A Sense-Based Translation Model for Statistical Machine Translation

Deyi Xiong and **Min Zhang***
Provincial Key Laboratory for Computer Information Processing Technology
Soochow University, Suzhou, China 215006
{dyxiong, minzhang}@suda.edu.cn

Abstract

The sense in which a word is used determines the translation of the word. In this paper, we propose a sense-based translation model to integrate word senses into statistical machine translation. We build a broad-coverage sense tagger based on a nonparametric Bayesian topic model that automatically learns sense clusters for words in the source language. The proposed sense-based translation model enables the decoder to select appropriate translations for source words according to the inferred senses for these words using maximum entropy classifiers. Our method is significantly different from previous word sense disambiguation reformulated for machine translation in that the latter neglects word senses in nature. We test the effectiveness of the proposed sense-based translation model on a large-scale Chinese-to-English translation task. Results show that the proposed model substantially outperforms not only the baseline but also the previous reformulated word sense disambiguation.

1 Introduction

One of very common phenomena in language is that a plenty of words have multiple meanings. In the context of machine translation, such different meanings normally produce different target translations. Therefore a natural assumption is that word sense disambiguation (WSD) may contribute to statistical machine translation (SMT) by providing appropriate word senses for target translation selection with context features (Carpuat and Wu, 2005).

This assumption, however, has not been empirically verified in the early days. Carpuat and Wu (2005) adopt a standard formulation of WSD: predicting word senses that are defined on an ontology for ambiguous words. As they apply WSD to Chinese-to-English translation, they predict word senses from a Chinese ontology HowNet and project the predicted senses to English glosses provided by HowNet. These glosses, used as the sense predictions of their WSD system, are integrated into a word-based SMT system either to substitute for translation candidates of their translation model or to postedit the output of their SMT system. They report that WSD degenerates the translation quality of SMT.

In contrast to the standard WSD formulation, Vickrey et al. (2005) reformulate the task of WSD for SMT as predicting possible target translations rather than senses for ambiguous source words. They show that such a reformulated WSD can improve the accuracy of a simplified word translation task.

Following this WSD reformulation for SMT, Chan et al. (2007) integrate a state-of-the-art WSD system into a hierarchical phrase-based system (Chiang, 2005). Carpuat and Wu (2007) also use this reformulated WSD and further adapt it to multi-word phrasal disambiguation. They both report that the redefined WSD can significantly improve SMT.

Although this reformulated WSD has proved helpful for SMT, one question is not answered yet: are pure word senses useful for SMT? The early WSD for SMT (Carpuat and Wu, 2005) uses projected word senses while the reformulated WSD sidesteps word senses. In this paper we would like to re-investigate this question by resorting to word sense induction (WSI) that is related to but different from WSD.[1] We use

*Corresponding author

[1] We will discuss the relation and difference between WSI and WSD in Section 2.

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 1459–1469,
Baltimore, Maryland, USA, June 23-25 2014.

WSI to obtain word senses for large-scale data. With these word senses, we study in particular: 1) whether word senses can be directly integrated to SMT to improve translation quality and 2) whether WSI-based model can outperform the reformulated WSD in the context of SMT.

In order to incorporate word senses into SMT, we propose a sense-based translation model that is built on maximum entropy classifiers. We use a nonparametric Bayesian topic model based WSI to infer word senses for source words in our training, development and test set. We collect training instances from the sense-tagged training data to train the proposed sense-based translation model. Specially,

- Instead of predicting target translations for ambiguous source words as the previous reformulated WSD does, we first predict word senses for ambiguous source words. The predicted word senses together with other context features are then used to predict possible target translations for these words.

- Instead of using word senses defined by a prespecified sense inventory as the standard WSD does, we incorporate word senses that are automatically learned from data into our sense-based translation model.

We integrate the proposed sense-based translation model into a state-of-the-art SMT system and conduct experiments on Chines-to-English translation using large-scale training data. Results show that automatically learned word senses are able to improve translation quality and the sense-based translation model is better than the previous reformulated WSD.

