Improving Translation Model by Monolingual Data[*]

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

Abstract

We use target-side monolingual data to extend the vocabulary of the translation model in statistical machine translation. This method called "reverse self-training" improves the decoder's ability to produce grammatically correct translations into languages with morphology richer than the source language esp. in small-data setting. We empirically evaluate the gains for several pairs of European languages and discuss some approaches of the underlying back-off techniques needed to translate unseen forms of known words. We also provide a description of the systems we submitted to WMT11 Shared Task.

1 Introduction

Like any other statistical NLP task, SMT relies on sizable language data for training. However the parallel data required for MT are a very scarce resource, making it difficult to train MT systems of decent quality. On the other hand, it is usually possible to obtain large amounts of monolingual data.

In this paper, we attempt to make use of the monolingual data to reduce the sparseness of surface forms, an issue typical for morphologically rich languages. When MT systems translate into such languages, the limited size of parallel data often causes the situation where the output should include a word form never observed in the training data. Even though the parallel data do contain the desired word in other forms, a standard phrase-based decoder has no way of using it to generate the correct translation.

Reverse self-training addresses this problem by incorporating the available monolingual data in the translation model. This paper builds upon the idea outlined in Bojar and Tamchyna (2011), describing how this technique was incorporated in the WMT Shared Task and extending the experimental evaluation of reverse self-training in several directions – the examined language pairs (Section 4.2), data size (Section 4.3) and back-off techniques (Section 4.4).

2 Related Work

The idea of using monolingual data for improving the translation model has been explored in several previous works. Bertoldi and Federico (2009) used monolingual data for adapting existing translation models to translation of data from different domains. In their experiments, the most effective approach was to train a new translation model from "fake" parallel data consisting of target-side monolingual data and their machine translation into the source language by a baseline system.

Ueffing et al. (2007) used a boot-strapping technique to extend translation models using monolingual data. They gradually translated additional source-side sentences and selectively incorporated them and their translations in the model.

Our technique also bears a similarity to de Gispert et al. (2005), in that we try to use a back-off for surface forms to generalize our model and produce translations with word forms never seen in the original parallel data. However, instead of a rule-based approach, we take advantage of the available

[*] This work has been supported by the grants EuroMatrixPlus (FP7-ICT-2007-3-231720 of the EU and 7E09003 of the Czech Republic), P406/10/P259, and MSM 0021620838.

Proceedings of the 6th Workshop on Statistical Machine Translation, pages 330–336,
Edinburgh, Scotland, UK, July 30–31, 2011. ©2011 Association for Computational Linguistics

	Source English		Target Czech	Czech Lemmatized
Parallel (small)	a cat chased…	=	**kočka** honila…	*kočka honit…*
	I saw a cat	=	viděl jsem **kočku**	*vidět být kočka*
	I read about a dog	=	četl jsem o psovi	*číst být o pes*
Monolingual (large)	?		četl jsem o **kočce**	*číst být o kočka*
	I read about a cat	←	Use reverse translation backed-off by lemmas.	

Figure 1: The essence of reverse self-training: a new phrase pair ("about a cat" = "o **kočce**") is learned based on a small parallel corpus and large target-side monolingual texts.

data and learn these forms statistically. We are therefore not limited to verbs, but our system is only able to generate surface forms observed in the target-side monolingual data.

3 Reverse Self-Training

Figure 1 illustrates the core of the method. Using available parallel data, we first train an MT system to translate from the target to the source language. Since we want to gather new word forms from the monolingual data, this reverse model needs the ability to translate them. For that purpose we use a factored translation model (Koehn and Hoang, 2007) with two alternative decoding paths: form→form and back-off→form. We experimented with several options for the back-off (simple stemming by truncation or full lemmatization), see Section 4.4. The decoder can thus use a less sparse representation of words if their exact forms are not available in the parallel data.

We use this reverse model to translate (much larger) target-side monolingual data into the source language. We preserve the word alignments of the phrases as used in the decoding so we directly obtain the word alignment in the new "parallel" corpus. This gives us enough information to proceed with the standard MT system training – we extract and score the phrases consistent with the constructed word alignment and create the phrase table.

