Improving Statistical Machine Translation with Selectional Preferences

Haiqing Tang, Deyi Xiong*, Min Zhang and Zhengxian Gong
Soochow University, Suzhou, China
hqtang@stu.suda.edu.cn
{dyxiong, minzhang, zhxgong}@suda.edu.cn

Abstract

Long-distance semantic dependencies are crucial for lexical choice in statistical machine translation. In this paper, we study semantic dependencies between verbs and their arguments by modeling selectional preferences in the context of machine translation. We incorporate preferences that verbs impose on subjects and objects into translation. In addition, bilingual selectional preferences between source-side verbs and target-side arguments are also investigated. Our experiments on Chinese-to-English translation tasks with large-scale training data demonstrate that statistical machine translation using verbal selectional preferences can achieve statistically significant improvements over a state-of-the-art baseline.

1 Introduction

Lexical translation error is one of the most urgent issues for statistical machine translation (SMT). Although phrase-based SMT can deal with local context dependencies well, it performs rather poorly with long-distance dependencies and therefore causes a lot of lexical translation errors. Verbs and their arguments form such long-distance dependencies and play important roles in translation as they build skeletons of sentences. However, many SMT systems are not sufficient to capture long-distance dependencies between arguments and their dominating verbs. Verbs and arguments are often either incorrectly translated or not translated at all according to the error study by Wu and Fung (2009a).

In order to address this issue, predicate-argument structures (PAS), which identify semantic frames within sentences by marking predicates, and labeling arguments with semantic roles, have been explored for SMT via various approaches in recent years. Wu and Fung (2009b) employ target-side PAS to pick out the most suitable translations among translation candidates after the decoding procedure is completed. Gildea (2010) integrates the PAS knowledge into decoding through projecting source-side PAS to the target-side via word alignments. In this paper, we are particularly interested in long-distance dependencies between verbs and their arguments in a predicate-argument structure. We propose to utilize selectional preferences (SPs) to handle these verb-argument dependencies for SMT.

Selectional preferences place semantic restrictions on words, with which words can co-occur in different syntactic patterns. To be more specific, the SPs of a verb can characterize the semantic restrictions that the verb imposes on its arguments. Violating these restrictions inevitably makes sentence senses odd or implausible. For example, in the sentence *"The ball drinks a potato."*, both subject and object preferences for the verb "drink" are violated. SPs have proven useful for numerous applications, e.g., semantic role labeling (Gildea and Jurafsky, 2002), pronoun resolution (Bergsma et al., 2008), textual inference (Pantel et al., 2007), word-sense disambiguation (Resnik, 1997) and many more. Therefore, we have sufficient theoretical foundation to believe that SPs between verbs and arguments can be used to alleviate translation errors that we pointed out above.

Our work consists of two parts: modeling SPs for verbs and incorporating SPs into an SMT system. In particular, we focus on the verb-object (v, obj) and verb-subject $(v, subj)$ selectional preference instances which can be extracted from our target-side corpus. SPs are computed in two ways: conditionally

*Corresponding author

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 2154–2163, Osaka, Japan, December 11-17 2016.

probabilistic SPs and topic-based SPs. The former calculates conditional probabilities between verbs and arguments as the strengths of SPs in a traditional way. The latter builds a class-based SP model using topics as semantic classes of arguments. All these calculated SPs are monolingual SPs. Since we model SPs for translation, we are also interested in cross-lingual SPs, i.e., selectional preferences of source-side verbs over corresponding target-side arguments. Taking (v, obj) semantic restriction as an example, we want to extract source-side verbs v_s and target-side objects obj_t from our word-aligned bilingual corpus. With these semantic instances, we define a bilingual SP model to calculate cross-lingual SP strength that a source-side v_s impose on its target-side obj_t. We integrate SPs into a state-of-the-art phrase-based SMT system. Experiments on large-scale translation display that SPs can achieve an improvement of up to 0.83 BLEU points over our baseline.

To the best of our knowledge, this is the first attempt to successfully incorporate selectional preferences into SMT. Our contributions are as follows.

- We propose various models to incorporate target-side monolingual selectional preferences into SMT.
- We also present a model for cross-lingual selectional preferences.
- In order to address the unknown word issue in SP modeling, we further introduce a word embedding based similarity model.
- Finally, we conduct experiments and in-depth analysis to demonstrate how these SP models work for SMT.