The remainder of this paper proceeds as follows. Section 2 introduces how we obtain word senses for our large-scale training data via a WSI-based broad-coverage sense tagger. Section 3 presents our sense-based translation model. Section 4 describes how we integrate the sense-based translation model into SMT. Section 5 elaborates our experiments on the large-scale Chinese-to-English translation task. Section 6 introduces related studies and highlights significant differences from them. Finally, we conclude in Section 7 with future directions.

2 WSI-Based Broad-Coverage Sense Tagger

In order to obtain word senses for any source words, we build a broad-coverage sense tagger that relies on the nonparametric Bayesian model based word sense induction. We first describe WSI, especially WSI based on the Hierarchical Dirichlet Process (HDP) (Teh et al., 2004), a nonparametric version of Latent Dirichlet Allocation (LDA) (Blei et al., 2003). We then elaborate how we use the HDP-based WSI to predict sense clusters and to annotate source words in our training/development/test sets with these sense clusters.

2.1 Word Sense Induction

Before we introduce WSI, we differentiate **word type** from **word token**. A word type refers to a unique word as a vocabulary entry while a word token is an occurrence of a word type. Take the first sentence of this paragraph as an example, it has 11 word tokens but 9 word types as there are two word tokens of the word type "we" and two tokens of the word type "word".

Word sense induction is a task of automatically inducing the underlying senses of word tokens given the surrounding contexts where the word tokens occur. The biggest difference from word sense disambiguation lies in that WSI does not rely on a predefined sense inventory. Such a prespecified list of senses is normally assumed by WSD which predicts senses of word tokens using this given inventory. From this perspective, WSI can be treated as a clustering problem while WSD a classification one.

Various clustering algorithms, such as k-means, have been previously used for WSI. Recently, we have also witnessed that WSI is cast as a topic modeling problem where the sense clusters of a word type are considered as underlying topics (Brody and Lapata, 2009; Yao and Durme, 2011; Lau et al., 2012). We follow this line to tailor a topic modeling framework to induce word senses for our large-scale training data.

In the topic-based WSI, surrounding context of a word token is considered as a **pseudo document** of the corresponding word type. A pseudo document is composed of either a bag of neighboring words of a word token, or the Part-to-Speech tags of neighboring words, or other contextual information elements. In this paper, we define a pseudo

document as $\pm N$ neighboring words centered on a given word token. Table 1 shows examples of pseudo documents for a Chinese word "wǎngluò" (network). These two pseudo documents are extracted from a sentence listed in the first row of Table 1. Here we set $N = 5$. We can extract as many pseudo documents as the number of word tokens of a given word type that occur in training data. The collection of all these extracted pseudo documents of the given word type forms a corpus. We can induce topics on this corpus for each pseudo document via topic modeling approaches.

Figure 1(a) shows the LDA-based WSI for a given word type W. The outer plate represents replicates of pseudo documents which consist of N neighboring words centered on the tokens of the given word type W. $w_{j,i}$ is the i-th word of the j-th pseudo document of the given word type W. $s_{j,i}$ is the sense assigned to the word $w_{j,i}$. The conventional topic distribution θ_j for the j-th pseudo document is taken as the the distribution over senses for the given word type W. The LDA generative process for sense induction is as follows: 1) for each pseudo document D_j, draw a per-document sense distribution θ_j from a Dirichlet distribution Dir(α); 2) for each item $w_{j,i}$ in the pseudo document D_j, 2.1) draw a sense cluster $s_{j,i} \sim$ Multinomial(θ_j); and 2.2) draw a word $w_{j,i} \sim \varphi_{s_{j,i}}$ where $\varphi_{s_{j,i}}$ is the distribution of sense $s_{j,i}$ over words drawn from a Dirichlet distribution Dir(β).