We combine this enlarged translation model with a model trained on the true parallel data and use Minimum Error Rate Training (Och, 2003) to find the balance between the two models. The final model has four separate components – two language models (one trained on parallel and one on monolingual data) and the two translation models.

We do not expect the translation quality to improve simply because more data is included in training – by adding translations generated using known data, the model could gain only new combinations of known words. However, by using a back-off to less sparse units (e.g. lemmas) in the factored target→source translation, we enable the decoder to produce previously unseen surface forms. These translations are then included in the model, reducing the data sparseness of the target-side surface forms.

4 Experiments

We used common tools for phrase-based translation – Moses (Koehn et al., 2007) decoder and tools, SRILM (Stolcke, 2002) and KenLM (Heafield, 2011) for language modelling and GIZA++ (Och and Ney, 2000) for word alignments.

For reverse self-training, we needed Moses to also output word alignments between source sentences and their translations. As we were not able to make the existing version of this feature work, we added a new option and re-implemented this funcionality.

We rely on automatic translation quality evaluation throughout our paper, namely the well-established BLEU metric (Papineni et al., 2002). We estimate 95% confidence bounds for the scores as described in Koehn (2004). We evaluated our translations on lower-cased sentences.

4.1 Data Sources

Aside from the WMT 2011 Translation Task data, we also used several additional data sources for the experiments aimed at evaluating various aspects of reverse self-training.

JRC-Acquis

We used the JRC-Acquis 3.0 corpus (Steinberger et al., 2006) mainly because of the number of available languages. This corpus contains a large amount

Source	Target	Corpus Size (k sents)		Vocabulary Size Ratio	Baseline	+Mono LM	+Mono TM
		Para	Mono				
English	Czech	94	662	1.67	40.9±1.9	43.5±2.0	*44.3±2.0
English	Finnish	123	863	2.81	27.0±1.9	27.6±1.8	28.3±1.7
English	German	127	889	1.83	34.8±1.8	36.4±1.8	37.6±1.8
English	Slovak	109	763	2.03	35.3±1.6	37.3±1.7	37.7±1.8
French	Czech	95	665	1.43	39.9±1.9	42.5±1.8	43.1±1.8
French	Finnish	125	875	2.45	26.7±1.8	27.8±1.7	28.3±1.8
French	German	128	896	1.58	38.5±1.8	40.2±1.8	*40.5±1.8
German	Czech	95	665	0.91	35.2±1.8	37.0±1.9	*37.3±1.9

Table 1: BLEU scores of European language pairs on JRC data. Asterisks in the last column mark experiments for which MERT had to be re-run.

of legislative texts of the European Union. The fact that all data in the corpus come from a single, very narrow domain has two effects – models trained on this corpus perform mostly very well in that domain (as documented e.g. in Koehn et al. (2009)), but fail when translating ordinary texts such as news or fiction. Sentences in this corpus also tend to be rather long (e.g. 30 words on average for English).

CzEng

CzEng 0.9 (Bojar and Žabokrtský, 2009) is a parallel richly annotated Czech-English corpus. It contains roughly 8 million parallel sentences from a variety of domains, including European regulations (about 34% of tokens), fiction (15%), news (3%), technical texts (10%) and unofficial movie subtitles (27%). We do not make much use of the rich annotation in this paper, however we did experiment with using Czech lemmas (included in the annotation) as the back-off factor for reverse self-training.

4.2 Comparison Across Languages

In order to determine how successful our approach is across languages, we experimented with Czech, Finnish, German and Slovak as target languages. All of them have a rich morphology in some sense. We limited our selection of source languages to English, French and German because our method focuses on the target language anyway. We did however combine the languages with respect to the richness of their vocabulary – the source language has less word forms in almost all cases.

Czech and Slovak are very close languages, sharing a large portion of vocabulary and having a very similar grammar. There are many inflectional rules for verbs, nouns, adjectives, pronouns and numerals. Sentence structure is exhibited by various agreement rules which often apply over long distance. Most of the issues commonly associated with rich morphology are clearly observable in these languages.