The remainder of this paper is organized as follows. Section 2 introduces related studies about SPs induction and application. Section 3 elaborates our methods to learn verbal SPs from a large-scale corpus and three SP models for SMT. Section 4 discusses how to deal with unknown words in SPs. Section 5 describes how we integrate the verbal SPs into SMT. Section 6 reports the experimental results. In the last section, we conclude with future directions.

2 Related Work

Recent two decades have witnessed increasing efforts on automatic acquisition of SPs for verbs as well as wide applications of SPs in NLP tasks. Resnik (1996) is a pioneer on the induction of SPs from corpus, proposing a class-based approach named selectional association that uses WordNet synsets to provide conceptual classes for nouns co-occurring with a specific predicate in a particular relation. Li and Abe (1998) also rely on WordNet and use the principle of Minimum Description Length to find a suitable generalization level of a noun. But entirely relying on WordNet to generalize nouns to semantic classes has a fatal disadvantage because WordNet is lack of coverage of proper nouns. Therefore, Rooth et al. (1999) propose a probabilistic latent variable model using Expectation-Maximization (EM) clustering algorithm to induce class-based SPs. Erk (2007) investigates a similarity-based model which takes advantage of a corpus-based distributional similarity metrics between arguments for SPs. More recently, a number of researchers come up with methods modeling SPs via unsupervised topic models where topics express a set of latent classes for preferences with different grammatical relations. Séaghdha (2010) describes a model using latent Dirichlet allocation (LDA) (Blei et al., 2003) to compute SPs composed of a predicate and a single argument. In contrast, Ritter et al. (2010) study acquiring selectional preferences of a predicate and multiple arguments with topic models.

SPs are useful for numerous NLP tasks. Resnik (1997) uses automatically acquired SPs for word sense disambiguation. Zapirain et al. (2009) employ SPs to process semantic role classification in a large dataset. Many researchers apply SPs to conduct pseudo-disambiguation tasks (Van de Cruys, 2014; Erk, 2007) in order to evaluate the performance of their methods of acquiring SPs. In contrast to plenty of applications of SPs in monolingual tasks, rather few efforts are devoted to incorporate SPs into SMT. To the best of our knowledge, we are the first to model SPs in the context of SMT.

From the perspective of verb and argument translation, the most related work to ours is Xiong et al. (2012). They propose two translation models to incorporate source-side PAS into SMT. One is the predicate translation model exploring both lexical and semantic contexts to predict target-side predicates. The other is the argument reordering model which estimates the direction of target-side arguments movement relative to their predicates. The significant difference is that they separately model the translation of verbs and arguments while we model them in a unified fashion via SPs.

3 Selectional Preference Model

Most approaches represent SPs for verbs as a function $\sigma : (v, r, c) \rightarrow s$ that maps each verb v and the semantic class c of its argument with respect to role r to a real-valued selectional preference strength s (Light and Greiff, 2002). The higher the value of s is, the more arguments semantically fit their dominating verb. In this paper, we are interested in the degree to which an object or subject semantically fits a given verb. Additionally, we are wondering which semantic relation is more helpful for a phrase-based machine translation. We propose two approaches: a conditional probability-based method and a topic-based method to model verbal SPs. Due to the space limit, we only describe how we compute the SP strength of (v, obj). The strength of $(v, subj)$ can be calculated in a similar way.

3.1 Conditional Probability-Based SPs

The conditional probability-based method is the most primitive corpus-based way to capture SPs that a verb imposes on its arguments. The conditional probability can be computed as follows.

$$P(n|v, r) = \frac{f(v, r, n)}{f(v, r)} \tag{1}$$

where $f(v, r, n)$ represents the number of times that a noun n co-occurs with a verb v in a grammatical relation r. Considering r as a relation of a direct object of v, n is correspondingly specified as the headword of r. Thus we simplify formula (1) to calculate the SP strength between a verb and its object as follows.

$$P_c(obj|v) = \frac{f(v, obj)}{f(v)} \tag{2}$$

where obj is the headword of the object of v.

