Utilizing Source Context in Statistical Machine Translation

A. Tamchyna

Charles University, Faculty of Mathematics and Physics, Prague, Czech Republic.

Abstract. Current methods for statistical machine translation typically utilize only a limited context in the input sentence. Many language phenomena thus remain out of their reach, for example long-distance agreement in morphologically rich languages or lexical selection often require information from the whole source sentence. In this work, we present an overview of approaches for including wider context in SMT and describe our first experiments.

Introduction

Phrase-based statistical machine translation (PBSMT, Koehn *et al.* 2003) is arguably the most widespread approach to MT and represents the state-of-the-art for many language pairs [Bojar *et al.*, 2012; Hasler *et al.*, 2012; Huck *et al.*, 2012].

Phrase-based systems are trained from parallel sentence-aligned data. First, word alignment is estimated, i.e. for each word (token) in the parallel corpus, we learn which word(s) in the other language correspond to it. Unsupervised techniques are typically employed, which are based mostly on IBM models [Brown *et al.*, 1993]. Given the word alignment, pairs of phrases $(\boldsymbol{e}, \boldsymbol{f})$[1] are extracted from the training data using a simple heuristic. The "dictionary" of a phrase-based system is a phrase table: a list of phrase pairs and their probabilities $P(\boldsymbol{e}|\boldsymbol{f}), P(\boldsymbol{f}|\boldsymbol{e})$.[2] Complementary to the translation model, a language model is used for the probability of target-side word sequences, i.e. $P(\boldsymbol{e})$.

Finally, a log-linear combination of these individual features is used to score translations. The importance (weight) of each component is typically chosen to maximize the quality of translation of a development data set. The "translation quality" here corresponds to the score of an automatic metric, e.g. BLEU [Papineni *et al.*, 2002], which rewards the similarity of MT output to reference translations. Minimum error rate training (MERT, Och 2003) is the most common algorithm for parameter optimization.

During translation (decoding), input sentences are segmented into phrases, their possible translations are collected from the phrase table and a search is carried out for the most probable hypothesis according to the model.

PBSMT has many inherent limitations. In this work, we address one of them, namely the limited source-side context it considers. Because both the phrase table and the language model have a fixed maximum scope (maximum phrase length and n-gram order, respectively), many language phenomena simply cannot be captured by them. For English-Czech translation (and translation into morphologically rich languages in general), morphological coherence is a typical problem with which PBSMT struggles.

In the first example of Figure 1, the gender of the word "rády" is translated correctly by a top-performing MT system only when it is adjacent to the noun "děti". When these words are separated, the decoder (i.e. the MT system) selects the wrong surface form of the adverb, "rádi". This distance is naturally well within the maximum phrase scope, however the data provided insufficient statistics for this particular sequence of words (note that no abstraction

[1] The variables $\boldsymbol{e}, \boldsymbol{f}$ originally stood for English and French. Today, $\boldsymbol{e}$ is used for the target language and $\boldsymbol{f}$ stands for "foreign".

[2] Lexical weights $\text{lex}(\boldsymbol{e}|\boldsymbol{f}), \text{lex}(\boldsymbol{f}|\boldsymbol{e})$ are also part of the phrase table. They correspond to word-based translation probabilities and are used for smoothing.

Input	Google Translate	
Kids like to play football.	Děti **rády** hrají fotbal. *Kids$_{fem}$* ***gladly***$_{fem}$ *play football.*	✓
Kids mostly like to play football.	Děti většinou **rádi** hrají fotbal. *Kids$_{fem}$ mostly* ***gladly***$_{mas}$ *play football.*	✗
Shooting of the film.	**Natáčení** filmu. ***Shooting***$_{camera}$ *of_a_film.*	✓
Shooting of the expensive film.	**Natáčení** drahé filmu. ***Shooting***$_{camera}$ *of_an_expensive film.*	✓
Shooting of the least expensive film.	**Střelba** z nejlevnějších filmu. ***Shooting***$_{gun}$ *from the_cheapest film.*	✗

Figure 1. Examples of problems of PBMT: morphological coherence and lexical selection.

from word forms occurs in PBMT).

Another problem which cannot be handled without considering wider context is lexical selection. The second example in Figure 1 attempts to illustrate the problem on the ambiguous word "shoot". In Czech, the two meanings (shoot a film, shoot from a gun) have entirely different translations. Google Translate is capable of selecting the correct one only up to a certain distance from the cue word "film". Once the statistics become too sparse, the decoder switches to the other (presumably a priori more likely) translation.