German also has some inflection, albeit much less complex. The main source of German vocabulary size are the compound words. Finnish serves as an example of agglutinative languages well-known for the abundance of word forms.

Table 1 contains the summary of our experimental results. Here, only the JRC-Acquis corpus was used for training, development and evaluation. For every language pair, we extracted the first 10 percent of the parallel corpus and used them as the parallel data. The last 70 percent of the same corpus were our “monolingual” data. We used a separate set of 1000 sentences for the development and another 1000 for testing.

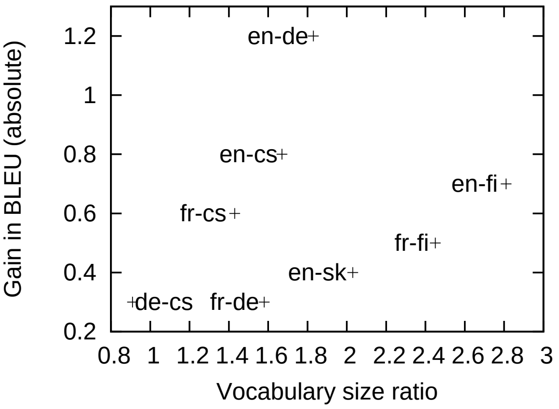
Sentence counts of the corpora are shown in the columns Corpus Size Para and Mono. The table also shows the ratio between observed vocabulary size of the target and source language. Except for the German→Czech language pair, the ratios are higher than 1. The Baseline column contains the BLEU score of a system trained solely on the parallel data (i.e. the first 10 percent). A 5-gram language model was used. The “+Mono LM” scores were achieved by adding a 5-gram language model trained on the monolingual data as a separate component (its weight was determined by MERT). The last column contains the scores after adding the translation model self-trained on target monolingual data. This model was also added as another component and the weights associated with it were found by MERT.

For the back-off in the reverse self-training, we used a simple suffix-trimming heuristic suitable for fusional languages: cut off the last three characters of each word always keeping at least the first three characters. This heuristic reduces the vocabulary size to a half for Czech and Slovak but it is much less effective for Finish and German (Table 2), as can be expected from their linguistic properties.

Language	Vocabulary reduced to (%)
Czech	52
Finnish	64
German	73
Slovak	51

Table 2: Reduction of vocabulary size by suffix trimming

Figure 2: Vocabulary ratio and BLEU score



We did not use any linguistic tools, such as morphological analyzers, in this set of experiments. We see the main point of this section in illustrating the applicability of our technique on a wide range of languages, including languages for which such tools are not available.

We encountered problems when using MERT to balance the weights of the four model components. Our model consisted of 14 features – one for each language model, five for each translation model (phrase probability and lexical weight for both directions and phrase penalty), word penalty and distortion penalty. The extra 5 weights of the reversely trained translation model caused MERT to diverge in some cases. Since we used the `mert-moses.pl` script for tuning and kept the default parameters, MERT ran for 25 iterations and stopped. As a result, even though our method seemed to improve translation performance in most language pairs, several experiments contradicted this observation. We simply reran the final tuning procedure in these cases and were able to achieve an improvement in BLEU as well. These language pairs are marked with a '*' sign in Table 1.

A possible explanation for this behaviour of MERT is that the alternative decoding paths add a lot of possible derivations that generate the same string. To validate our hypothesis we examined a diverging run of MERT for English→Czech translation with two translation models. Our n-best lists contained the best 100 derivations for each translated sentence from the development data. On average (over all 1000 sentences and over all runs), the n-best list only contained 6.13 different translations of a sentence. The result of the same calculation applied on the baseline run of MERT (which converged in 9 iterations) was 34.85 hypotheses. This clear disproportion shows that MERT had much less information when optimizing our model.

Overall, reverse self-training seems helpful for translating into morphologically rich languages. We achieved promising gains in BLEU, even over the baseline including a language model trained on the monolingual data. The improvement ranges from roughly 0.3 (e.g. German→Czech) to over 1 point (English→German) absolute. This result also indicates that suffix trimming is a quite robust heuristic, useful for a variety of language types.