3.2 Topic-Based SPs

Topic-based SP is a typical of class-based SP that models how well a particular class of words fits a verb. We use latent topics that are learned from a collection of documents as our semantic classes. We choose the most widely used LDA (Blei et al., 2003) topic model to infer topics for our arguments. Each word in our corpus is assigned a topic. Then we compute the SP for a verb and its object headword as follows.

$$\begin{aligned} P_t(obj|v) &= \sum_{tp \in T} P(tp|v) P(obj|v, tp) \\ &\approx \sum_{tp \in T} P(tp|v) P(obj|tp) \end{aligned} \tag{3}$$

where T denotes the collection of topics that the current obj belongs to and tp stands for a topic assignment of the object. The first part $P(tp|v)$ can be calculated with relative counts that a verb co-occurs with objects that are assigned a topic tp. The second part $P(obj|tp)$ can be directly retrieved from the per-topic word distribution of topic tp over words computed by the LDA topic model.

3.3 Bilingual SPs

The two models introduced above are used to calculate SPs for verbs only on the target side. We also want to model cross-lingual SPs that source-side verbs impose on their corresponding target-side arguments.

Selectional Pairs \ Translation Category	#0#	#1#	#2#	#3#
target-side (v, obj)	10.67%	27.46%	48.32%	13.54%
target-side $(v, subj)$	4.95%	34.45%	48.79%	11.80%

Table 1: Proportion of source-side verb-argument translation categories on the development set.

We therefore adapt the above two models to compute bilingual SPs. The conditionally probabilistic bilingual SP variant is calculated as follows.

$$P_{cbil}(obj_t|v_s) = \frac{f(v_s, obj_t)}{f(v_s)} \tag{4}$$

where obj_t is the target translation of obj_s. If obj is translated into a multi-word phrase, we use the first word of the phrase as obj_t.

For the bilingual topic-based SP model, we still use the LDA topic model to infer topics on the target language. We compute bilingual topic-based SPs via the following formula.

$$\begin{aligned} P_{tbil}(obj_t|v_s) &= \sum_{tp \in T} P(tp|v_s)P(obj_t|v_s, tp) \\ &\approx \sum_{tp \in T} P(tp|v_s)P(obj_t|tp) \end{aligned} \tag{5}$$

where T denotes the set of topics that the current obj_t belongs to and tp is the topic assigned to the object by LDA. $P(tp|v_s)$ is calculated with counts that the source-side verb v_s co-occurs with an object whose target-side counterpart is labeled with a topic tp. $P(obj_t|tp)$ is calculated by the LDA model.

4 SPs of Unseen Words

Conditional probability-based SPs cannot make any predictions for object headwords that have never occurred in our extracted selectional preference instances (v, obj). As our corpus cannot cover any phenomena in real world, a zero co-occurrence count between v and obj is not sufficient to show that the v has no selectional preference for that obj as its object. The method we employ to obtain source-side verb-argument pairs corresponding target-side verb-argument pairs during decoding has an obvious defect that word alignments directly affect the generation of translations. In this case, there may be many unseen word combinations. In order to investigate the proportion of unseen word combinations during decoding, we define four labels to classify these generated phrases. Label #0# represents phrases whose verb or object is translated into "#NULL#" due to incorrect word alignments. Phrases whose verbs or arguments are unseen in our trained SP models are labeled with #1#, #2# respectively. The remaining phrases appearing in our trained SP models are annotated with label #3#.

Table 1 shows the distribution of these phrases over the four categories on our development set (see details in Section 6.1). There are 53,268 (v, obj) pairs and 42,385 $(v, subj)$ pairs generated on the target side during decoding. Among these phrases, Only 13.54% (v, obj) pairs and 11.80% $(v, subj)$ pairs appear in our trained SP models. Most phrases, accounting for nearly 50%, are those whose argument headwords are unseen for our trained SP models. Hence, it is quite necessary to take some measure to model the SPs that verbs impose on their unseen argument headwords.

Instead of assigning a uniform value as the selectional strength for those unseen verb-argument combinations, we exploit a similarity-based model to compute SPs of a verb for an unseen argument headword during decoding, similar to the model by Erk (2007). Assuming (v, w_{un}) is a generated selectional preference instance according to word alignment information and w_{un} is an unseen object headword of v. The formulation to compute the selectional strength that v imposes on w_{un} is as follows.

$$P_c(w_{un}|v) = \sum_{w \in Seen(obj)} sim(w_{un}, w) \times wt_{obj}(w) \tag{6}$$

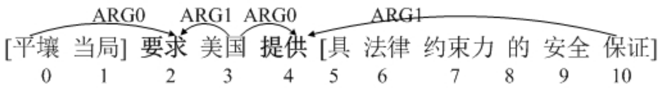

Figure 1: A source sentence with its predicate-argument structure. The verbs in the sentence are bold.

where $Seen(obj)$ is the set of seen headwords for an argument obj of a verb v, $sim(w_{un}, w)$ is the similarity between the seen and potential headword, and $wt_{obj}(w)$ is the weight of a seen headword w.

