Extracting features from text to improve statistical machine translation

A. P. Molchanov

Abstract. In this paper we investigate the technique of extending the Moses Statistical Machine Translation (SMT) system default set of features using shallow linguistic information from source and target phrases. Although a typical SMT system uses a phrase table with 5 default features, most systems are scalable and support any number of additional features. We assume that linguistic information extracted from the source and target phrases can improve the overall translation quality, i.e. make the system more robust and reduce the number of instances of incorrect word choice, punctuation mistakes and other problems SMT systems are prone to. First, we build a baseline SMT system. Then we extract shallow linguistic features directly from source and target phrases of the baseline system’s phrase table. The features are precomputed and stored in the phrase table, so they can be regarded as stateless dense features. We develop and examine 19 features incorporating information from source and target phrases. We explore features commonly used in monolingual and parallel data filtering techniques. The features we investigate include source and target phrase lengths, word, number and punctuation symbol count, word frequencies according to large monolingual corpora etc. For each feature, we build and evaluate a separate SMT system. We conduct a series of experiments on the English-Russian language pair and obtain statistically significant improvements of up to 0.4 BLEU compared to baseline configuration.

Keywords: machine translation, statistical machine translation, SMT, Moses, feature extraction, translation quality.

Introduction

Most modern SMT systems use a phrase-table with 5 common features: forward and backward phrase probability, forward and backward lexical weight and phrase penalty, the latter usually being a constant. The Moses (Koehn et al. 2007) SMT decoder architecture is scalable and supports an unlimited number of additional features. In this paper, we investigate how shallow linguistic information from source and target phrases can improve overall translation quality. We show improvements of up to 0.4 BLEU.

The body of the paper is organized as follows: In Section 2, we briefly outline prior research. In Section 3, we describe baseline system configuration, investigated features and experimental setup. The results of the experiments are presented in Section 4. Section 5 concludes the paper and proposes some ideas for future work.

Prior Research

The impact of various features on SMT quality has been extensively studied in recent years. For instance, Och et al. (2004) have tried using various features for reranking the n-best lists of translations. However, it seems less promising than introducing new features at the time of decoding. A paper by Chiang et al. (2009) describes how combining multiple sparse features (i.e. rare features like specific lexical instantiations of a general feature) can improve syntax-based SMT. Hasler et al. (2012) explore the application of sparse lexicalized features for domain adaptation. Another common technique is adding a (usually) binary feature indicating whether a phrase originates from in-domain corpus or not (Dandapat et al. 2010; Pinnis, Skadiņš 2012).
The work by Cer et al. (2010) addresses the problem of lexical reordering and describes a feature which examines the reordering blocks of adjacent phrases. The same problem is investigated in (Collin 2013) using sparse features.

We focus on dense features for a standard phrase-based SMT system. We assume that we can extract the information directly from aligned phrase pairs and use it to make our system more robust and reduce the number of instances of incorrect word choice, punctuation mistakes and other problems SMT systems are prone to. We extracted shallow linguistic information from source and target phrases to improve overall translation quality of an SMT system. The features were precomputed and stored in the phrase table; they can thus be regarded as stateless dense features.

**System Description**

*Baseline System*

We conducted our experiments on the English-Russian language pair. We used the OPUS parallel corpora (Tiedemann 2012) to train the translation models and the 2014, 2015 news corpora from statmt.org to train the language model. We used the Moses open-source toolkit as the decoder. Moses is a state-of-the-art Statistical Machine translation system which uses bilingual parallel corpora to train a statistical translation model which is in turn used to produce translations from the source to the target language. We used MGIZA (Gao, Vogel 2008) to generate word alignments. We built the phrase tables and lexical reordering tables using the Moses toolkit. The IRSTLM toolkit (Federico et al. 2008) was used to build 3-gram language models, which were scored using KenLM (Heafield 2011) in the decoding process. We used ZMERT (Zaidan 2009) for weights optimization. The texts were tokenized and lowercased before training. The statistics on the training, development and test corpora are presented in Table 1. We used the whole GlobalVoices, News-Commentary11, Tatoeba and TED2013 corpora and randomly selected 1M parallel sentences for the MultiUN corpus. As for development and testing, we used the first 6006 lines of the WMT corpus for tuning and the rest 3000 lines as a test set.

| Corpus           | #token S     | #token T     |
|------------------|--------------|--------------|
| GlobalVoices     | 1819472      | 1602635      |
| MultiUN          | 27581770     | 25013175     |
| News-Commentary11| 4899256      | 4531701      |
| Tatoeba          | 787903       | 683608       |
| TED2013          | 2641553      | 2258270      |
| Tune             | 137501       | 122529       |
| Test             | 69301        | 62069        |
| Overall          | 37936756     | 34273987     |

**Feature Extraction**

We assume that we can focus on the similarity of the source and target phrases when designing our features. Thus, we explored features used in monolingual and parallel data filtering
techniques (see Khadivi, Ney 2005; Rarrick et al. 2011; Taghipour et al. 2011; Mahesh et al. 2011).