Figure 2 illustrates the relationship between vocabulary size ratio of the language pair and the improvement in translation quality. Although the points are distributed quite irregularly, a certain tendency towards higher gains with higher ratios is observable. We assume that reverse self-training is most useful in cases where a single word form in the source language can be translated as several forms in the target language. A higher ratio between vocabulary sizes suggests that these cases happen more often, thus providing more space for improvement using our method.

4.3 Data Sizes

We conducted a series of English-to-Czech experiments with fixed parallel data and a varying size of monolingual data. We used the CzEng corpus, 500 thousand parallel sentences and from 500 thousand up to 5 million monolingual sentences. We used two separate sets of 1000 sentences from CzEng for development and evaluation. Our results are summarized in Figure 3. The gains in BLEU become more significant as the size of included monolingual data increases. The highest improvement can be observed when the data are largest – over 3 points absolute. Figure 4 shows an example of the impact on translation quality – the "Mono" data are 5 million sentences.

When evaluated from this point of view, our method can also be seen as a way of considerably improving translation quality for languages with little available parallel data.

We also experimented with varying size of parallel data (500 thousand to 5 million sentences) and its effect on reverse self-training contribution. The size of monolingual data was always 5 million sentences. We first measured the percentage of test data word forms covered by the training data. We calculated the value for parallel data and for the combination of parallel and monolingual data. For word forms that appeared only in the monolingual data, a different form of the word had to be contained in the parallel data (so that the model can learn it through the back-off heuristic) in order to be counted in. The difference between the first and second value can simply be thought of as the upper-bound estimation of reverse self-training contribution. Figure 5 shows the results along with BLEU scores achieved in translation experiments following this scenario.

Our technique has much greater effect for small parallel data sizes; the amount of newly learned word forms declines rapidly as the size grows. Similarly, improvement in BLEU score decreases quickly and becomes negligible around 2 million parallel sentences.

4.4 Back-off Techniques

We experimented with several options for the back-off factor in English→Czech translation. Data from training section of CzEng were used, 1 million parallel sentences and another 5 million sentences as target-side monolingual data. As in the previous section, the sizes of our development and evaluation sets were a thousand sentences.

Figure 3: Relation between monolingual data size and gains in BLEU score

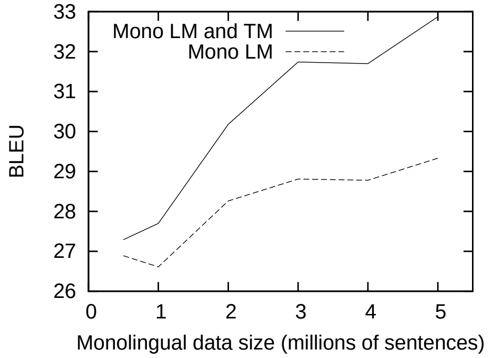
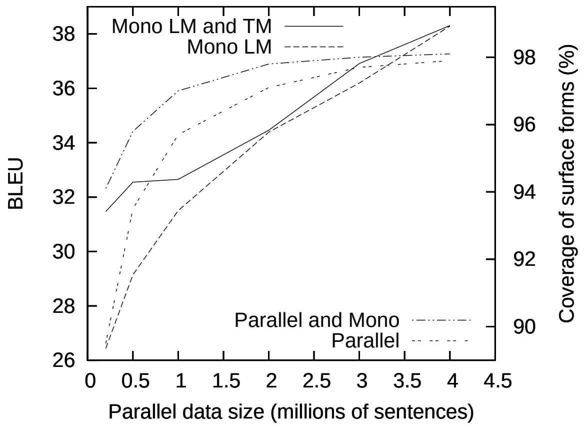


Figure 5: Varying parallel data size, surface form coverage ("Parallel", "Parallel and Mono") and BLEU score ("Mono LM", "Mono LM and TM")



CzEng annotation contains lexically disambiguated word lemmas, an appealing option for our purposes. We also tried trimming the last 3 characters of each word, keeping at least the first 3 characters intact. Stemming of each word to four characters was also evaluated (Stem-4).

Table 3 summarizes our results. The last column shows the vocabulary size compared to original vocabulary size, estimated on lower-cased words.