For the headword weight $wt_{obj}(w)$, we employ the selectional preference that the verb v imposes on the seen headword w to compute the value. $sim(w_{un}, w)$ is calculated with word2vec[1] and the similarity metric: Cosine. After each word on the target-side corpus is projected into a multidimensional vector space, $sim(w_{un}, w)$ is computed as follows.

$$Sim(\overrightarrow{w_{un}}, \overrightarrow{w}) = \frac{\overrightarrow{w_{un}} \bullet \overrightarrow{w}}{||\overrightarrow{sw_{un}}|| \times ||\overrightarrow{w}||} = \frac{\sum\limits_{i} (a_i \times b_i)}{\sqrt{\sum\limits_{i} {a_i}^2 \times \sum\limits_{i} {b_i}^2}} \tag{7}$$

where a_i and b_i are the value of ith dimension of their word embeddings.

5 Decoding

In this section, we mainly elaborate how to integrate the proposed SP models into a phrase-based SMT system built on bracketing transduction grammars (BTG) (Wu, 1997). Before we introduce the integration algorithm for SP models, we define two functions F and G on a source sentence and its predicate-argument structure following Xiong et al. (2012). We use the sentence in Figure 1 as an example to make the two functions easier to be understood.

- $F(i, j)$: The function is used to find positions of all verbs and their object headwords pairs from the predicate-argument structure. These pairs are completely located within the source span (i, j). For example, in Figure 1, $F(0, 4)$={(2,3)}, $F(0, 10)$={(2,3), (4,10)} while $F(0, 2)$={} because the object headword "美国" is located outside of the span $(0, 2)$ and $F(5, 10)$={} for the reason that the verb "提供" is located outside of the span $(5, 10)$.

- $G(i, k, j)$: The function finds positions of all verbs and their object headwords pairs that cross two neighboring spans (i, k) and $(k+1, j)$. It can also be formulated as $F(i, j) - (F(i, k) \cup F(k+1, j))$. In Figure 1, $G(0, 4, 10)$=$F(0, 10) - (F(0, 4) \cup F(5, 10))$={(4,10)}.

In order to calculate SP strengths of target-side verbs and arguments as well as bilingual verb-argument pairs, we store word alignment information for each phrase pair in the phrase table. Given a source sentence with its predicate-argument structure, if a BTG lexical rule is applied to translate a source phrase c spanning (i, j) to a target phrase e, we use $F(i, j)$ to detect all verb-object pairs and build a translation set $A(i, j) = \{(v_t, obj_t), (\cdots), \cdots\}$ to store corresponding verb-object translations on the target side through word alignments. Since our decoder is a log-linear model which is easy to incorporate new features, we define another function P_r to calculate the score of SPs as a new feature over span (i, j) as follows.

$$P_r(A(i, j)) = \prod_{(v_t, obj_t) \in A(i,j)} P_{.}(obj_t|v_t) \tag{8}$$

where $P_{.}(obj_t|v_t)$ can be the conditionally probabilistic model P_c or topic-based model P_t. For the bilingual SP models, we only need to change v_t to its source-side counterpart v_s.

If a BTG merging rule is applied to combine its two sub-spans (i, k), $(k + 1, j)$ in a straight $((i, k) + (k + 1, j) \rightarrow (i, j))$ or inverted order $((k + 1, j) + (i, k) \rightarrow (i, j))$, we directly use $P_r(A(i, k))$ and $P_r(A(k + 1, j))$ that have been already computed for the two sub-spans (i, k) and $(k + 1, j)$ in