As LDA needs to manually specify the number of senses (topics), a better idea is to let the training data automatically determine the number of senses for each word type. Therefore we resort to the HDP, a natural nonparametric generalization of LDA, for the inference of both sense clusters and the number of sense clusters following Lau et al. (2012) and Yao and Durme (2011). The HDP for WSI is shown in Figure 1(b). The HDP generative process for word sense induction is as follows: 1) sample a base distribution G_0 from a Dirichlet process DP(γ, H) with a concentration parameter γ and a base distribution H; 2) for each pseudo document D_j, sample a distribution $G_j \sim$ DP(α_0, G_0); 3) for each item $w_{j,i}$ in the pseudo document D_j, 3.1) sample a sense cluster $s_{j,i} \sim G_j$; and 3.2) sample a word $w_{j,i} \sim \varphi_{s_{j,i}}$. Here G_0 is a global distribution over sense clusters that are shared by all G_j. G_j is a per-document sense distribution over these sense clusters, which has its own document-specific proportions of these sense clusters. The hyperparameter γ, α_0 in the HDP are both concentration parameters which control the variability of senses in the global distribution G_0 and document-specific distribution G_j.

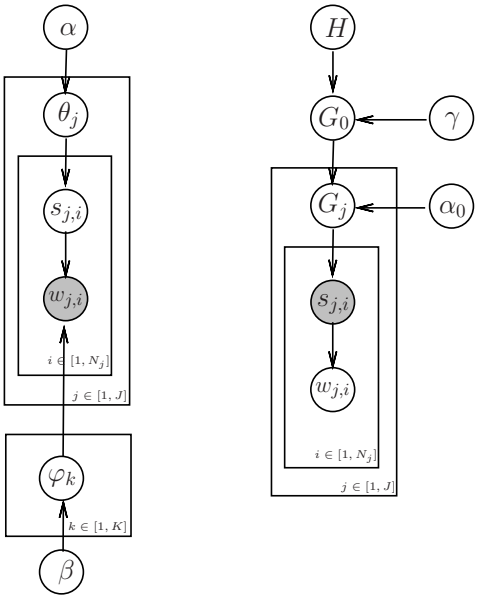

Figure 1: Graphical model representations of (a) Latent Dirichlet Allocation for WSI, (b) Hierarchical Dirichlet Process for WSI.

The HDP/LDA-based WSI complies with the distributional hypothesis that states that words occurring in the same contexts tend to have similar meanings. We want to extend this hypothesis to machine translation by building sense-based translation model upon the HDP-based word sense induction: words with the same meanings tend to be translated in the same way.

2.2 Word Sense Tagging

We adopt the HDP-based WSI to automatically predict word senses and use these predicted senses to annotate source words. We individually build a HDP-based WSI model per word type and train these models on the training data. The sense for a word token is defined as the most probable sense according to the per-document sense distribution G_j estimated for the corresponding pseudo document that represents the surrounding context of the word token. In particular, we take the following steps.

tā tíxǐng wǒguó wǎngluò yùnyíng zhě zhùyì fángfàn hēikè gōngjī ， quèbǎo wǎngluò ānquán 。
Pseudo Documents for word “wǎngluò”
tā tíxǐng wǒguó wǎngluò yùnyíng zhě zhùyì fángfàn hēikè fángfàn hēikè gōngjī ， quèbǎo wǎngluò ānquán 。

Table 1: Examples of pseudo documents extracted from a Chinese sentence (written in Chinese Pinyin).

- **Data preprocessing** We preprocess the source side of our bilingual training data as well as development and test set by removing stop words and rare words.
- **Training Data Sense Annotation** From the preprocessed training data, we extract all possible pseudo documents for each source word type. The collection of these extracted pseudo documents is used as a corpus to train a HDP-based WSI model for the source word type. In this way, we can train as many HDP-based WSI models as the number of word types kept after preprocessing. The sense with the highest probability output by the HDP-based WSI model for each pseudo document is used as the sense cluster to label the corresponding word token.
- **Test/Dev Data Sense Annotation** From the preprocessed test data, we can also extract pseudo documents for each source word type that occur in the test/dev set. Using the trained HDP-based WSI model that correspond to the source word type in question, we can obtain the best sense assignment for each pseudo document of the word type, which in turn is used to annotate the corresponding word token in the test/dev data.

