Sentence Similarity-Based Source Context Modelling in PBSMT

Rejwanul Haque, Sudip Kumar Naskar, Andy Way
CNGL, School of Computing
Dublin City University
Dublin 9, Ireland
{*rhaque, snaskar, away*}*@computing.dcu.ie*

Marta R. Costa-jussà
Barcelona Media Research Center
Speech and language
Av Diagonal 177, 08018 Barcelona
marta.ruiz@barcelonamedia.org

Rafael E. Banchs
Institute for Infocomm Research
Human Language Technology
A-STAR, Singapore 138632
rembanchs@i2r.a-star.edu.sg

***Abstract*—Target phrase selection, a crucial component of the state-of-the-art phrase-based statistical machine translation (PBSMT) model, plays a key role in generating accurate translation hypotheses. Inspired by context-rich word-sense disambiguation techniques, machine translation (MT) researchers have successfully integrated various types of source language context into the PBSMT model to improve target phrase selection. Among the various types of lexical and syntactic features, lexical syntactic descriptions in the form of supertags that preserve long-range word-to-word dependencies in a sentence have proven to be effective. These rich contextual features are able to disambiguate a source phrase, on the basis of the local syntactic behaviour of that phrase. In addition to local contextual information, global contextual information such as the grammatical structure of a sentence, sentence length and *n*-gram word sequences could provide additional important information to enhance this phrase-sense disambiguation. In this work, we explore various sentence similarity features by measuring similarity between a source sentence to be translated with the source-side of the bilingual training sentences and integrate them directly into the PBSMT model. We performed experiments on an English-to-Chinese translation task by applying sentence-similarity features both individually, and collaboratively with supertag-based features. We evaluate the performance of our approach and report a statistically significant relative improvement of 5.25% BLEU score when adding a sentence-similarity feature together with a supertag-based feature.**

***Keywords*-sentence similarity; source context information; statistical machine translation**

I. Introduction

The state-of-the-art PBSMT [1] model uses a translation model (TM) and a language model (LM) to translate a new source sentence. The TM usually comprises a phrase translation table (t-table) which is a collection of (source and target) phrase pairs extracted from a large bilingual training corpus. Translation of a source sentence begins by generating all possible source phrases and gathering all candidate phrases (target phrases) for each of the source phrases. In the translation process, thousands of translation hypotheses are statistically generated using the candidate phrases to form candidate translations. Therefore, the best candidate translation comprises the most appropriate candidate phrases selected from the pool of all candidate phrases. Interestingly, candidate phrase selection during translation is strongly driven by the LM which predicts how likely it is for a sequence of words to appear in the target language. In other words, it can be argued that phrase selection in PBSMT is modelled suboptimally, as it completely ignores the contexts in which a source phrase appears. Various types of lexical and syntactic features have been explored as effective source language contexts to improve target phrase selection in PBSMT, such as position-specific neighbouring words and POS tags [2], full sentential context [3], shallow and deep syntactic features [4], lexical syntactic descriptions in the form of supertags [5] and grammatical dependency relations [6]. We refer interested readers to [7] for more details of MT research on incorporating source contexts into SMT models. Among the source language contexts, supertags [8] which inherently capture long-range word-to-word dependencies in a sentence, have been demonstrated to perform better than words and POS tags in target phrase selection. Supertags include rich knowledge sources to disambiguate a source phrase, although they provide only local syntactic behaviour of a source phrase. However, sometimes, global contextual features such as the similarity measure between an input source sentence to be translated with the source-side of the bilingual training sentences could provide useful evidence to choose more appropriate candidate phrases. In addition to local contextual information, global contextual information could be an additional important source of information to enhance the sense disambiguation task. Costa-Jussà and Banchs [9] integrate source context information into the PBSMT model by incorporating a feature function estimated using a cosine distance similarity metric. Their feature function is computed for each of the phrase-pairs of the t-table by measuring cosine distance between the input sentence to be translated and the source sentences of the bilingual training corpus from which those phrase-pairs were extracted. A slight improvement was reported over the PBSMT system Moses[1] baseline on an English-to-Spanish experimental corpus. In this work, following [9] we explore various similarity features including cosine distance by measuring the similarity between a source sentence to be translated with the source-side of the parallel training

[1] http://www.statmt.org/moses/

sentences and integrate them directly into the state-of-the-art log-linear PBSMT [1] model. Furthermore, we perform experiments by combining sentence-similarity features (global contexts) with supertag-based features (local contexts) [5].

The remainder of the paper is organized as follows. In section 2 we describe the context-informed features contained in our baseline PBSMT model. Section 3 presents the results obtained, and offers a brief qualitative analysis. In Section 4 we formulate our conclusions, and offer some avenues for further work.