[1]https://code.google.com/archive/p/word2vec/

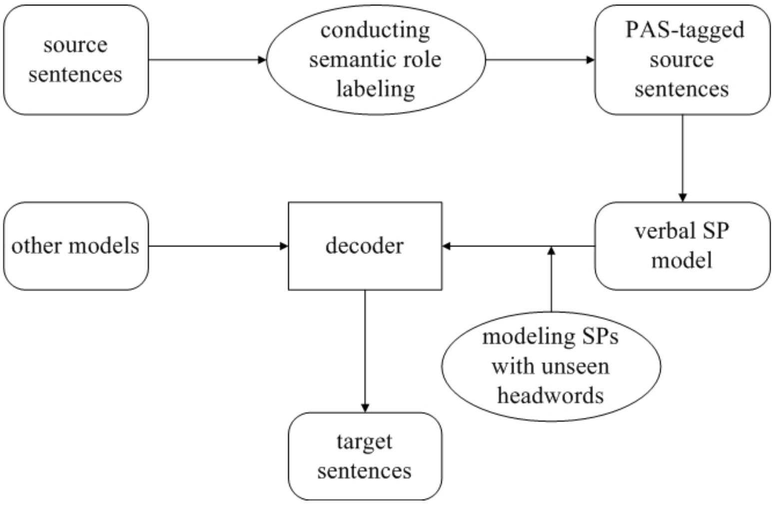

Figure 2: Architecture of SMT system equipped with verbal SPs.

the dynamic programming decoding algorithm. In this way, we only need to set another translation set $B(i,j) = \{(v_t, obj_t), (\cdots), \cdots\}$ to store the translations of source-side verb-object pairs found by $G(i,k,j)$ according to word alignments and calculate $P_r(B(i,j))$ for verb-object pairs that cross the two sub-spans.

In order to expedite the decoding process, we compute corresponding SPs for each (v, obj) semantic pairs extracted from the training corpus before decoding and load them when decoding instead of computing them on the fly. As for unseen object headword w_{un} of a verb, we use Eq. (6) to model SPs when integrating conditional probability-based SP model and Eq. (3) to model SPs when integrating topic-based SP model. We store the selectional strength of (v, w_{un}) for each unknown word so as to avoid repetitive computation. Figure 2 shows the architecture of the SMT system equipped with verbal SPs translation model. Since the system we used is based on a CKY-style decoder, the integration algorithm introduced here can be easily adapted to other CKY-based decoding systems such as the hierarchical phrasal system (Chiang, 2007).

6 Experiments

In order to validate the effectiveness of our SMT system enhanced with SPs, we perform a series of experiments on Chinese-to-English translation, which are trained with massive data. Specially, we aim at investigating:

- Whether integrating SPs into SMT can improve the system translation accuracy.
- Which can achieve better performance, conditionally probabilistic SP model or topic-based SP model?
- Whether semantic similarity-based approach is more reasonable than assigning a uniform value as the selectional strength that a verb imposes on its unseen argument headwords.
- Whether bilingual SPs are more effective than monolingual SPs for SMT.

6.1 Setup

The baseline is a state-of-the-art BTG-based phrasal system (Xiong et al., 2006). Our training data corpora[2] consist of 2.9M sentence pairs with 80.9M Chinese words and 86.4M English words. We ran GIZA++ on these corpora in both directions and then applied the "grow-diag-final" refinement rule to obtain final word alignments. Then we used all these word-aligned corpora to generate our phrase table.

[2]The corpora include LDC2003E14, LDC2004T07, LDC2005T06, LDC2005T10 and LDC2004T08 (Hong Kong Hansards/Laws/News).

Model	NIST04	NIST05
Base	36.40	33.69
Base+$P_c(obj_t\|v_t)$	36.93*	34.22**
Base+$P_c(obj_t\|v_t)$+$P_c(obj_{t_{un}}\|v_t)$	37.09*	34.43**
Base+$P_c(sub_t\|v_t)$	36.89	34.19*
Base+$P_c(sub_t\|v_t)$+$P_c(sub_{t_{un}}\|v_t)$	36.99*	34.37**
Base+$P_{cbil}(obj_t\|v_s)$	37.15**	34.21**

Table 2: Results of conditionally probabilistic SPs with two selectional relations: (v, obj) and (v, sub). **/*: significantly better than the baseline at $p < 0.01$ and $p < 0.05$ respectively.

Our 4-gram language model was trained on the Xinhua section of the English Gigaword corpus using the SRILM toolkit with modified Kneser-Ney smoothing.