3 Sense-Based Translation Model

In this section we present our sense-based translation model and describe the features that we use as well as the training process of this model.

3.1 Model

The sense-based translation model estimates the probability that a source word c is translated into a target phrase $\tilde{e}$ given contextual information, including word senses that are obtained using the HDP-based WSI as described in the last section. We allow the target phrase $\tilde{e}$ to be either a phrase of length up to 3 words or NULL so that we can capture both multi-word and null translations. The essential component of the model is a maximum entropy (MaxEnt) based classifier that is used to predict the translation probability $p(\tilde{e}|\mathcal{C}(c))$. The MaxEnt classifier can be formulated as follows.

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

where h_is are binary features, θ_is are weights of these features, $\mathcal{C}(c)$ is the surrounding context of c.

We define two groups of binary features: 1) **lexicon features** and 2) **sense features**. All these features take the following form.

$$h(\tilde{e}, \mathcal{C}(c)) = \begin{cases} 1, & \text{if } \tilde{e} = \square \text{ and } \mathcal{C}(c).\mu = \nu \\ 0, & \text{else} \end{cases} \quad (2)$$

where $\square$ is a placeholder for a possible target translation (up to 3 words or NULL), μ is the name of a contextual (lexicon or sense) feature for the source word c, and the symbol ν represents the value of the feature μ.

We extract both the lexicon and sense features from a $\pm k$-word window centered on the word c. The lexicon features are defined as the preceding k words, the succeeding k words and the word c itself: $\{c_{-k}, ..., c_{-1}, c, c_1, ..., c_k\}$. The sense features are defined as the predicted senses for these words: $\{s_{c_{-k}}, ..., s_{c_{-1}}, s_c, s_{c_1}, ..., s_{c_k}\}$.

As we also use these neighboring words to predict word senses in the HDP-based WSI, the information provided by the lexicon and sense features may overlap. This is not a issue for the MaxEnt classifier as it can deal with arbitrary overlapping features (Berger et al., 1996). One may also wonder whether the sense features can contribute to SMT new information that can NOT be obtained from the lexicon features. First, we believe that the senses induced by the HDP-based WSI provide a different view of data than that of the lexicon features. Second, the sense features contain semantic distributional information learned by the HDP across contexts where lexical words occur. Third, we empirically investigate this doubt by comparing two MaxEnt-based translation models

in Section 5. One model only uses the lexicon features while the other integrates both the lexicon and sense features. The former model can be considered as a reformulated WSD for SMT as we described in Section 1.

Given a source sentence $\{c_i\}_1^I$, the proposed sense-based translation model M_s can be denoted as

$$M_s = \prod_{c_i \in \mathcal{W}} (\tilde{e_i}|\mathcal{C}(c_i)) \quad (3)$$

where $\mathcal{W}$ is a set of words for which we build MaxEnt classifiers (see the next subsection for the discussion on how we build MaxEnt classifiers for our sense-based translation model).

3.2 Training

The training of the proposed sense-based translation model is a process of estimating the feature weights θs in the equation (1). There are two strategies that we can use to obtain these weights. We can either build an all-in-one MaxEnt classifier that integrates all source word types c and their possible target translations $\tilde{e}$ or build multiple MaxEnt classifiers. If we train the all-in-one classifier, we have to predict millions of classes (target translations of length up to 3 words). This is normally intractable in practice. Therefore we take the second strategy: building multiple MaxEnt classifiers with one classifier per source word type.

