM from bilingual Canadian websites.
- *TransSearch*: EBMT system, Hansard corpus; linguistically motivated segments.

Methods for expanding TM

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

Subsegment generating

- From parallel corpus (OPUS),
- train translational model (GIZA++),
- build word matrix for segments from a TM and
- generate new subsegments,
- resulting pairs can be added directly to TM or
- can be combined together.



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 $_{sub}$TM 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 MT:
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).

	^{s}TM		$_{sub}$TM		^{s}TM+$_{sub}$TM	
	Seg	**%**	**Seg**	**%**	**Seg**	**%**
matches	576	6.4	1175	**13.33**	1240	**15.64**

Subsegments only

	^{s}TM		$_{subjoin}$TM		^{s}TM+$_{subjoin}$TM	
	Seg	**%**	**Seg**	**%**	**Seg**	**%**
matches	576	6.4	1193	**14.03**	1254	**16.15**

Subsegments joined

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.