Improving Translation Selection with Supersenses

Haiqing Tang[1], **Deyi Xiong**[1]*, **Oier Lopez de Lacalle**[2] **and Eneko Agirre**[2]
Soochow University, Suzhou, China[1]
University of the Basque Country, Donostia, Spain[2]
hqtang@stu.suda.edu.cn, dyxiong@suda.edu.cn
{oier.lopezdelacalle, e.agirre}@ehu.eus

Abstract

Selecting appropriate translations for source words with multiple meanings still remains a challenge for statistical machine translation (SMT). One reason for this is that most SMT systems are not good at detecting the proper sense for a polysemic word when it appears in different contexts. In this paper, we adopt a supersense tagging method to annotate source words with coarse-grained ontological concepts. In order to enable the system to choose an appropriate translation for a word or phrase according to the annotated supersense of the word or phrase, we propose two translation models with supersense knowledge: a maximum entropy based model and a supersense embedding model. The effectiveness of our proposed models is validated on a large-scale English-to-Spanish translation task. Results indicate that our method can significantly improve translation quality via correctly conveying the meaning of the source language to the target language.

1 Introduction

Phrase-based SMT has achieved better performance than word-based SMT. One of the reasons is that continuous phrases, rather than single words, are used as translation units so that useful context information can be captured for selecting appropriate translations. Even so, when translating sentences containing ambiguous words, which have multiple meanings, the state-of-the-art phrase-based SMT is still suffering from inaccurate lexical choice which makes translations unable to correctly convey the meaning of source sentences. Recent studies show that in order to improve translation quality, one must correctly identify the most likely senses of source-side ambiguous words when selecting target translation (Gao et al., 2013; Zou et al., 2013; Zhang et al., 2014).

One common approach to deal with ambiguity is to incorporate a word sense disambiguation (WSD) system into SMT system. At first, Carpuat and Wu (2005) attempt to use the senses of source ambiguous words predicted by a standard formulation of WSD directly in a word-based SMT but results are disappointing. They are skeptical of the assumption that WSD systems are useful for SMT. Instead of the standard WSD task, Vickrey et al. (2005) propose a novel formulation of WSD for SMT: directly predicting possible target translation candidates as senses for ambiguous source words. This reformulated WSD has been shown to help SMT by several subsequent studies, including later work by Carpuat and Wu (2007). Following this WSD reformulation for SMT, they integrate the WSD training, where sense definitions are drawn automatically from all phrasal translation candidates rather than from a predefined sense inventory into a phrase-based SMT.

In addition to WSD, topic model (Blei et al., 2003) is yet another technique used to detect most likely senses of source words. Various topic-specific lexicon translation models are proposed to improve translation quality. These models can be classified into two categories: word-level translation models (Zhao and Xing, 2006; Tam et al., 2007) and phrase-level models (Xiao et al., 2012). Especially, Xiong and Zhang (2014) propose a sense-based translation model that integrates hidden word senses into machine

*Corresponding author

translation to investigate whether hidden senses are useful for SMT. They resort to word sense induction (WSI) and build a broad-coverage sense tagger that relies on the nonparametric Bayesian model to obtain hidden senses for each source word in large-scale corpora. Different from what the previous reformulated WSD does, they first predict word senses which are automatically learned from data for ambiguous words and then make use of predicted word senses along with other context features to predict possible target translations for these words. They conclude that word senses automatically induced by WSI are very useful for SMT in dealing with inaccurate lexical choice.

However, all the models mentioned above are focusing on investigating how to exploit fine-grained word senses (e.g., predefined senses, target translation candidates, hidden senses) in SMT to improve translation quality. Then how about coarse-grained word senses such as supersenses that are WordNet (Fellbaum, 1998) semantic labels grouping semantically close synsets into a coarse-grained ontology? Are they useful for SMT? Since supersense tagging is the task of assigning high-level ontological classes to open-class words such as nouns, verbs, it is thus a coarse-grained word sense disambiguation task. To the best of our knowledge, we are the first to be dedicated to systematically investigating whether supersenses can be used in SMT to alleviate source word sense translation errors. Specifically, we try to model supersenses for SMT in the following two ways:

- A maximum entropy (MaxEnt) based model: Building multiple MaxEnt classifiers with one classifier per source word type, which incorporate supersenses as features.