In order to train these classifiers, we have to collect training events from our word-aligned bilingual training data where source words are annotated with their corresponding sense clusters predicted by the HDP-based WSI as described in Section 2. A training event for a source word c consists of all contextual elements in the form of $\mathcal{C}(c).\mu = \nu$ defined in the last subsection and the target translation $\tilde{e}$. Using these collected events, we can train our multiple classifiers. In practice, we do not build MaxEnt classifiers for source words that occur less than 10 times in the training data and run the MaxEnt toolkit in a parallel manner in order to expedite the training process.

4 Decoding with Sense-Based Translation Model

The sense-based translation model described above is integrated into the log-linear translation model of SMT as a sense-based knowledge source. The weight of this model is tuned by the minimum error rate training (MERT) (Och, 2003) together with other models such as the language model.

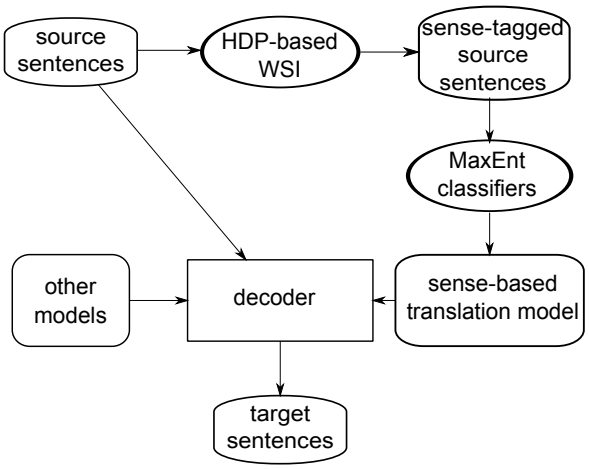

Figure 2: Architecture of SMT system with the sense-based translation model.

Figure 2 shows the architecture of the SMT system enhanced with the sense-based translation model. Before we translate a source sentence, we use the HDP-based WSI models trained on the training data to predict senses for word tokens occurring in the source sentence as discussed in Section 2.2. Note that the HDP-based WSI does not predict senses for all words due to the following two reasons.

- We do not train HDP-based WSI models for word types for which we extract more than T pseudo documents.[2]
- In the test/dev set, there are some words that are unseen in the training data. These unseen words, of course, do not have their HDP-based WSI models.

For these words, we set a default sense (i.e. $s_c = s_1$).

Sense tagging on test sentences can be done in a preprocessing step. Once we get sense clusters for word tokens in test sentences, we load pre-trained MaxEnt classifiers of the corresponding word types. During decoding, we keep word alignments for each translation rule. Whenever a new source word c is translated, we find its translation $\tilde{e}$ via the kept word alignments. We then calculate the translation probability $p(\tilde{e}|\mathcal{C}(c))$ according to the equation (1) using the corresponding loaded classifier. In this way, we can easily calculate the sense-based translation model score.

[2] we set $T = 20,000$.

5 Experiments

In this section, we carried out a series of experiments on Chinese-to-English translation using large-scale bilingual training data. In order to build the proposed sense-based translation model, we annotated the source part of the bilingual training data with word senses induced by the HDP-based WSI. With the trained sense-based translation model, we would like to investigate the following two questions:

- Do word senses automatically induced by the HDP-based WSI improve translation quality?
- Does the sense-based translation model outperform the reformulated WSD for SMT?

5.1 Setup

Our baseline system is a state-of-the-art SMT system which adapts Bracketing Transduction Grammars (Wu, 1997) to phrasal translation and equips itself with a maximum entropy based reordering model (Xiong et al., 2006). We used LDC corpora LDC2004E12, LDC2004T08, LDC2005T10, LDC2003E14, LDC2002E18, LDC2005T06, LDC2003E07, LDC2004T07 as our bilingual training data which consists of 3.84M bilingual sentences, 109.5M English word tokens and 96.9M Chinese word tokens. We ran Giza++ on the training data in two directions and applied the "grow-diag-final" refinement rule (Koehn et al., 2003) to obtain word alignments. From the word-aligned data, we extracted weighted phrase pairs to generate our phrase table. We trained a 5-gram language model on the Xinhua section of the English Gigaword corpus (306 million words) using the SRILM toolkit (Stolcke, 2002) with the modified Kneser-Ney smoothing (Chen and Goodman, 1996).