We are not surprised by stemming performing the

System	Translation	Gloss
Baseline	Jsi tak zrcadla?	Are you$_{SG}$ so mirrors? (ungrammatical)
+Mono LM	Jsi neobjednávejte zrcadla?	Did you$_{SG}$ don't order$_{PL}$ mirrors? (ungrammatical)
+Mono TM	Už sis objednal zrcadla?	Have you$_{SG}$ ordered$_{SG}$ the mirrors (for yourself) yet?

Figure 4: Translation of the sentence "Did you order the mirrors?" by baseline systems and a reversely-trained system. Only the last one is able to generate the correct form of the word "order".

worst – the equivalence classes generated by this simple heuristic are too broad. Using lemmas seems optimal from the linguistic point of view, however suffix trimming outperformed this approach in our experiments. We feel that finding well-performing back-off techniques for other languages merits further research.

Back-off	BLEU	Vocabulary Size (%)
Baseline	31.82±3.24	100
Stem-4	32.73±3.19	19
Lemma	33.05±3.40	54
Trimmed Suffix	**33.28±3.32**	47

Table 3: Back-off BLEU scores comparison

4.5 WMT Systems

We submitted systems that used reverse self-training (`cu-tamchyna`) for English→Czech and English→German language pairs.

Our parallel data for German were constrained to the provided set (1.9 million sentences). For Czech, we used the training sections of CzEng and the supplied WMT11 News Commentary data (7.3 million sentences in total).

In case of German, we only used the supplied monolingual data, for Czech we used a large collection of texts for language modelling (i.e. unconstrained). The reverse self-training used only the constrained data – 2.3 million sentences in German and 2.2 in Czech. In case of Czech, we only used the News monolingual data from 2010 and 2011 for reverse self-training – we expected that recent data from the same domain as the test set would improve translation performance the most.

We achieved mixed results with these systems – for translation into German, reverse self-training did not improve translation performance. For Czech, we were able to achieve a small gain, even though the reversely translated data contained less sentences than the parallel data. Our BLEU scores were also affected by submitting translation outputs without normalized punctuation and with a slightly different tokenization.

In this scenario, a lot of parallel data were available and we did not manage to prepare a reversely trained model from larger monolingual data. Both of these factors contributed to the inconclusive results.

Table 4 shows case-insensitive BLEU scores as calculated in the official evaluation.

Target Language	Mono LM	+Mono TM
German	14.8	14.8
Czech	15.7	15.9

Table 4: Case-insensitive BLEU of WMT systems

5 Conclusion

We introduced a technique for exploiting monolingual data to improve the quality of translation into morphologically rich languages.

We carried out experiments showing improvements in BLEU when using our method for translating into Czech, Finnish, German and Slovak with small parallel data. We discussed the issues of including similar translation models as separate components in MERT.

We showed that gains in BLEU score increase with growing size of monolingual data. On the other hand, growing parallel data size diminishes the effect of our method quite rapidly. We also documented our experiments with several back-off techniques for English to Czech translation.

Finally, we described our primary submissions to the WMT 2011 Shared Translation Task.

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.

Ondřej Bojar and Aleš Tamchyna. 2011. Forms Wanted: Training SMT on Monolingual Data. Abstract at Machine Translation and Morphologically-Rich Languages. Research Workshop of the Israel Science Foundation University of Haifa, Israel, January.

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

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.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh, UK, July. Association for Computational Linguistics.

Philipp Koehn and Hieu Hoang. 2007. Factored Translation Models. In *Proc. of EMNLP*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL*. The Association for Computer Linguistics.

Philipp Koehn, Alexandra Birch, and Ralf Steinberger. 2009. 462 machine translation systems for europe. In *MT Summit XII*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *EMNLP*, pages 388–395. ACL.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. pages 440–447, Hongkong, China, October.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *ACL*, pages 160–167.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Daniel Varga. 2006. The JRC-acquis: A multilingual aligned parallel corpus with 20+ languages. *CoRR*, abs/cs/0609058. informal publication.

Andreas Stolcke. 2002. Srilm — an extensible language modeling toolkit, June 06.

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