- And a supersense embedding model: Projecting word supersenses to a multidimensional vector space with word2vec[1], and inducing source phrase supersense embeddings for phrasal translation.

These two supersenses-based translation models are integrated into a state-of-the-art SMT system and a series of experiments are conducted on English-to-Spanish translation based on large-scale training data. Results show that supersenses, high-level ontological concepts, are capable of improving translation quality and the supersense embedding model outperforms the MaxEnt classifiers-based model.

Our work is different from previous standard WSD, reformulated WSD and WSI with hidden senses for SMT. The first uses fine-grained senses predefined by WordNet. The second explores surrounding words to disambiguate word senses. And the third integrates topics inferred from pseudo documents as hidden senses. We employ coarse-grained predefined categories from WordNet to disambiguate ambiguous words for SMT.

The remainder of this paper is organized as follows. Section 2 introduces related studies exploiting word sense knowledge to improve lexical selection in SMT and various applications of supersenses and word embeddings. Section 3 elaborates how we obtain supersense tags for words in large-scale data. Section 4 describes our supersense-based translation models. Section 5 presents the way that we integrate supersense-based translation models into SMT. Section 6 discusses our experiments and results. In section 7 we summarize our findings and directions for future work.

2 Related Work

The problem of accurate lexical choice is an unsolved challenge for phrase-based SMT. Much work has been done to identify proper senses of source ambiguous words to aid system in choosing appropriate translations. Integrating WSD into an SMT system is typical of this work as described in Section 1 (Carpuat and Wu, 2005; Vickrey et al., 2005; Carpuat and Wu, 2007; Chan et al., 2007). Exploring topic model for SMT is another attempt. Gong et al. (2010) introduce document-level topics to help SMT generate target translations. They use a monolingual LDA model to assign a specific topic to the document to be translated. Similarly, each phrase pair is also assigned with one specific topic. A phrase pair will be filtered from phrase table if its topic mismatches the document topic. Xiao et al. (2012) propose a topic similarity model which incorporates the rule-topic distributions on both the source and target side into a hierarchical phrase-based system for rule selection.

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

noun	Tops	act	animal	artifact	attribute
	body	cognition	communication	event	feeling
	food	group	location	motive	object
	person	phenomenon	plant	possession	process
	quantity	relation	shape	state	substance
	time				
verb	body	change	cognition	communication	competition
	consumption	contact	creation	emotion	motion
	perception	possession	social	stative	weather

Table 1: Supersense labels for nouns and verbs in WordNet.

Supersenses are useful and have been used as high-level features in various tasks. Ciaramita and Altun (2006) define a tagset based on WordNet supersenses to perform broad-coverage word sense disambiguation and information extraction which they approach as a unified tagging problem. They achieve considerable improvements over the first sense baseline. Koo and Collins (2005) utilize supersense reranking that provides a partial disambiguation step in syntactic parse to build useful latent semantic features. Other tasks like preposition sense disambiguation (Ye and Baldwin, 2007), noun compound interpretation (Tratz and Hovy, 2010) can also employ supersenses to improve performance.

Word embeddings, also called distributed word representations, are used in many natural language processing areas such as information retrieval (Manning et al., 2008), search query expansions (Jones et al., 2006), or representing semantics of words (Reisinger and Mooney, 2010). As to SMT, Zou et al. (2013) propose a method to learn bilingual word embeddings for recognizing and quantifying semantic similarities across languages.

Our work is to integrate supersenses into a phrase-based SMT. We adopt two ways to train our supersense-based models: one is a MaxEnt classifier-based model which is closely related to Xiong and Zhang's (2014) work, the other is a model built on supersense embeddings that are different from word embeddings in that supersense embeddings can provide more high-level semantic information than word embeddings.

3 Supersense Tagging

Supersenses, first defined by Ciaramita and Johnson (2003), are coarse-grained semantic labels used by lexicographers to facilitate the development of WordNet. There are 45 supersense labels, 26 for nouns, 15 for verbs, 3 for adjectives and 1 for adverbs, used in WordNet to classify synsets into several domains based on syntactic category and semantic coherence. Normally, an ambiguous word belongs to several synsets. Since supersense labels are assigned to synsets, word sense ambiguity can be preserved to a certain degree at this level. In this paper, we focus on noun and verb supersenses. Table 1 shows the corresponding supersense labels in WordNet.