We trained our HDP-based WSI models via the C++ HDP toolkit[3] (Wang and Blei, 2012). We set the hyperparameters $\gamma = 0.1$ and $\alpha_0 = 1.0$ following Lau et al. (2012).We extracted pseudo documents from a $\pm$10-word window centered on the corresponding word token for each word type following Brody and Lapata (2009). As described in Section 2.2, we preprocessed the source part of our bilingual training data by removing stop words and infrequent words that occurs less than 10 times in the training data. From the preprocessed data, we extracted pseudo documents for each word type to train a HDP-based WSI model per word type. Note that we do not build WSI models for highly frequent words that occur more than 20,000 times in order to expedite the HDP training process.

We trained our MaxEnt classifiers with the off-the-shelf MaxEnt tool.[4] We performed 100 iterations of the L-BFGS algorithm implemented in the training toolkit on the collected training events from the sense-annotated data as described in Section 3.2. We set the Gaussian prior to 1 to avoid overfitting. On average, we obtained 346 classes (target translations) per source word type with the maximum number of classes being 256,243. It took an average of 57.5 seconds for training a Maxent classifier.

We used the NIST MT03 evaluation test data as our development set, and the NIST MT05 as the test set. We evaluated translation quality with the case-insensitive BLEU-4 (Papineni et al., 2002) and NIST (Doddington, 2002). In order to alleviate the impact of MERT (Och, 2003) instability, we followed the suggestion of Clark et al. (2011) to run MERT three times and report average BLEU/NIST scores over the three runs for all our experiments.

	Training	Test
# Word Types	67,723	4,348
# Total Pseudo Documents	27.73M	11,777
# Avg Pseudo Documents	427.79	2.71
# Total Senses	271,770	24,162
# Avg Senses	4.01	5.56

Table 2: Statistics of the HDP-based word sense induction on the training and test data.

5.2 Statistics and Examples of Word Senses

Before we present our experiment results of the sense-based translation model, we study some statistics of the HDP-based WSI on the training and test data. We show these statistics in Table 2. There are 67,723 and 4,348 unique word types in the training and test data after the preprocessing step. For these word types, we extract 27.73M and 11,777 pseudo documents from the training and test set respectively. On average, there are 427.79

[3] http://www.cs.cmu.edu/~chongw/resource.html

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

System	BLEU(%)	NIST
STM (±5w)	34.64	9.4346
STM (±10w)	**34.76**	**9.5114**
STM (±15w)	-	-

Table 4: Experiment results of the sense-based translation model (STM) with lexicon and sense features extracted from a window of size varying from ±5 to ±15 words on the development set.

pseudo documents per word type in the training data and 2.71 in the test set. The HDP-based WSI learns 271,770 word senses in total using the pseudo documents collected from the training data and infers 24,162 word senses using the pseudo documents extracted from the test set. There are 4.01 different senses per word type in the training data and 5.56 in the test set on average.

Table 3 illustrates six different senses of the word "运营 (operate)" learned by the HDP-based WSI in the training data. We also show the most probable 10 words for each sense cluster. Sense s_1 represents the operations of company or organization, sense s_2 denotes country/institution/internation operations, sense s_3 refers to market operations, sense s_4 corresponds to business operations, sense s_5 to public facility operations, and finally s_6 to economy operations.

5.3 Impact of Window Size k used in MaxEnt Classifiers

Our first group of experiments were conducted to investigate the impact of the window size k on translation performance in terms of BLEU/NIST on the development set. We extracted both the lexicon and sense features from a $\pm k$-word window for our MaxEnt classifiers. We varied k from 5 to 15. Experiment results are shown in Table 4. We achieve the best performance when $k = 10$. This suggests that a ±10-word window context is sufficient for predicting target translations for ambiguous source words. We therefore set $k = 10$ for all experiments thereafter.

