2010 Failures in Czech-English Phrase-Based MT

Ondřej Bojar, Kamil Kos; bojar@ufal.mff.cuni.cz, kamilkos@email.cz; Charles University in Prague, MFF, ÚFAL

Overview

Problems of English-to-Czech MT

- Large Target-Side Vocabulary
- Higher Out-of-Vocabulary Rate
- Low Reachability of Human Translations

Possible Solutions Examined

- Two-Step Translation
- Coarser Optimisation Metric
  - SemPOS (Kos and Bojar, 2009)

Two-Step Translation

Overcome Target-Side Data Sparseness

Moses run twice in a row:

1. Translate English to lemmatized Czech augmented to preserve important semantic properties known from the source phrase
2. Generate fully inflected Czech.

Src		after a sharp drop	
Mid	po+6	ASA1.prudký	NSA-.pokles
Gloss	after+voc	adj+sg…sharp	noun+sg…drop
Out	po	prudkém	poklesu

Simple vs. Two-Step Translation

Data Size		Simple		Two-Step		Change	
Parallel	**Mono**	**BLEU**	**SemPOS**	**BLEU**	**SemPOS**	**BLEU**	**SemPOS**
Small	**Small**	10,28±0,40	29,92	10,38±0,38	30,01	↗	↗
Small	**Large**	12,50±0,44	31,01	12,29±0,47	31,40	↘	↗
Large	**Large**	14,17±0,51	33,07	14,06±0,49	32,57	↘	↘

Minor gain in Small-Small, minor loss in Large-Large setting.
Interesting mixed result in Small-Large:

- Indicates that large LM can improve BLEU score without addressing the cross-lingual data sparseness (tackled by Two-Step model and appreciated by SemPOS).
- Note that large monolingual data were used also as the LM in the first step.

Manual Annotation

- 150 sentences manually annotated by two annotators (Small-Large setting).
- Each of them mildly prefers Two-Step model.
- Equal result (23) when limited to sentences where they agree.

	Two-Step	Both Fine	Both Wrong	Simple	Total
Two-Step Better	**23**	4	8	-	**35**
Both Fine	7	14	17	5	43
Both Wrong	8	1	28	2	39
Simple Better	-	3	7	**23**	33
Total	**38**	22	60	30	150

Rich Morphology

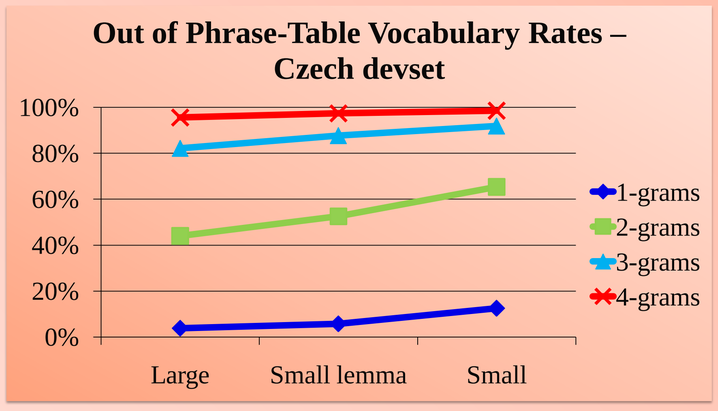
- Two training corpora:
  - Large – 7,5M sentences
  - Small – 126,1k sentences

Large Vocabulary

- Although the training data contains roughly the same number of Czech and English tokens, Czech vocabulary has many more entries.
- Czech vocabulary has approximately *double* the size compared to Czech lemmas.

	Large	Small
Sentences	7.5M	126.1k
Czech Tokens	79.2M	2.6M
English Tokens	89.1M	2.9M
Czech Vocabulary	923.1k	138.7k
English Vocabulary	646.3k	64.7k
Czech Lemmas	553.5k	60.3k
English Lemmas	611.4k	53.8k

Higher Out-of-Vocabulary Rates



- Out-of-Vocabulary (OOV) rate of unigrams and bigrams significantly increases for Small data setting.
- OOV can be reduced by lemmatizing word forms if we have only limited amount of data.
- Almost 12% of the Czech devset tokens are not available in the extracted phrase tables.

Lower Reachability of Reference Translations

- Percentage of reference translations reachable by exhaustive search (Schwartz, 2008).
- Exhaustive search influenced by:
  - *distortion limit* (default: 6),
  - *number of translation options* (default: 50).
- It is much harder to reach reference translation in Czech than in English.

		Distortion Limit		
	Topts	3	6	30
Czech	1	0,2%	0,3%	0,3%
	50	1,2%	**1,5%**	1,7%
	100	1,2%	1,5%	1,7%
English	1	0,4%	0,4%	0,4%
	50	4,9%	**6,7%**	8,6%
	100	5,3%	7,6%	9,4%

Optimisation towards Coarser Metric – SemPOS

SemPOS:

- Operates on lemmas of content words.
- Ignores word order.
- Reference lemmas matched only if semantic parts-of-speech (Hajič et al., 2004) agree.
- Czech and English supported so far.

Weights	BLEU	SemPOS
1:0	14,08±0,50	32,44
1:1	13,79±0,55	33,17

SemPOS-tuned parameters with MERT widely range in final quality.
(6.96 – 9.46 BLEU for Small data)

Linear combination of BLEU and SemPOS

Large data

Small data

Weights	BLEU	SemPOS
1:0	10,42±0,38	29,91
1:1	10,15±0,39	29,81
1:1	9,42±0,37	29,30
2:1	10,37±0,38	29,95
3:1	10,30±0,39	30,03
10:1	10,17±0,40	29,58
1:2	10,11±0,38	29,80
1:10	9,44±0,40	29,74

Our WMT10 System Configuration

English-to-Czech

- Standard GIZA++ word alignment based on both source and target lemmas.
- Two alternative decoding paths; forms always truecased:
  form+tag → form
  form → form.
  The first path is more specific and helps to preserve core syntactic elements in the sentence.
  Without the tag, ambiguous English words could often all translate as e.g. nouns, leading to no verb in the Czech sentence. The default path serves as a back-off.
- Significance filtering of the phrase tables (Johnson et al.,2007) implemented for Moses by Chris Dyer; default settings of filter value a+e and the cut-off 30.
- Lexicalized reordering (or-bi-fe) based on forms.
- Two separate 5-gram Czech LMs of truecased forms each of which interpolates models trained on the following datasets; the interpolation weights were set automatically using SRILM (Stolcke, 2002) based on the target side of the development set:
  - Interpolated CzEng domains: news, web, fiction. The rationale behind the selection of the domains is that we prefer prose-like texts for LM estimation (and not e.g. technical documentation) while we want as much parallel data as possible.
  - Interpolated monolingual corpora: WMT09 monolingual, WMT10 monolingual, Czech National Corpus (Kocek et al., 2000) sections SYN2000+2005+2006PUB.
- Standard Moses MERT towards BLEU.

Czech-to-English

Far fewer configurations tested, this is the final one:

- Two alternative decoding paths; forms always truecased:
  - form → form
  - lemma → form.
- Significance filtering as in English-to-Czech.
- 5-gram English LM based on CzEng English side only.
- Lexicalized reordering (or-bi-fe) based on forms.
- Standard Moses MERT towards BLEU.

Using Gigaword LM as compiled by Chris Callison-Burch caused a significant loss in quality, probably due to different tokenization rules.

Full text, acknowledgement and the list of references in the proceedings of ACL 2010 Workshop on Statistical Machine Translation (WMT10).