The set of selected features is as follows:

- **len_ratio** — ratio of source and target phrase lengths in tokens;
- **avg_tok_len_ratio** — average token length ratio;
- **punct_ratio** — punctuation symbol count ratio;
- **punct_identical_ratio** — identical punctuation symbol count ratio (i.e., ratio of identical punctuation symbol count to the length of the shorter phrase. The motivation for this was to investigate how indicating an unusually large number of punctuation symbols can affect translation quality);
- **alpha_ratio** — word (i.e., tokens containing only alphabetic characters, a hyphen and the apostrophe symbol for phrases in English) count ratio;
- **alpha_identical_ratio** — identical word count ratio;
- **no_alpha_ratio** — non-alphabetic token (i.e., tokens not containing alphabetic symbols) count ratio;
- **no_alpha_identical_ratio** — identical non-alphabetic token count ratio;
- **numbers_ratio** — number count ratio;
- **numbers_identical_ratio** — identical number count ratio;
- **mixed_ratio** — mixed token (i.e., tokens containing both alphabetic and non-alphabetic symbols) count ratio;
- **mixed_identical_ratio** — identical mixed token count ratio (i.e., the number of identical mixed tokens in both source and target phrases);
- **t_mean_frq** — mean frequency of target words (as described in the alpha_ratio feature) according to the target-domain frequency list. The frequency list was built from the 2014, 2015 Russian News corpora used for target language model training;
- **mean_frq_ratio** — source and target word mean frequency ratio (the source frequency list was built from the 2014, 2015 English News corpora from statmt.org);
- **alpha_len_ratio** — ratio of the source and target phrase lengths in words (the difference from the len_ratio feature is that only words are considered);
- **avg_alpha_tok_len_ratio** — average word length ratio;
- **mixed_s_ratio** — special symbol (neither alphabetic, nor numbers or punctuation) count ratio;
- **mixed_s_identical_ratio** — identical special symbol count ratio;
- **ppl_ratio** — source and target phrase perplexity ratio (we use the baseline News language model for the target and a language model built from the 2014, 2015 English News corpora for the source).

All the features were precomputed for all the phrase pairs; the scores were added directly to the phrase-table. Before computing the scores, we normalized the punctuation symbols to their standard ASCII versions.

**Experimental setup**

The baseline system only includes the five default features mentioned in the Introduction Section. First, we conducted a separate experiment for each additional feature. We used BLEU (Papineni et al. 2002) to automatically evaluate the results. We used the bootstrap resampling technique as described in (Koehn 2004) to see if the difference in BLEU scores for the baseline system and each of the improved systems is significant. We used the BLEU Kit to apply the bootstrap resampling method with the p-level of 0.05. Finally, we conducted an experiment combining all additional features, which showed positive results in terms of BLEU.
Extracting features from text to improve statistical machine translation

Results

The results of the experiments are presented in Table 2.

Table 2. Results of the experiments. Difference Significant stands for the bootstrap resampling results, Improvements stands for whether the feature improved the translation quality in terms of BLEU

| System          | Feature                | BLEU   | Difference significant | Improvement |
|-----------------|------------------------|--------|------------------------|-------------|
| Baseline        | —                      | 24.43  | —                      | —           |
| +feat1 len_ratio| 24.48                  | yes    | yes                    |
| +feat2 avg_tok_len_ratio | 24.52 | yes    | yes                    |
| +feat3 punct_ratio| 24.44                | yes    | yes                    |
| +feat4 punct_identical_ratio | 24.37 | no     | no                     |
| +feat5 alpha_ratio| 24.37                 | no     | no                     |
| +feat6 alpha_identical_ratio | 24.39 | no     | no                     |
| +feat7 no_alpha_ratio | 24.36             | yes    | no                     |
| +feat8 no_alpha_identical_ratio | 24.36 | yes    | no                     |
| +feat9 numbers_ratio | 24.43            | no     | no                     |
| +feat10 numbers_identical_ratio | 24.43   | no     | no                     |
| +feat11 mixed_ratio | 24.39        | yes    | no                     |
| +feat12 mixed_identical_ratio | 24.42        | yes    | no                     |
| +feat13 t_mean_frq | 24.65       | yes    | yes                    |
| +feat14 mean_frq_ratio | 24.8          | yes    | yes                    |
| +feat15 alpha_len_ratio | 24.41          | yes    | no                     |
| +feat16 avg_alpha_tok_len_ratio | 24.34        | yes    | no                     |
| +feat17 mixed_s_ratio | 24.37        | yes    | no                     |
| +feat18 mixed_s_identical_ratio | 24.37        | yes    | no                     |
| +feat19 ppl_ratio | 24.42       | yes    | no                     |

Five features showed BLEU improvements, but only two of them (t_mean_frq, mean_frq_ratio) can be considered successful. This can be explained by the fact that these features account for 1) the translation phrase reliability in general (t_mean_frq) and the reliability of the translation phrase according to the source phrase (mean_frq_ratio).