Related Work

Word Sense Disambiguation

Word sense disambiguation (WSD) is an NLP task where the system selects from possible word meanings the one which applies in the current sentence. We mention this task here to highlight the connection between WSD and our work.

There is a large body of research on stand-alone WSD systems, but the main application of this task lies in using its output in some larger NLP system, such as a machine translation engine. In this case, it is useful to take possible translations of a word as the senses to be disambiguated [Vickrey *et al.*, 2005]. This step essentially turns WSD into a discriminative MT system which relies solely on source-side features for selecting word translations. Thus when we consider the goal of our work – including wider source-context information in a statistical (phrase-based) MT system – features and the overall know-how of WSD is a natural source of inspiration.

Using Source-Context in SMT

One of the first attempts at including wider context in phrase-based MT was Stroppa *et al.* [2007]. The authors used a *memory-based classifier* (an IGTree, Daelemans *et al.* [1997]) to score possible translations of phrases based on their source-side context. More specifically, the classifier features were:

- words on the left and right side of the phrase (a fixed-length window),
- part-of-speech (POS) tags of these context words.

The classifier produced a "similarity score" (i.e., how similar is the current context to the contexts where the given translation was observed). This score was normalized and added as a feature in the log-linear model. An additional feature was implemented which was 0 for all

translations except for the most probable one (according to the classifier), where it was set to 1. Weights of both features were tuned in the standard way (by MERT).

From a machine learning perspective, the setting was similar to 1-nearest-neighbor classification[3] with information gain as distance metric. No model of the data was therefore created. The authors reported significant improvement in BLEU score on Chinese→English and Italian→English language pairs.

Giménez and Màrquez [2007] built on the work of Vickrey *et al.* [2005], extending it in two crucial aspects: they moved from words to phrases (dubbing this task "discriminative phrase translation", DPT) and they incorporated their module in a full MT system, thus being able to evaluate its impact on translation quality. While no improvement of BLEU score was achieved, they carried out manual evaluation and confirmed that DPT helps MT quality. They used a rich feature set based on the WSD system of Yarowsky *et al.* [2001]. Local context was captured by a window of fixed length 5, from which words, POS tags and phrase-chunking labels were extracted. Global context was modeled as a bag of lemmas of the whole source sentence.

The learning setting was very different from Stroppa *et al.* [2007]: they used local linear support vector machines (SVM). For each source phrase S_i and for each of its possible translations T_j, they trained a binary one-against-all classifier $C_{i,j}$. Sentences where S_i was translated as T_j served as positive examples and all other occurrences of S_i as negative examples. Training such models is very taxing in terms of computing time and storage space; a problem common to nearly all the methods described in this paper.

Carpuat and Wu [2007] described an approach related to Giménez and Màrquez [2007], naming the task "phrase sense disambiguation" (PSD). The PSD module was based on their WSD system [Carpuat *et al.*, 2004] and utilized an ensemble of four machine-learning models: a naive Bayes model, a maximum-entropy model, a boosting model and a kernel-PCA model. The features were also based on their state-of-the-art WSD system. They carried out a large-scale evaluation using realistic data sizes across several language pairs. Their results showed consistent improvement of scores according to a wide range of MT metrics (including BLEU).

An interesting contrast to the previous approaches was developed by Gimpel and Smith [2008]. In the previous cases, context features were used as input for a classifier. The classifier produced a score C which was in turn used as a feature in the decoder, see Figure 2 (a). Gimpel and Smith [2008] (Figure 2 b) bypassed the classifier and defined rich context features directly in the decoder.

All of their features were probabilities, obtained from parallel data using maximum likelihood estimation (MLE):

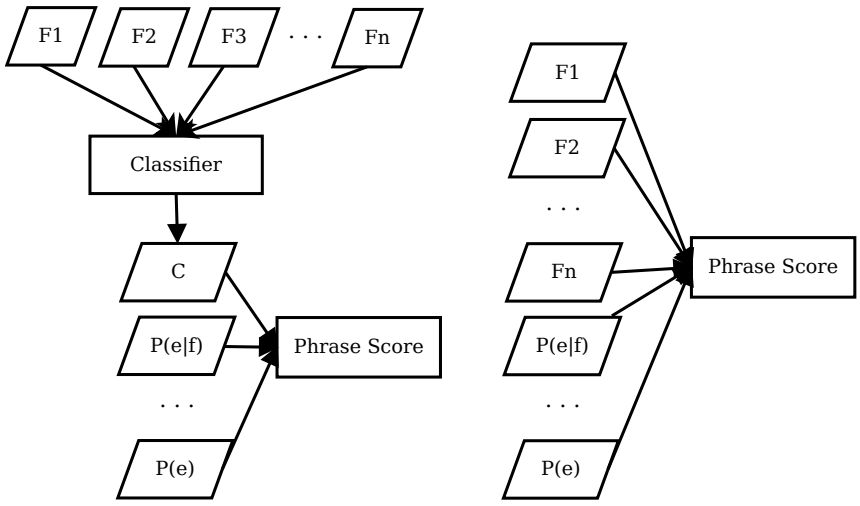
$$P(\boldsymbol{e}|\boldsymbol{f}, \boldsymbol{f}_{context}) = \frac{\text{count}(\boldsymbol{e}|\boldsymbol{f}, \boldsymbol{f}_{context})}{\sum_{e'} \text{count}(\boldsymbol{e'}|\boldsymbol{f}, \boldsymbol{f}_{context})}$$