We use supersenses as our semantic classes to tag our training data for obtaining contextual information for the following advantages. First, this set of semantic labels is fairly general and therefore small. The reasonable size of the label set makes it possible to have only one model. In contrast, we have one model per word for fine-grained WSD. Second, the sensible semantic categories are easily recognizable and not too abstract. Since similar words tent to be merged together, these semantic categories seem promising to be used in MT. Third, while individual glossary can not embody too much about the narrow concept it is attached to, at the supersense level this information accumulates. In order to be more intuitive, we take the verb ***"help"*** as an example to show how fine-grained senses are grouped in supersenses, which is presented in Table 2.

In this paper, supersense tagging is carried out with a model based on Ciaramita and Altun's (2006) work. We deploy the implementation provided by Michael Heilman[2]. The model takes a sequence labeling approach to learn a model for supersense tagging. Specifically, we employ a sequential labeller

[2]http://www.ark.cs.cmu.edu/mheilman/questions/SupersenseTagger-10-01-12.tar.gz

Supersense	WN Senses	Gloss
social	1	give help or assistance; be of service
social	6	contribute to the furtherance of
body	2	improve the condition of
stative	3	be of use
stative	4	abstain from doing; always used with a negative
change	8	improve; change for the better
consumption	5	help to some food; help with food or drink
consumption	7	take or use

Table 2: An example of grouping different fine-grained senses of a word into supersenses. WN senses: senses from WordNet.

that is based on a Hidden Markov Model (HMM) trained in a discriminative way. That is, the model can be seen as a perceptron-trained HMM (Collins, 2002) that jointly models observation/label sequences. The model is trained on Semcor Corpus (Miller et al., 1993) following the experimental setting described in Ciaramita and Altun (2006). WordNet fine-grained senses are mapped to their corresponding supersense. In our case, only nouns and verbs are mapped, labeling as "NULL" the rest of the tokens (including adjectives and adverbs). In some cases "noun.Tops" refers to more specific supersenses, such as "food", "person", or "animal". In those cases we substitute the "noun.Tops" with more specific label (e.g "animal" as "noun.animal"). Although the tagger learns 41 semantic categories, we included (B) beginning and (I) continuation as supersense prefixes to learn more categories. Thus, actual label space to be learned increases to 83 (including "NULL").

Semcor is divided in three parts: "brown1" and "brown2", in which nouns, verbs, adjectives and adverbs are annotated. But the section "brownv", contains annotations only for verbs. To avoid many nouns being labeled with "NULL" we take the same procedure as Ciaramita and Altun (2006) to extract the text segment including a verb but not a noun.

Regarding features, our implementation replicates the features used in the original work, which include words and part-of-speech tags occurring in a context-window, word-shapes of the surrounding words, the first-sense of the word and the surrounding words, and the previous label. Please refer to the original work for a more detailed description of the model.

4 Supersense-based Translation Model

In this section, we present two methods to build the proposed supersense-based translation models.

4.1 A MaxEnt Classifier-based Model

Given a source word c with its contextual information including supersenses, we resort to a MaxEnt classifier to estimate the probability $p(e|C(c))$ of a target phrase e. The MaxEnt classifier is formulated as follows.

$$P(e|C(c)) = \frac{exp(\sum_i \theta_i h_i(e, C(c)))}{\sum_{e'} exp(\sum_i \theta_i h_i(e', C(c)))} \quad (1)$$

where h_i are binary features, θ_i are weights of these features.

We use two groups of features: lexicon features and supersense features, which are used to define $C(c)$ as follows:

$$C(c) = \{c_{-k}, s_{c_{-k}}, ..., c_{-1}, s_{c_{-1}}, c, s_c, c_1, s_{c_1}, ..., c_k, s_{c_k}\} \quad (2)$$

where c represents words and s for supersenses. In this way, not only centered word c and its supersenses s_c are included, but the preceding and succeeding k words with their corresponding supersenses are also involved. Particularly c_{-k} is the kth preceding word of c and $s_{c_{-k}}$ is the supersense of word c_{-k}. We extract all training events defined by $C(c)$ for each source word c and train multiple MaxEnt classifiers with one classifier per source word.