II. Context-Informed System

In PBSMT, selection of an appropriate candidate phrase for a source phrase depends only on the source phrase itself, regardless of its contexts, and the target language model which in turn represents target language context. In particular, disambiguation of a source phrase can be enhanced by taking into account the source language contextual information. In our work, a source sentence similarity score serves as a feature to represent global contextual information, and structural contextual information in the form of supertags serves as a feature to represent local contextual information.

A. Sentence-Similarity Features

A PBSMT t-table contains a list of source phrases and their corresponding translations (target phrases) with associated translation scores. Source and target phrases are extracted from a large bilingual training corpus to build the t-table. During training, additionally we keep track of the source sentences of the training corpus from which phrase pairs are extracted.

To translate a source test sentence, first we generate all possible source phrases and gather their corresponding candidate translations (target phrases). Then, we collect the source training sentences that are linked to the source phrase of each of the phrase-pairs. Thereafter, we measure the similarity between the source test sentence to be translated with the source training sentences. There could be two possible cases. Firstly, a phrase pair could be extracted from only one training sentence pair, in which case we calculate the similarity score between the source test sentence and only that particular training sentence. Secondly, phrase pairs could be extracted from many training sentence pairs, in which case we calculate the similarity score between the source test sentence and each of the training sentences separately, and then take the average of the scores. Finally, we normalize these similarity scores to convert them into probabilities ($\mathrm{P_{SIM}}$). Thus, we derive a log-linear feature $\hat{h}_{\mathrm{SIM}}$ to represent global context information as in (1):

$$\hat{h}_{\mathrm{SIM}} = \log \mathrm{P_{SIM}} \tag{1}$$

We have considered different similarity functions to measure the similarity between two sentences, including: (i) cosine distance as used in [9], (ii) dice coefficient (DC), and (iii) the METEOR automatic MT evaluation metric [10].

B. Supertag-Based Features

As in [5], we express a memory-based supertag feature ($\hat{h}_{\mathrm{MEM}}$) as the conditional probability of the target phrases ($\hat{e}_k$) given the source phrase ($\hat{f}_k$) and its *supertag information* (SI), as in (2):

$$\hat{h}_{\mathrm{MEM}} = \log \mathrm{P}(\hat{e}_k|\hat{f}_k, \mathrm{SI}(\hat{f}_k)) \tag{2}$$

As in [2], we used a memory-based classifier [11] to estimate $\mathrm{P}(\hat{e}_k|\hat{f}_k, \mathrm{SI}(\hat{f}_k))$ while avoiding sparseness problems. In our experiments two kinds of supertags are employed: those from lexicalized tree-adjoining grammar (LTAG) [8] and combinatory categorial grammar (CCG) [12]. As in [2], in addition to $\hat{h}_{\mathrm{MEM}}$, we derive a simple binary feature $\hat{h}_{\mathrm{best}}$ based on the $\mathrm{P}(\hat{e}_k|\hat{f}_k, \mathrm{SI}(\hat{f}_k))$ scores.

We performed experiments by integrating these log-linear features ($\hat{h}_{\mathrm{SIM}}$, $\hat{h}_{\mathrm{MEM}}$ and $\hat{h}_{\mathrm{best}}$) directly into the PBSMT model. Their weights are optimized using minimum error-rate training[2] on a held-out development set for each of the experiments. Like [5], we implemented a calling framework to take into account these features for the translation of the development and test set.

III. Results and Analysis

Experiments were carried out on an English-to-Chinese task. The training text contains 100,000 sentence pairs of newswire translation genres from LDC. We used the NIST MT05 1,082 test set sentences (devset) for tuning and the NIST MT08 1,357 'current' test set sentences (testset) for evaluation. Both the devset and testset have only one reference translation. Average sentence lengths in the English and Chinese training set are 45.35 and 36.79 words respectively; similar trends are observed for the testset and devset sentences. Our intention to combine sentence similarity-based features with supertag-based features forced us to choose English as the source language, given that supertaggers are readily available only for English.

A. Automatic Evaluation

Chinese translations generated by the systems are evaluated using BLEU[3] which has been applied at the word level. Additionally we performed statistical significance tests using bootstrap resampling [13]. The confidence level (%) of the improvements obtained by the context-informed systems with respect to the PBSMT baseline are reported in the result tables below.