Conclusion and Future Work

We have presented a way to improve SMT translation quality by adding features using superficial linguistic information from source and target phrase pairs. We show moderate improvements in terms of BLEU for 5 out of 19 features: length ratio of the source and target phrases (in words), average token length ratio, punctuation count ratio, mean frequency of the target
words according to a general domain frequency list, and source and target word mean frequency ratio. We plan to further investigate the scoring and normalization techniques to improve the feature performance. We should also examine how the combination of successful features will affect translation quality.

References

Cer, D., Galley, M., Jurafsky, D., Manning, Ch. D. (2010) Phrasal: A Toolkit for statistical machine translation with facilities for extraction and incorporation of arbitrary model features. In: Human Language Technologies: the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Demonstration Session. Proceedings of the conference (NAACL-HLT 2010), pp. 9–12. (In English)

Collin, Ch. (2013) Improved reordering for phrase-based translation using sparse features. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013), pp. 22–31. (In English)

Chiang, D., Knight, K., Wang, W. (2009) 11,001 new features for statistical machine translation. In: Proceedings of the 2009 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2009), pp. 218–226. (In English)

Dandapat, S., Forcada, M. L., Groves, D. et al. (2010) OpenMaTrEx: A free/open-source marker-driven example-based machine translation system. In: Proceedings of the 7th International Conference on Natural Language Processing (IceTAL 2010), pp. 121–126. (In English)

Federico, M., Bertoldi, N., Cettolo, M. (2008) IRSTLM: An open source toolkit for handling large scale language models. In: Proceedings of the 9th Annual Conference of the International Speech Communication Association 2008 (INTERSPEECH 2008), pp. 1618–1621. (In English)

Gao, Q., Vogel, S. (2008) Parallel implementations of word alignment tool. In: Software Engineering, Testing, and Quality Assurance for Natural Language Processing, (SETQA-NLP 2008), pp. 49–57. (In English)

Hasler, E., Haddow, B., Koehn, Ph. (2012) Sparse lexicalised features and topic adaptation for SMT. In: Proceedings of the 9th International Workshop on Spoken Language Translation (IWSLT 2012), pp. 268–275. (In English)

Heafield, K. (2011). Kenlm: Faster and smaller language model queries. In: Proceedings of the Sixth Workshop on Statistical Machine Translation, pp. 187–197. (In English)

Khadivi, Sh., Ney, H. (2005). Automatic filtering of bilingual corpora for statistical machine translation. In: Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB 2005), pp. 263–274. (In English)

Koehn, Ph. (2004) Statistical significance tests for machine translation evaluation. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (A meeting of SIGDAT, a Special Interest Group of the ACL held in conjunction with ACL 2004), pp. 388–395. (In English)

Koehn, Ph., Hoang, H., Birch, A. et al. (2007). Moses: Open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions (ACL 2007), pp. 177–180. (In English)

Kavitha, K. M., Gomes, L., Lopes, G. P. (2011) Using SVMs for filtering translation tables. In: Proceedings of the 15th Portuguese Conference in Artificial Intelligence (EPIA 2011), pp. 690–702. (In English)

Och, F. J., Gildea, D., Khudanpur, S. et al. (2004) A smorgasbord of features for statistical machine translation. In: Proceedings of the Human Language Technologies Conference of the Association for Computational Linguistics: (HLT-NAACL 2004), pp. 161–168. (In English)

Papineni, K., Roukos, S., Ward, T. et al. (2002) BLEU: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002), pp. 311–318. (In English)

Pinnis, M., Skadiņš, R. (2012) MT adaptation for under-resourced domains — what works and what not. In: Proceedings of the Fifth International Conference “Human Language Technologies — The Baltic Perspective” (Baltic HLT 2012), pp. 176–184. (In English)

Rarrick, S., Quirk, Ch., Lewis, W. (2011) MT detection in Web-scraped parallel corpora. In: Proceedings of the 13th Machine Translation Summit (MT Summit XIII), pp. 422–430. (In English)

Taghipour, K., Khadivi, Sh., Xu, J. (2011) Parallel corpus refinement as an outlier detection algorithm. In: Proceedings of the 13th Machine Translation Summit (MT Summit XIII), pp. 414–421. (In English)

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012), pp. 2214–2218. (In English)
Zaidan, O. F. (2009). Z-MERT: A fully configurable open source tool for minimum error rate training of machine translation systems. *The Prague Bulletin of Mathematical Linguistics*, 91: 79–88. (In English)

**Author:**
Alexander P. Molchanov, e-mail: Alexander.Molchanov@promt.ru

**For citation:** Molchanov, A. P. (2019) Extracting features from text to improve statistical machine translation. *Journal of Applied Linguistics and Lexicography*, 1 (1), 12–17. DOI: 10.33910/2687-0215-2019-1-1-12-17

**Received** 30 May 2019; reviewed 3 July 2019; accepted 8 July 2019.

**Copyright:** © The Author (2019). Published by Herzen State Pedagogical University of Russia. Open access under CC BY-NC License 4.0.