4.2 A Supersense Embedding Model

A traditional way to generate a phrase table is to use the training component of Moses[3] that allows you to automatically train translation models for any language pairs. Normally, each entry in the phrase table contains a source phrase, a target phrase, their word alignments, and five types of translation scores. From the training corpora where the source side is annotated with supersenses, we want to learn the distribution of the supersenses tagged for a source-side phrase. In order to achieve this goal, we have made some changes on the phrase extraction and scoring module of Moses training system to make them output source word supersenses. The number of extracted phrase pairs are expanded greatly since the same source phrase may correspond to several sequences of tagged supersenses.

With word2vec, we can train word sense embeddings on the supersense-tagged corpus. Assuming that a source phrase src containing n words $(w_1, w_2, ..., w_n)$ has k supersense sequences: $ps_1, ps_2, ..., ps_k$. Since each word supersense in the phrase constitutes the phrase supersense sequence, we can use Eq. (3) to represent ps_i, and Eq. (4) to represent the supersense sequence embedding.

$$ps_i = (w_1|ws_1\ w_2|ws_2\ ...\ w_n|ws_n) \tag{3}$$

$$\overrightarrow{ps_i} = \overrightarrow{w_1|ws_1} + \overrightarrow{w_2|ws_2} + ... + \overrightarrow{w_n|ws_n} \tag{4}$$

where ws represents supersense labels and | is a separator between the word and its supersense label.

Each supersense sequence has a unique sense embedding according to the calculating method above. Then how can we represent the supersense embeding of a source phrase in the phrase table? We adopt a dividing and merging method.

The dividing method A translation rule in the original phrase table is divided into several rules for the reason that the source phrase of the rule has several supersense sequences. Direct and indirect phrase translation probability calculated in Eq. (5) and (6) are reformulated and recalculated according to Eq. (7) and (8) respectively.

$$P(e|f) = \frac{Count(e, f)}{Count(f)} \tag{5}$$

$$P(f|e) = \frac{Count(f, e)}{Count(e)} \tag{6}$$

$$P(e|f, ps) = \frac{Count(e, f, ps)}{Count(f, ps)} \tag{7}$$

$$P(f, ps|e) = \frac{Count(f, ps, e)}{Count(e)} \tag{8}$$

where f, e, ps stands for source phrase, target phrase, source phrase supersense sequence respectively. In this way, a source phrase may have multiple supersense embeddings.

The merging method Supposing a source phrase src has k supersense sequences: $ps_1, ps_2, ..., ps_k$, we study the probability distribution of each supersense sequence according to the following formula.

$$P(ps_i|src) = \frac{Count(ps_i, src)}{Count(src)} \tag{9}$$

We assign each source phrase a unique sense embedding $\overrightarrow{s_{src}}$ using the following formula (10).

$$\overrightarrow{s_{src}} = \lambda_1\overrightarrow{ps_1} + \lambda_2\overrightarrow{ps_2} + ... + \lambda_k\overrightarrow{ps_k} \tag{10}$$

where λ_i represents the probability of the ith supersense sequence calculated according to Eq. (9).

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

5 Decoding

We integrate the proposed supersense-based translation models described above into a log-linear translation framework of SMT as a new knowledge source to disambiguate source words. Both supersense-based translation models require that each sentence should be sense-tagged to annotate each word with supersenses before being translated. We adopt an integration strategy similar to that introduced by Xiong and Zhang (2014) to incorporate the MaxEnt classifier-based model into our SMT system. During decoding, once a new source word c is translated, we find its target phrase e according to word alignments that are kept in the phrase table. Then we compute the translation probability $p(e|C(c))$ via the equation (1) using the corresponding classifier. As to integrating the supersense embedding model, we load the pre-trained word sense embeddings to calculate the source phrase sense embeddings according to Eq. (4) when translating a sentence. We compute the similarity between a source phrase in a source sentence and the corresponding matched phrase from the phrase table using the following formula in order to select appropriate translation rules.

$$Sim(\overrightarrow{s_{ssrc}}, \overrightarrow{s_{tsrc}}) = \frac{\overrightarrow{s_{ssrc}} \bullet \overrightarrow{s_{tsrc}}}{||\overrightarrow{s_{ssrc}}|| \times ||\overrightarrow{s_{tsrc}}||} = \frac{\sum_i (a_i \times b_i)}{\sqrt{\sum_i {a_i}^2 \times \sum_i {b_i}^2}} \tag{11}$$

where s_{ssrc} is a phrase in a source sentence, s_{tsrc} is the same phrase in the phrase table, a_i and b_i are the value of ith dimension of their sense embeddings.