The results obtained with the similarity features, compared to the Moses baseline, are shown in Table I. BLEU scores are very low due perhaps to the fact that we worked

[2]http://www.statmt.org/moses/?n=FactoredTraining.Tuning

[3]ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v13a.pl

on a small training set and we had only one reference translation for evaluation. We see that cosine distance

Experiments	BLEU
Baseline	*8.52*
Cosine	8.50
Monogram Overlap DC	8.69 (94%)
Bigram Overlap DC	**8.70** (88%)
Trigram Overlap DC	8.50
METEOR	8.52
Monogram + Bigram Overlap DC	8.49

Table I
EXPERIMENTAL RESULTS APPLYING SENTENCE-SIMILARITY FEATURES.

and METEOR similarity features are unable to show any improvement over the baseline. However, monogram and bigram overlap dice coefficient (DC) similarity features produce 0.17 BLEU points (1.79% relative) and 0.18 BLEU points (1.89% relative) improvements respectively over the baseline, and these improvements are very close to the significance level (95%); by contrast, the trigram overlap DC similarity feature is unable to show any improvement over the baseline. Furthermore, we applied two features (monogram and bigram overlap DC) together in the model as two different log-linear features to see whether further improvement could be achieved. However, using the combined features does not show any improvement.

The results obtained with the supertag-based features, compared to the Moses baseline, are shown in Table II. Experimental results in Table II clearly show that both CCG

Experiments	BLEU
Baseline	*8.52*
CCG	**8.74** (88%)
LTAG	8.71 (80%)

Table II
EXPERIMENTAL RESULTS APPLYING SUPERTAG-BASED FEATURES.

and LTAG supertag-based features improve over the baseline (0.22 BLEU points; 2.31% relative and 0.19 BLEU points; 2% relative improvements respectively). Improvements are quite close to the significance level (95%).

As was our main intention, we performed experiments by integrating sentence-similarity features and supertag-based features together into the PBSMT model as different log-linear features to see whether further improvements could be achieved. Hence, the best performing sentence-similarity set-ups (monogram and bigram overlap DC) are combined with CCG and LTAG supertag-based features. Experimental results obtained combining both types of features are shown in Table III. Improvement obtained by adding monogram overlap DC with LTAG is the highest (0.50 BLEU points improvement; 5.25% relative) over the baseline, which is statistically significant; by contrast, adding monogram overlap

Experiments	BLEU
Baseline	*8.52*
Monogram Overlap DC + CCG	8.48
Monogram Overlap DC + LTAG	**9.02** (99.9%)
Bigram Overlap DC + CCG	8.71 (88%)
Bigram Overlap DC + LTAG	8.73 (90%)

Table III
EXPERIMENTAL RESULTS APPLYING COMBINED FEATURES.

DC to CCG features does not produce any improvement over the baseline. Adding bigram overlap DC to CCG and LTAG features add a 0.19 BLEU point (2% relative increase) and a 0.21 BLEU point (2.2% relative increase) improvements over the baseline score.

B. Translation Analysis

We also performed a sentence-level automatic evaluation of the translations produced by our best-performing context-informed (CI) system (monogram overlap DC + LTAG) and those by the Moses baseline. Among the 1,357 test set sentences, the best system obtains a higher BLEU score than the baseline in 177 sentences, while the baseline obtains a higher BLEU score than the best system in 133 sentences. We performed a manual qualitative analysis of the translations of the two systems by randomly sampling a few (25 sentences) of those translations. Figure 1 shows two such translation examples which demonstrates how our context-informed system can improve over the baseline.

(1) input: he called for intensive talks in the coming weeks .
reference: 他 呼吁 在 未来 几 周 进行 密集 谈判 。
CI System: 他 呼吁 在 未来 数 周 密集 会谈 。
baseline: 他 呼吁 密集 会谈 在 未来 数 周 。

(2) input: i have too many things waiting for me that i cannot do now .
reference: 我 有 太 多 现在 无法 做 的 事情 等待 着 我 去 做 。
CI System: 我 有 太 多 的 事 等待 我 ， 我 现在 不 做 。
baseline: 我 有 太 多 的 事 等待 我 ， 我 不 能 在 。

Figure 1. Example Translations

We observe that our best-performing system produces more fluent translations (like example (1) and (2) in Figure 1) than the baseline Moses translations. The improvements of our system over the baseline system are two-fold: better word reordering (like example (1) in Figure 1) and better lexical selection (like example (2) in Figure 1).

We also looked at a few (25 sentences) of those translations (133 sentences) where the baseline system generates better translations than ours. We observe that both the baseline and our best-performing context-informed (CI) system show poor lexical choice and bad word order for most of those translations, due perhaps to the fact that the small training set does not contain the correct translation. Manual evaluation also reveals that our CI system tends to generate

additional words (i.e., preposition, verb) to form more fluent and grammatical translations than the baseline. However, sometimes these additional words seem to be unnecessary; an investigation on such anomalies is left for future work.