Individual features then differed in the definition of $\boldsymbol{f}_{context}$. For lexical features, the context was a window of length 1 or 2 of words to the left and/or right of the source phrase. Shallow syntactic features were the same, except POS tags were used instead of word forms. The authors also described a number of syntactic features (such as "is the phrase strictly to the left of the root word") and positional features ("is the phrase at the beginning of the sentence").

While their ML estimates were not smoothed, the authors argued that some features served as back-off for others (e.g. shorter window or POS tags instead of words) and that tuning (MERT) in fact learned back-off weights. Reported results were mixed: while a significant improvement of BLEU was achieved for Chinese→English, the method did not help for German→English and English→German.

Chan *et al.* [2007] included source-context information in a hierarchical MT system. They used SVMs to predict the probability of translation given source context and included the scores

[3]The IGTree, while conceptually related to decision trees, was used here for compression and efficient lookup.



(a) Typical discriminative setting. (b) Gimpel and Smith [2008].

Figure 2. Integration of context features in the MT decoder.

as a feature in the decoder (limiting the predictions to short phrases). They also reported gains in BLEU score.

Mauser *et al.* [2009] introduced a technique which attempts to predict, for each word in the target vocabulary, whether or not it should appear in the translation of a given sentence. The authors trained a binary classifier for each target word, using as features only the bag of words (from the whole source sentence). Sentences where the target word occurred were used as positive examples, other sentences served as negative examples. During decoding, all classifiers were queried and translation hypotheses were rewarded based on the "scores" of words that they contained. Significant improvements of BLEU score were reported.

First Experiments

Context Similarity Feature

We developed an additional feature for the log-linear model in the decoder which scores phrases without the need for a classifier, relying on a simple measure of similarity. Specifically, given a phrase pair $(\boldsymbol{e}, \boldsymbol{f})$, our feature computes the cosine similarity between the current context and the contexts in training data where $\boldsymbol{f}$ was translated as $\boldsymbol{e}$. If we disregard word counts in the context vector, cosine similarity is reduced to:

$$\mathrm{sim}_1(A, B) = \frac{|A \cap B|}{\sqrt{|A|} \times \sqrt{|B|}}$$

We experimented with this simplified measure and with the formula which takes word frequencies into account:

$$\mathrm{sim}_2(A, B) = \frac{\sum_i A_i \cdot B_i}{\sqrt{\sum_i A_i^2} \times \sqrt{\sum_i B_i^2}}$$

Context vectors are sparse, so internally, we represent them as (C++) maps, allowing for efficient computation of the similarity score.

We evaluated a range of experimental settings, however the feature was never beneficial. Even parameter tuning assigned a very low weight to our feature (practically zero). While this result requires further investigation, we can attribute it (at least partially) to the following problems:

- The feature heavily depends on phrase segmentation.
- No abstraction from surface forms is done.
- Function words have the same weight as content words.

The feature is very unstable: if the decoder chooses a slightly different segmentation of input, observed contexts of the individual phrases change dramatically and the feature score is very different. We could mitigate this issue if we disregard function words and punctuation and perhaps also if we use e.g. lemmas instead of word forms, making our observations less sparse.

Beyond Phrase Sense Disambiguation

At the 2012 summer workshop at CLSP, Johns Hopkins University, a slightly modified approach to phrase sense disambiguation was developed and integrated in the Moses decoder [Carpuat *et al.*, 2012]. The learning setting is different: one global linear model is trained which predicts for each phrase pair the *loss* of selecting it in the current context. The model is trained with positive examples having zero loss and negative examples (all other possible translations of the given source phrase) having loss of one. During decoding, the classifier returns a loss for each possible translation of the current source phrase, the (inverse) losses are then normalized to form a probability distribution. The work is ongoing and no results of end-to-end MT evaluation are available yet.