Given a source sentence $\{c_i\}_1^N$, We define the score of supersense embedding model as follows.

$$Score_{M_s} = \sum_{ssrc_i \in \Gamma} Sim(\overrightarrow{s_{ssrc_i}}, \overrightarrow{s_{tsrc_i}}) \tag{12}$$

where Γ is a set of source phrases which have translation rules in the phrase table.

6 Experiments

In this section, we conducted a series of experiments on English-to-Spanish translation using massive training data. With the trained supersense-based translation model, we would like to investigate the following two questions:

- Whether coarse-grained supersenses can improve translation quality.
- Whether supersense embeddings can play a role in lexical selection.

6.1 Setup

Our baseline is a state-of-the-art SMT system which adapts Bracketing Transduction Grammars (Wu, 1997) to phrasal translation and augment itself with a maximum entropy based reordering model (Xiong et al., 2006). Our training corpora are English-Spanish sentences from the Europarl parallel corpus (Koehn, 2005) consisting of 1.9M sentence pairs with 51M English words and 54M Spanish words. We ran GIZA++ on the training data in both directions and then applied the "grow-diag-final" refinement rule (Koehn et al., 2003) to obtain final word alignments. Our phrase table was generated according to the word-aligned data. As to the language model, we trained a separate 5-gram LM using the SRILM toolkit (Stolcke, 2002) with modified Kneser-Ney smoothing (Chen and Goodman, 1996) on each subcorpus[4] and then interpolated them according to the corpus used for tuning.

We trained our MaxEnt classifiers with the off-the-shelf MaxEnt tool.[5] We performed 100 iterations of the L-BFGS algorithm implemented in the training toolkit on the collected training events from the sense-annotated data. We set the Gaussian prior to 1 to avoid overfitting.

[4]There are 12 subcorpora: commoncrawl, europarl, kde4, news2007, news2008, news2009, news2010, news2011, news2012, newscommentary, openoffice, un

[5]http://homepages.inf.ed.ac.uk/lzhang10/maxenttoolkit.html

System	batch2	batch2a	batch2q
Base	37.86	36.06	42.59
Max_ss	38.25**	36.33	43.10*
Max_hs	38.28**	36.47**	42.94

Table 3: Results of MaxEnt-based sense models with supersenses (Max_ss) vs. hidden senses (Max_hs) against the baseline. **/*: significantly better than the baseline at $p < 0.01$ and $p < 0.05$ respectively.

	System	batch2	batch2a	batch2q
	Base	37.86	36.06	42.59
Dividing	SSTM(100)	38.51**	36.55*	43.22**
	SSTM(200)	38.42**	36.32	43.23**
Merging	SSTM(100)	38.05	36.17	42.93
	SSTM(200)	38.64**	36.68**	43.17*

Table 4: Results of using the dividing and merging method to train supersense embedding-based translation model (SSTM) with vector dimensionality varying from 100 to 200. **/*: significantly better than the baseline at $p < 0.01$ and $p < 0.05$ respectively.

The method used to learn supersense embeddings, word2vec, in this paper was implemented based on continuous bag-of-words model (Mikolov et al., 2013). We only varied vector dimensionality from 100 to 200 and set the value of threshold for occurrence of words to 0.00001. Default values of other parameters such as the training algorithm and the size of the window were all taken.

