Forms Wanted: Training SMT on Monolingual Data*

Ondřej Bojar and Aleš Tamchyna
bojar@ufal.mff.cuni.cz, a.tamchyna@gmail.com
Charles University in Prague, Institute of Formal and Applied Linguistics

Abstract

We propose and evaluate a simple technique of "reverse self-training" for statistical machine translation. The technique allows to extend target-side vocabulary of the MT system using target-side monolingual data and it is especially aimed at translation to morphologically rich languages.

1 Introduction

Machine translation to morphologically rich languages such as Czech faces a severe problem with target-side vocabulary. Statistical approaches such as phrase-based translation (SMT) so far are unable to produce forms never seen in parallel data. The baseline setup has no such generational capacity at all, and only very limited or disputable success was observed with factored translation as implemented in the Moses translation system (Bojar and Kos, 2010; Bojar et al., 2009), although some empirical methods were already proposed (de Gispert et al., 2005).

We propose a rather simple but effective approach to extend target-side vocabulary using target-side monolingual data. We call our approach "reverse self-training" for a similarity to self-training techniques (Ueffing et al., 2007). Experiments similar to ours were also conducted by Bertoldi and Federico (2009) for domain adaptation without specific aim at target-side vocabulary.

Given a baseline parallel corpus, we train a factored SMT system in the reverse direction, translate a large (ultimately target-side) monolingual corpus using this system "back" to the source language and add the output to our parallel data.[1] Unlike in self-training, there is no urge for the filtration of the MT-generated parallel corpus, because its target side is known to be correct text.

1.1 Learning to Use New Word Forms

Regular self-training helps MT because it can provide the system with new output phrases composed of known word forms. We set up out reverse self-training so that we can actually learn to produce new word forms, i.e. word forms never seen in the original parallel data. We achieve this by ensuring that the reverse MT system attempts to translate also unseen word forms (these will become the newly learned target word forms).

So far, we experimented only with using word lemmas as the fall-back for unseen word forms, but many other options are conceivable and needed. Specifically, we use Moses alternative decoding paths as developed by Birch et al. (2007) to translate either from the form or the lemma, whichever scores better.[2]

2 Experimental Results for English-to-Czech Translation

We use "the standard Moses pipeline" for our experiments, i.e. simple phrase-based translation using heuristically extracted phrases based on GIZA++ word alignments. Only the reverse translation uses the two source factors as described.

Table 1(a) documents the gradual gain in BLEU scores by various combinations of the baseline 126k parallel sentences (the news section of CzEng 0.9, (Bojar and Žabokrtský, 2009)) and 2M sentences from the WMT10 monolingual Czech news[3]. We tune our model on WMT08 test set and evaluate on WMT09 test set, all in the news domain.

*The work on this project was supported by the grants EuroMatrixPlus (FP7-ICT-2007-3-231720 of the EU and 7E09003 of the Czech Republic), GAČR P406/10/P259, and MSM 0021620838.

[1] We re-align this new corpus for the time being but we believe both the quality and efficientcy can be improved by using the alignments as produced by the MT system.

[2] We did not correct the scoring of phrases available in one corpus only, as noted by Bertoldi and Federico (2009), but we are planning to correct this issue as well.

[3] http://www.statmt.org/wmt10/translation-task.html#download

Table 1: BLEU scores (a) and a preliminary manual evaluation (b) when training on 126k parallel sentences or when also using 2M target-side monolingual sentences. Simple corpus concatenation is denoted as ".", interpolation in MERT is denoted as "+".

BLEU	TM	LM
10.56±0.39	para	para
10.70±0.40	mono	mono
10.98±0.38	mono	para+mono
11.06±0.40	mono	para.mono
12.20±0.40	para	para+mono
12.24±0.44	para	para.mono
12.27±0.41	para.mono	para+mono
12.33±0.43	para.mono	para.mono
12.65±0.42	para+mono	para.mono

	Baseline	TM para.mono
	12.24±0.44	12.33±0.43
One system is better	19	29
Equally fine	6	
Equally wrong	46	
	Baseline	TM para+mono
	12.24±0.44	12.65±0.42
One system is better	27	35
Equally fine	10	
Equally wrong	28	

The use of monolingual data only in the language model (LM) already significantly increases the performance (from 10.56±0.39 to 12.24±0.44). A further increase to 12.65±0.42 is achieved with our reverse self-training approach which incorporates the 2M sentences in the translation model (TM) as well. Note that for a significant increase in the BLEU score, it was essential to supply the additional training data as an independent phrase table to let the MERT procedure find a proper balance of translation model weights. For the language model, we observe a little loss if we use MERT to balance the two LMs.

We ran two independent small manual evaluations (blindly) comparing random 100 sentences produced by the "clever baseline" system 12.24±0.44 and two variants of the reverse self-training. In both cases, we confirm the improvement in translation quality.

3 Conclusion and Future Research

The technique of reverse self-training proved helpful in a small data setting. The utility of the same approach with larger parallel data available has yet to be investigated. Similarly, we will explore various back-off options in the reverse translation (e.g. lemmatization, simple stemming, synonyms, or no back-off at all) and their impact on the forward translation performance.

So far, we have tested our approach only on English-to-Czech translation. We will soon apply it to other language pairs. We expect gains for highly inflected languages where the reverse translation can be relatively easily backed off by lemmas or stems. Languages with agglutinative properties or word formation by composition will be harder to tackle, because the word form is not recognized even when treated as source word and back-off techniques are not that straightforward.

4 References

Nicola Bertoldi and Marcello Federico. 2009. Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 182–189, Athens, Greece, March. Association for Computational Linguistics.

Alexandra Birch, Miles Osborne, and Philipp Koehn. 2007. CCG Supertags in Factored Statistical Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 9–16, Prague, Czech Republic, June. Association for Computational Linguistics.

Ondrej Bojar and Kamil Kos. 2010. 2010 Failures in English-Czech Phrase-Based MT. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 60–66, Uppsala, Sweden, July. Association for Computational Linguistics.

Ondřej Bojar and Zdeněk Žabokrtský. 2009. CzEng 0.9: Large Parallel Treebank with Rich Annotation. *Prague Bulletin of Mathematical Linguistics*, 92:63–83.

Ondřej Bojar, David Mareček, Václav Novák, Martin Popel, Jan Ptáček, Jan Rouš, and Zdeněk Žabokrtský. 2009. English-Czech MT in 2008. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, March. Association for Computational Linguistics.

Adrià de Gispert, José B. Mariño, and Josep M. Crego. 2005. Improving statistical machine translation by classifying and generalizing inflected verb forms. In *Eurospeech 2005*, pages 3185–3188, Lisbon, Portugal.

Nicola Ueffing, Gholamreza Haffari, and Anoop Sarkar. 2007. Semi-supervised model adaptation for statistical machine translation. *Machine Translation*, 21(2):77–94.