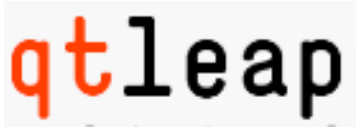
Factored Models for Deep Machine Translation

Kiril Simov, Iliana Simova, Velislava Todorova, Petya Osenova

Institute for Information and Communication Technology
Bulgarian Academy of Sciences

Deep Machine Translation Workshop 2015, Prague, 3-4 September 2015



Plan of the Talk

- Preliminary notes
- Related work
- Data preparation
- Linguistic Preprocessing and Factor-based SMT Model
- Experiments
- Conclusions

Preliminary Notes

- We focus on the *Bulgarian-to-English* and *English-to-Bulgarian* translation
- We build on the SMT baseline, which is already augmented with linguistic features (POS, grammatical features) in an entry-deep level of experiments
- We explore the impact of the bilingual morphological lexicons in the translation process

Related Work

- Our work is closely connected to the transfer-based MT models
  - Deep linguistic grammars in DELPH-IN infrastructure, which are in HPSG framework and uses MRS. The transfer in this setting is usually implemented in the form of rewriting rules
  - TectoMT approach. It is the more abstract level of language representation, which then is used for the transfer step within the MT systems

Data Preparation

- Two types of data are used in our experiments:
  - parallel news and other domain data, which is the training data (SETIMES, EuroParl and LibreOffice Document Foundation)
  - parallel QTLeap data in the IT domain, which is the training and test data (QTLeap corpus is composed of 4 000 pairs of questions and respective answers in the domain of ICT troubleshooting for both hardware and software)

Linguistic Preprocessing and Factor-based SMT Model (1)

- The data in the training datasets was analyzed at two levels – POS tagging and Lemmatization
- We built our approach on top of the factor-based SMT model proposed by Koehn and Hoang (2007a), as an extension of the traditional phrase-based SMT framework
- The process is quite similar to supertagging (Bangalore and Joshi, 1999), which assigns "rich descriptions (supertags) that impose complex constraints in a local context".

Linguistic Preprocessing and Factor-based SMT Model (2)

- The morphosyntactic factors for both languages:
  - WF - word form, which is the original text token.
  - LEMMA - the lexical invariant of the original word form.
  - POS - part-of-speech of the word.
  - LING - other linguistic features derived from the POS tag in the BulTreeBank tagset

Linguistic Preprocessing and Factor-based SMT Model (3)

- In comparison to the experiments described in ((Wang et al., 2012a), (Wang et al., 2012b)) the number of the linguistic factors were reduced to the ones that contributed best to the improvement of the translation results.
- Thus, we have excluded all the factors based on dependency parsing of the data.

Linguistic Preprocessing and Factor-based SMT Model (4)

- Similarly to our previous experiments, here we use only the RMRS relation and the type of the main argument as features to the translation model:
  - EP – the name of the elementary predicate, which usually indicates an event or an entity from a semantic point of view.
  - EOV – indicates the current EP as either an event, a reference variable or their subtypes (now with subtypes for various POS!).

Entry Level Deep Experiments (1)

- We make use of the Moses open source toolkit to build a factored SMT model (Koehn and Hoang, 2007b)

- The best performing model featuring a semantic factor for the direction **BG->EN** includes four factors: *word form, lemma, POS* and *variable type*; a word and POS based language model

Entry Level Deep Experiments (2)

- For the translation direction **EN->BG** the model includes three factors: *word form*, *part of speech*, and *variable type*. In the translation step, the source word, POS, and variable type are translated into the target word form.

Factors	LM	Translation	Generation	Decoding	BLEU	
					BG→EN	EN→BG
WF, EP, EoV	0	0,1,2-0	–	–	31.53	24.00
WF, POS, EoV	0	0,1,2-0	–	–	32.07	24.13
WF, LEMMA, EP, EoV	0	1-1+2-2+3-3	1,2,3-0	–	23.94	13.69
WF, LEMMA, POS	0,2	0-0,2+1-0,2	–	t0:t1	32.59	22.86
WF, LEMMA, POS, LING	0,2	1-1+3-2+0-0,2	1-2+1,2-0	t0,g0,t1,g1:t2	32.78	22.73
WF, LEMMA, POS, EoV	0,2	0,3-0,2+1,3-0,2	–	t0:t1	32.59	22.77

Table 1: A subset of the results from the factored experiments, evaluated on the second half of the QTLeap data set.

Results for the Baseline and the Entry-Deep Translation

The BG-to-EN direction was evaluated on questions. Here are the numbers for Pilot 0 and Pilot 1:

- 1. BLEU Pilot 0 (29.7); Pilot 1 (27.7)
- 2. wordF Pilot 0 (22.8); Pilot 1 (22.4)
- 3. chartF Pilot 0 (46.7); Pilot 1 (**47.4**)

The EN-to-BG direction was evaluated on the answers:

- 1. BLEU Pilot 0 (25.3); Pilot 1 (24.5)
- 2. wordF Pilot 0 (25.6); Pilot 1 (25.0)
- 3. chartF Pilot 0 (46.7); Pilot 1 (46.6)

Preliminary Experiments with a Parallel Morphological Lexicon

- One of the main problems in the translation in both directions are the so-called out-of-training word forms
- Thus, a morphological lexicon was constructed:
  - BTB-Morphological lexicon containing all wordforms for more than 110 000 Bulgarian lemmas
  - BTB-bilingual Bulgarian-English lexicons (with about 8000 entries)
  - English Wiktionary

Preparation Steps (1)

- From Wiktionary the English wordforms were extracted for the English lemmas
- The wordform lexicons for both languages were mapped to the corresponding part of the bilingual lexicon
- The corresponding wordforms were aligned on the basis of their morphological features like *number* (singular, plural); *degree* (comparative, superlative); *definiteness* (definite, indefinite)

Preparation Steps (1)

- The lexicon represents more than 70 000 aligned wordforms
- Only *the noun* and *the adjective* parts-of-speech from the wordform aligned bilingual lexicon were used

Example

Bulgarian	English
visok|visok|a	a|a|d high|high|g
visok|visok|a	high|high|g
visok|visok|a	a|a|d tall|tall|g
visok|visok|a	tall|tall|g
—	—
naj-visokata|visok|a	highest|highest|g
naj-visokata|visok|a	the|the|d highest|highest|g
naj-visokata|visok|a	tallest|tallest|g
naj-visokata|visok|a	the|the|d tallest|tallest|g

Results with Morphological Lexicon

	without lexicon	with lexicon; with only indefinite forms	with lexicon; with all forms
BG→EN	32.59	33.02	32.88
EN→BG	22.86	23.91	22.97

Table 3: Preliminary experiments with parallel morphological lexicons.

Conclusions and Outlook (1)

- We reported our initial work towards building deep statistical machine translation models between Bulgarian and English in both directions.
- The paper showed that the addition of a wordform aligned parallel lexicon improved the results in both translation directions.

Conclusions and Outlook (2)

- In entry-level deep experiments we extended the semantic factors with new types of main arguments for MRS elementary predicates, which improved the results in *English-to-Bulgarian* direction and shows promising results for the *Bulgarian-to-English* direction
- Directions for improvement: incorporation also of other parts-of-speech, compositional and multiword phrases.

Thank you!

qtleap