The QTLeap corpus[6] was divided into two parts equally to be used as our development set *batch1* and test set *batch2*. The corpus was composed by 4000 question and answer pairs in the domain of computer and IT troubleshooting for both hardware and software. This material was collected using a support service via chat, this implies that the corpus is composed by naturally occurring utterances produced by users while interacting with a service. Only interactions composed by one question and the respective answer were included in the corpus. We also divided our test set *batch2* into two parts equally ***batch2a*** and ***batch2q*** respectively. In other words, we used three test sets to verify the effectiveness of our proposed models. We adopted the case-insensitive BLEU-4 (Papineni et al., 2002) as evaluation metric and ran MERT (Och, 2003) three times to alleviate the instability. We reported average BLEU scores over the three runs as final results.

6.2 Results

Our first group of experiments are designed to investigate whether supersenses can be modeled like hidden senses using a MaxEnt classifier. We use the same experiment settings as Xiong and Zhang (2014) did. Especially, we also find that 10-word window is the most suitable window for extracting semantic information according to experiments. Table 3 shows the experimental results for the two SMT systems equipped with multiple MaxEnt classifiers trained on supersenses and hidden senses respectively.

We can easily find that resorting to a MaxEnt classifier, supersenses can also be integrated into the SMT system and achieve an improvement over the baseline, which is comparable to that obtained by hidden senses.

Inspired by the idea that distribution word representations can convey contextual information, we conduct our second group of experiments to investigate whether distributed supersense representations can be used to improve SMT. We are also concerned about the potential impact of the sense embedding dimensionality on the performance of the supersense embedding model. Hence, in addition to the different methods to represent source phrase supersense embeddings, we consider the dimension as 100 and 200 when training word supersense embeddings. Experimental results are listed in Table 4. From the table, we can observe that

[6]http://metashare.metanet4u.eu/go2/qtleapcorpus

	System	batch2	batch2a	batch2q
	Base	37.86	36.06	42.59
Dividing	SSTM	38.51**	36.55*	43.22**
	HSTM	38.20*	36.25	42.85
Merging	SSTM	38.64**	36.68**	43.17*
	HSTM	38.15*	36.29	42.95

Table 5: Results of using the dividing and merging method to obtain supersense embedding-based translation model (SSTM) vs. hidden sense embedding-based translation model (HSTM). **/*: significantly better than the baseline at $p < 0.01$ and $p < 0.05$ respectively.

- No matter which method we use to calculate source phrase supersense embeddings, the supersense embedding-based translation model is able to achieve an average of 0.7 BLEU points over the baseline on three test sets.
- When using the dividing method to calculate source phrase sense embeddings, the sense embedding dimension has a slight impact on the translation quality in terms of BLEU.
- On the contrary, when using the merging method to obtain source phrase supersense embeddings, training supersense embeddings with 100-dimension performs worse than 200-dimension on the test set. The BLEU score drops by 0.6 points on average. This may be because the merging method uses multiple supersense sequences to compute the final supersense embeddings (see Eq. (10)). The fluctuations caused by the dimensionality of embeddings may be amplified by the summation in Eq. (10).

Our final group of experiments is to study 1) whether hidden senses can be integrated into the SMT system in the way similar to supersense embeddings and 2) which kind of word senses can perform better. When using the dividing method to obtain source phrase hidden sense embeddings, we set the dimension value to 100 in which case supersense embeddings perform a little better than 200-dimension according to Table 4. As for the merging method, we set the dimension value to 200 to make an equitable comparison with supersenses. Table 5 shows the results. We find that supersense embedding-based model performs better than hidden sense embedding-based model in all cases. The reason for this may be that hidden senses have already encoded distributional information in themselves.

7 Conclusion

We have exploited coarse-grained supersenses which are semantic labels defined by WordNet to conduct high-level word sense disambiguation for SMT. After each source word in the training data is tagged with a supersense, we take two strategies to train our supersense-based translation models. One is utilizing a maximum entropy classifier to predict the target translation for a source word given its surrounding words and their corresponding supersenses. The other is taking advantage of a word2vec tool to learn supersense embeddings on corpus annotated with supersenses and then calculating the semantic similarity between phrases in source sentence and matched phrases from phrase table. The supersense-based translation model is integrated into a phrase-based SMT system.

We have conducted a series of experiments to validate the effectiveness of the proposed supersense-based translation models. Final experimental results show us that

- The supersense-based translation model is capable of improving translation quality significantly in terms of BLEU.
- When using a MaxEnt classifier to predict target translation for a source word given its surrounding semantic information, both supersenses and hidden senses perform well.
- When using distributed sense representations to build sense-based translation models, supersense embeddings perform better than hidden sense embeddings.