In order to automatically learn SPs for verbs, we first parsed all source sentences using Stanford Parser and then ran the Chinese semantic role labeler (Li et al., 2010) on all source parse trees to annotate semantic roles for all verbs. At the same time, we ran SENNA on the target side to not only parse all target sentences but also conduct semantic role labeling for all verbs. It is easy to extract (v_t, obj_t) pairs or (v_t, sub_t) pairs after we obtained semantic roles on both sides. As for extracting (v_s, obj_t) selectional tuples, we first extracted (v_s, obj_s) pairs from source sentences with PAS and then used word alignments to get the target-side translation obj_t of obj_s. We used GibbsLDA++ to infer topics for our topic-based SP models. We set the number of topics from 50 to 350 with an incremental interval 50. We found the best number of topics according to results on our development set.

We trained word embeddings with word2vec using continuous bag-of-words model (Mikolov et al., 2013). The word vector dimensionality was set to 200 and we set the value of threshold for occurrence of words to 0.00001. Values of other parameters such as the training algorithm and the size of the window were all set by default.

We adopted the NIST MT03 evaluation test data as our development set, and the NIST MT04, MT05 as the test sets. We used the case-insensitive BLEU-4 (Papineni et al., 2002) to evaluate translation quality and run MERT (Och, 2003) three times. We finally recorded average BLEU scores over the three runs for all our experiments and used MultEval toolkit[3] to perform the significance test.

6.2 Results

Our first group of experiments is to investigate whether a simple conditional probability method for modeling SPs is able to improve translation accuracy in terms of BLEU. Moreover, we also would like to know whether the similarity-based SP model for unseen argument headwords will achieve further improvements. Experimental results are shown in Table 1. From the experiments which are conducted only using monolingual SPs, we can find that the verb-object SP model $P_c(obj_t|v_t)$ performs slightly better than $P_c(sub_t|v_t)$ on both test sets. Using semantic similarity metric rather than a uniform value to evaluate the selectional preference of a verb for its unseen argument can achieve better performance. It can also be observed that bilingual SPs marginally outperform than monolingual SPs on average. All SP models in this table are statistically better than the baseline on the test set MT05.

Our second group of experiments is to validate whether the topic-based SPs are more effective than conditionally probabilistic SPs in improving the accuracy of lexical choice. Table 2 shows our results. First, we have observations similar to what we have found in Table 1: verb-object SPs are better than verb-subject SPs while cross-lingual SPs better than monlingual SPs. Second, comparing Table 2 against Table 1, we find that topic-based SPs are better than conditional probabilistic SPs with a uniform value for unseen headwords, but similar to that with a similarity-based SP model for unseen headwords.

Analysis on translations reveals that our SP models are helpful for reducing verb-argument translation errors. Due to the space limit, we only show two translation examples. Figure 3 displays a translation example which shows that the system equipped with verbal SP model can solve the problem that the

[3]https://github.com/jhclark/multeval

Model	NIST04	NIST05
Base	36.40	33.69
Base+$P_t(obj_t\|v_t)$	37.11*	34.36**
Base+$P_t(sub_t\|v_t)$	37.07*	34.30**
Base+$P_{tbil}(obj_t\|v_s)$	37.23**	34.35**

Table 3: Results of topic-based SPs with two relations: (v, obj) and (v, sub). **/*: significantly better than the baseline at $p < 0.01$ and $p < 0.05$ respectively.

Figure 3: A translation example shows that verbal SPs can help SMT system alleviate the translation error that verb is not translated at all. The verbs in the sentence are bold.

baseline is unable to translate each verb in the source sentence to a target string. From the example, we can easily find that the baseline cannot correctly translate a $(verb, obj)$ selectional tuple like（参与，任务）where only obj “任务” is translated. Instead, in the system enhanced with verbal SPs, in addition to the object, $verb$ “参与” is also correctly translated into a target string. Figure 4 shows another example to demonstrate that verbal SPs are useful for selecting the proper translation for an object. The source word “发放” is not translated at all in the baseline while it is translated into “release” by our SP model.

7 Conclusion

We have presented three different models to compute SPs on verb-object and verb-subject pairs and successfully integrate them into a phrase-based SMT system. From a series of experiments on Chinese-to-English translation, we have found:

- Verbal SPs can significantly improve SMT in alleviating translation errors of verbs and their arguments.
- Verb-subject SPs perform similarly to verb-object SPs but slightly worse.