As an additional analysis, we found that the average number of candidate phrases that are used for a source phrase to translate the devset is 97.54 in the baseline PBSMT model. On the other hand, the average number of candidate phrases for a source phrase that are used to translate the same set are 57.01 and 53.74 in the CCG and LTAG supertag-based models respectively (combined models use the same set of target phrases as supertag-based models). Hence, in addition to the memory-based context-dependent translation features, supertag-based models use reduced but more fine-grained sets of candidate phrases in translation. Moreover, integration of sentence-similarity features into the PBSMT model helps the model to choose more appropriate candidate phrases for a source phrase during translation. Therefore, integration of global (sentence-similarity) and local (supertags) contextual features together into the PBSMT model force the model to weed out bad hypotheses during decoding, which improves translation quality.

IV. Conclusions and Future Work

In this work we explored various sentence-similarity features, and successfully integrated them into the state-of-the-art PBSMT model. The system produces 1.79% and 1.89% relative improvements on BLEU score while adding monogram and bigram DC sentence-similarity features respectively into the model. Following earlier work we compared the integration of sentence-similarity features to the integration of supertag features into the PBSMT system. The system also produces moderate gains for supertags (with 2% and 2.22% relative gains in BLEU for CCG and LTAG respectively). Furthermore, we performed experiments by integrating sentence-similarity features and supertag-based features collaboratively into the PBSMT model. We observed the highest improvement over the baseline when the monogram overlap DC sentence-similarity feature was combined with the LTAG supertag-based feature (5.25% relative gain on BLEU). Our experiments have focused on a standard but small dataset. Despite the challenges of scaling up to larger datasets for such types of features, we intend to further validate our conclusions on larger datasets.

Acknowledgments

We would like to thank our colleagues Yifan He and Jie Jiang for carrying out the manual evaluation. We are grateful to SFI (http://www.sfi.ie) and the Spanish Department of Education and Science for sponsoring this research under the grant 07/CE/I1142 and the *Juan de la Cierva* fellowship program, respectively.

References

[1] P. Koehn, F. J. Och, and D. Marcu, Statistical Phrase-Based Translation, In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP'03)*, Edmonton, AB, 2003, pp. 48–54.

[2] N. Stroppa, A. van den Bosch, and A. Way, Exploiting Source Similarity for SMT using Context-Informed Features, In *Proceedings of the 11th International Conference on Theoretical anc Methodological Issues in Machine Translation (TMI'07)*, Skövde, Sweden, 2007, pp. 231–240.

[3] M. Carpuat and D. Wu, Improving Statistical Machine Translation Using Word Sense Disambiguation, In *Proceedings of the EMNLP-CoNLL'07*, Prague, Czech Republic, 2007, pp. 61–72.

[4] K. Gimpel and N. A. Smith, Grammatical Dependencies for Contextual Phrase Translation Disambiguation, In *Proceedings of the Third Workshop on Statistical Machine Translation, ACL'08:HLT*, Columbus, OH, 2008, pp. 9–17.

[5] R. Haque, S. K. Naskar, Y. Ma, and A. Way, Using Supertags as Source Language Context in SMT, In *EAMT-2009: Proceedings of the 13th Annual Conference of the European Association for Machine Translation*, Barcelona, Spain, 2009, pp. 234–241.

[6] R. Haque, S. K. Naskar, A. van den Bosch, and A. Way, Dependency Relations as Source Context in Phrase-Based SMT, In *The 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC'09)*, Hong Kong, China, 2009, pp. 170–179.

[7] R. Haque, S. K. Naskar, A. van den Bosch, and A. Way, Supertags as Source Language Context in Hierarchical Phrase-Based SMT, In *AMTA 2010: 9th Conference of the Association for Machine Translation in the Americas*, Denver, CO, to appear.

[8] A. K. Joshi and Y. Schabes, Tree Adjoining Grammars and Lexicalized Grammars, In M. Nivat and A. Podelski (eds.) *Tree Automata and Languages*, Amsterdam, The Netherlands: North-Holland, 1992, pp. 409–431.

[9] M. R. Costa-Jussà and R. E. Banchs, A vector-space dynamic feature for phrase-based statistical machine translation, Journal of Intelligent Information Systems, 2010.

[10] A. Lavie and A. Agarwal, METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments, In *Proceedings of the Second Workshop on Statistical Machine Translation, ACL'07*, Prague, Czech Republic, 2007, pp. 228–231.

[11] W. Daelemans and A. van den Bosch, *Memory-Based Language Processing*, Cambridge University Press, Cambridge, UK, 2005.

[12] Mark Steedman, *The Syntactic Process*, MIT Press: Cambridge, MA, 2005.

[13] P. Koehn, Statistical significance tests for machine translation evaluation, In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP'04)*, Barcelona, Spain, 2004, pp. 388–395.