The integration simplifies experimenting with PSD but also, more generally, with discriminative models in phrase-based MT (and also hierarchical PBMT). We are currently working on utilizing the developed framework to address the problems of morphological coherence and lexical selection in MT.

Specifically, we plan to use two separate classifiers. The first classifier will have features that reflect wider, sentence-level context (e.g. bag of lemmas) and will be used to predict which content words (or lemmas) should appear in the translation. The second classifier will be used to disambiguate morphological categories (i.e. Czech morphological tags). For this classifier, we will experiment with more local features, e.g. position-sensitive morphological and syntactic context in a small window around the source phrase.

Conclusion

We presented an overview of approaches for including wider source context in SMT. Most of the techniques use a discriminative learning setting and features inspired by word sense disambiguation to provide the MT decoder with context information. We described our first experiments in this area: the context similarity feature and a plan for adaptation of phrase-sense disambiguation for translating into morphologically rich languages.

Acknowledgement

This work was supported by the EU projects MosesCore (FP7-ICT-2011-7-288487) and Khresmoi (contract no. 257528) and by SVV project number 267 314.

References

Ondřej Bojar, Bushra Jawaid, and Amir Kamran. Probes in a taxonomy of factored phrase-based models. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 253–260, Montréal, Canada, June 2012. Association for Computational Linguistics.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, June 1993.

Marine Carpuat and Dekai Wu. Improving Statistical Machine Translation using Word Sense Disambiguation. Prague, Czech Republic, 2007.

Marine Carpuat, Weifeng Su, and Dekai Wu. Augmenting Ensemble Classification for Word Sense Disambiguation with a Kernel PCA Model. In *Proceedings of Senseval-3, Third International Workshop on Evaluating Word Sense Disambiguation Systems*, Barcelona, Spain, 2004. SIGLEX, Association for Computational Linguistics.

Marine Carpuat, Hal Daumé III, Alexander Fraser, Chris Quirk, Fabienne Braune, Ann Clifton, Ann Irvine, Jagadeesh Jagarlamudi, John Morgan, Majid Razmara, Aleš Tamchyna, Katharine Henry, and Rachel Rudinger. Domain adaptation in machine translation: Final report. In *2012 Johns Hopkins Summer Workshop Final Report*. 2012.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. Word Sense Disambiguation Improves Statistical Machine Translation. In *ACL*, 2007.

Walter Daelemans, Antal van den Bosch, and Ton Weijters. IGTree: Using Trees for Compression and Classification in Lazy Learning Algorithms, 1997.

Jesús Giménez and Lluís Màrquez. Context-aware Discriminative Phrase Selection for Statistical Machine Translation. pages 159–166, Prague, Czech Republic, 2007.

K. Gimpel and N. A. Smith. Rich Source-Side Context for Statistical Machine Translation. Columbus, Ohio, 2008.

Eva Hasler, Peter Bell, Arnab Ghoshal, Barry Haddow, Philipp Koehn, Fergus McInnes, Steve Renals, and Pawel Swietojanski. The UEDIN systems for the IWSLT 2012 evaluation. In *Proceedings of IWSLT*, 2012.

Matthias Huck, Stephan Peitz, Markus Freitag, Malte Nuhn, and Hermann Ney. The rwth aachen machine translation system for wmt 2012. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 304–311, Montréal, Canada, June 2012. Association for Computational Linguistics.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *HLT-NAACL*, 2003.

Arne Mauser, Sasa Hasan, and Hermann Ney. Extending Statistical Machine Translation with Discriminative and Trigger-Based Lexicon Models. pages 210–218, Suntec, Singapore, 2009.

Franz Josef Och. Minimum Error Rate Training in Statistical Machine Translation. In *Proc. of ACL*, pages 160–167, Sapporo, Japan, 2003. ACL.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. pages 311–318, 2002.

Nicolas Stroppa, Antal van den Bosch, and Andy Way. Exploiting source similarity for SMT using context-informed features. In *TMI 2007*, pages 231–240, Skövde, Sweden, 2007.

D. Vickrey, L. Biewald, M. Teyssier, and D. Koller. Word-Sense Disambiguation for Machine Translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Vancouver, Canada, October 2005.

David Yarowsky, Silviu Cucerzan, Radu Florian, Charles Schafer, and Richard Wicentowski. The Johns Hopkins SENSEVAL2 System Descriptions. In *In Proceedings of SENSEVAL2*, pages 163–166, 2001.