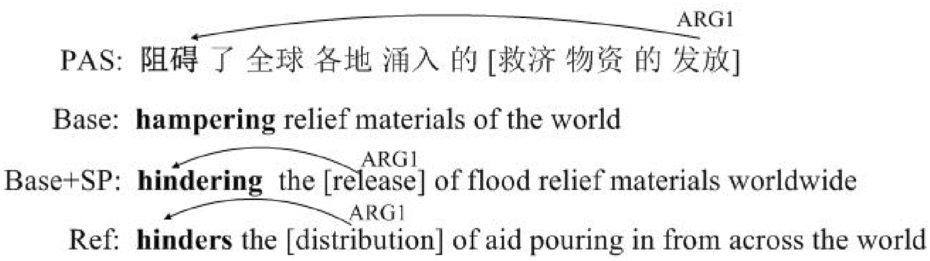

Figure 4: A translation example shows that verbal SPs can help SMT system alleviate the translation error that the argument of a verb is not translated at all. The verbs in the sentence are bold.

- Similarity-based SPs is helpful for conditional probability-based SP model to evaluate the SPs of a verb's unseen argument headword.

- Topic-based SPs are better than conditionally probabilistic SPs and bilingual SPs marginally better than monolingual SPs.

In the future, we would like to acquire bilingual SPs using a neural network approach (Van de Cruys, 2014). We also want to model SPs for verbs that are unseen in the training corpora and to explore a unified method to obtain SPs that a verb impose on its subject and object at the same time.

Acknowledgements

The authors were supported by National Natural Science Foundation of China (Grant Nos. 61403269, 61432013, 61525205, 61273319 and 61305088) and Natural Science Foundation of Jiangsu Province (Grant No. BK20140355). We also thank the anonymous reviewers for their insightful comments.

References

Shane Bergsma, Dekang Lin, and Randy Goebel. 2008. Discriminative learning of selectional preference from unlabeled text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 59–68. Association for Computational Linguistics.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of Machine Learning Research*, 3:993–1022.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

Katrin Erk. 2007. A simple, similarity-based model for selectional preferences. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, volume 45, page 216.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Daniel Gildea. 2010. Semantic role features for machine translation. In *International Conference on Computational Linguistics*, pages 716–724.

Hang Li and Naoki Abe. 1998. Generalizing case frames using a thesaurus and the mdl principle. *Computational Linguistics*, 24(2):217–244.

Junhui Li, Guodong Zhou, and Hwee Tou Ng. 2010. Joint syntactic and semantic parsing of chinese. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1108–1117. Association for Computational Linguistics.

Marc Light and Warren Greiff. 2002. Statistical models for the induction and use of selectional preferences. *Cognitive Science A Multidisciplinary Journal*, 26(3):269–281.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Computer Science*.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 160–167. Association for Computational Linguistics.

Patrick Pantel, Rahul Bhagat, Bonaventura Coppola, Timothy Chklovski, and Eduard H Hovy. 2007. Isp: Learning inferential selectional preferences. In *HLT-NAACL*, pages 564–571.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Philip Resnik. 1996. Selectional constraints: An information-theoretic model and its computational realization. *Cognition*, 61(1):127–159.

Philip Resnik. 1997. Selectional preference and sense disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How*, pages 52–57. Washington, DC.

Alan Ritter, Oren Etzioni, et al. 2010. A latent dirichlet allocation method for selectional preferences. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 424–434. Association for Computational Linguistics.

Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. 1999. Inducing a semantically annotated lexicon via em-based clustering. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 104–111. Association for Computational Linguistics.

Diarmuid O Séaghdha. 2010. Latent variable models of selectional preference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 435–444. Association for Computational Linguistics.

Tim Van de Cruys. 2014. A neural network approach to selectional preference acquisition. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 26–35.

Dekai Wu and Pascale Fung. 2009a. Can semantic role labeling improve smt? pages 218–225.

Dekai Wu and Pascale Fung. 2009b. Semantic roles for smt: a hybrid two-pass model. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics*, pages 13–16.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403.

Deyi Xiong, Qun Liu, and Shouxun Lin. 2006. Maximum entropy based phrase reordering model for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 521–528. Association for Computational Linguistics.

Deyi Xiong, Min Zhang, and Haizhou Li. 2012. Modeling the translation of predicate-argument structure for smt. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 902–911. Association for Computational Linguistics.

Beñat Zapirain, Eneko Agirre, and Lluís Màrquez. 2009. Generalizing over lexical features: Selectional preferences for semantic role classification. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 73–76. Association for Computational Linguistics.