In the future, we would like to investigate new neural models to learn embeddings for a single word supersense as well as a supersense sequence. We are also interested in incorporating the relations of supersenses (i.e., high-level ontological concepts) for machine translation.

Acknowledgements

The authors were supported by National Natural Science Foundation of China (Grant Nos. 61403269, 61432013, and 61525205) and Natural Science Foundation of Jiangsu Province (Grant No. BK20140355). In addition, this work was partially funded by MINECO (TUNER, TIN2015-65308-C5-1-R) and the European Commission (QTLeap, FP7-ICT-2013.4.1-610516). We also thank the anonymous reviewers for their insightful comments.

References

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

Marine Carpuat and Dekai Wu. 2005. Word sense disambiguation vs. statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 387–394. Association for Computational Linguistics.

Marine Carpuat and Dekai Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 61–72.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 33–40.

Stanley F Chen and Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 310–318. Association for Computational Linguistics.

Massimiliano Ciaramita and Yasemin Altun. 2006. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 594–602. Association for Computational Linguistics.

Massimiliano Ciaramita and Mark Johnson. 2003. Supersense tagging of unknown nouns in wordnet. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 168–175. Association for Computational Linguistics.

Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 1–8. Association for Computational Linguistics.

Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

Jianfeng Gao, Xiaodong He, Wen-tau Yih, and Li Deng. 2013. Learning semantic representations for the phrase translation model. *Computer Science*.

Zhengxian Gong, Yu Zhang, and Guodong Zhou. 2010. Statistical machine translation based on lda. In *Universal Communication Symposium (IUCS), 2010 4th International*, pages 286–290. IEEE.

Rosie Jones, Benjamin Rey, Omid Madani, and Wiley Greiner. 2006. Generating query substitutions. In *Proceedings of the 15th International Conference on World Wide Web*, pages 387–396. ACM.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*, volume 5, pages 79–86.

Terry Koo and Michael Collins. 2005. Hidden-variable models for discriminative reranking. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 507–514. Association for Computational Linguistics.

Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze, et al. 2008. *Introduction to information retrieval*, volume 43. Cambridge university press Cambridge.

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

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *The Workshop on Human Language Technology*, pages 303–308.

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

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.

Joseph Reisinger and Raymond J Mooney. 2010. Multi-prototype vector-space models of word meaning. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 109–117. Association for Computational Linguistics.

Andreas Stolcke. 2002. Srilm-an extensible language modeling toolkit. In *InterSpeech*, volume 2002, page 2002.

Yik Cheung Tam, Ian Lane, and Tanja Schultz. 2007. Bilingual lsa-based adaptation for statistical machine translation. *Machine Translation*, 21(4):187–207.

Stephen Tratz and Eduard Hovy. 2010. A taxonomy, dataset, and classifier for automatic noun compound interpretation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 678–687. Association for Computational Linguistics.

David Vickrey, Luke Biewald, Marc Teyssier, and Daphne Koller. 2005. Word-sense disambiguation for machine translation. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 771–778. Association for Computational Linguistics.

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

Xinyan Xiao, Deyi Xiong, Min Zhang, Qun Liu, and Shouxun Lin. 2012. A topic similarity model for hierarchical phrase-based translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 750–758. Association for Computational Linguistics.

Deyi Xiong and Min Zhang. 2014. A sense-based translation model for statistical machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*.

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.

Patrick Ye and Timothy Baldwin. 2007. Melb-yb: Preposition sense disambiguation using rich semantic features. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 241–244. Association for Computational Linguistics.

Min Zhang, Xinyan Xiao, Deyi Xiong, and Qun Liu. 2014. Topic-based dissimilarity and sensitivity models for translation rule selection. *Journal of Artificial Intelligence Research*, 50(1):1–30.

Bing Zhao and Eric P Xing. 2006. Bitam: Bilingual topic admixture models for word alignment. In *Proceedings of the COLING/ACL on Main Conference Poster Sessions*, pages 969–976. Association for Computational Linguistics.

Will Y Zou, Richard Socher, Daniel M Cer, and Christopher D Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398.