5.4 Effect of the Sense-Based Translation Model

Our second group of experiments were carried out to investigate whether the sense-base translation model is able to improve translation quality by comparing the system enhanced with our sense-based translation model against the baseline. We also studied the impact of word senses induced by the HDP-based WSI on translation performance by enforcing the sense-based translation model to use only sense features. Table 5 shows the experiment results. From the table, we can observe that

System	BLEU(%)	NIST
Base	33.53	9.0561
STM (sense)	34.15	9.2596
STM (sense+lexicon)	34.73	9.4184

Table 5: Experiment results of the sense-based translation model (STM) against the baseline.

System	BLEU(%)	NIST
Base	33.53	9.0561
Reformulated WSD	34.16	9.3820
STM	34.73	9.4184

Table 6: Comparison results of the sense-based translation model vs. the reformulated WSD for SMT.

- Our sense-based translation model achieves a substantial improvement of 1.2 BLEU points over the baseline. This indicates that the sense-based translation model is able to help select correct translations for ambiguous source words.
- If we only integrate sense features into the sense-based translation model, we can still outperform the baseline by 0.62 BLEU points. This suggests that automatically induced word senses alone are indeed useful for machine translation.

5.5 Comparison to Word Sense Disambiguation

As we mentioned in Section 3.1, our sense-based translation model can be degenerated to a reformulated WSD model for SMT if we only use lexicon features in MaxEnt classifiers. This allows us to directly compare our method against the reformulated WSD for SMT. Table 6 shows the comparison result.

From the table, we can find that the sense-based translation model outperforms the reformulated WSD by 0.57 BLEU points. This suggests that the HDP-based word sense induction is better than the reformulated WSD in the context of SMT. Furthermore, as the reformulated WSD is a degenerated version of our sense-based translation model which only uses the lexicon features,

s_1	s_2	s_3
运营 (operate)	运营 (operate)	运营 (operate)
设施 (facility)	卫星 (satellite)	市场 (market)
计划 (plan)	系统 (system)	企业 (enterprise)
基础 (foundation)	国家 (country)	竞争 (competition)
项目 (project)	提供 (supply)	资产 (assets)
公司 (company)	国际 (inter-nation)	利润 (profit)
结构 (structure)	机构 (institution)	造成 (cause)
服务 (service)	进行 (proceed)	费用 (cost)
组织 (organization)	中心 (center)	资金 (capital)
提供 (supply)	合作 (cooperate)	业务 (business)
s_4	s_5	s_6
费用 (cost)	城市 (city)	处于 (lie)
股价 (share price)	处理 (process)	拍照 (photograph)
27000	自来水 (tap-water)	119
科索沃 (Kosovo)	工厂 (factory)	DPRK
额外 (extra)	汽车 (car)	保险 (insurance)
工资 (wage)	铁路 (railway)	超支 (overspend)
美元 (dollar)	污水 (sewage)	地位 (position)
商业 (commerce)	办事处 (office)	经济 (economy)
收入 (income)	保本 (break-even)	竞争者 (competitor)
铁路局 (railway administration)	部件 (component)	平衡 (balance)

Table 3: Six different senses learned for the word "运营" from the training data.

the sense features used in our model do provide new information that can not be obtained by the lexicon features.

6 Related Work

In this section we introduce previous studies that are related to our work. For ease of comparison, we roughly divide them into 4 categories: 1) WSD for SMT, 2) topic-based WSI, 3) topic model for SMT and 4) lexical selection.

WSD for SMT As we mentioned in Section 1, WSD has been successfully reformulated and adapted to SMT (Vickrey et al., 2005; Carpuat and Wu, 2007; Chan et al., 2007). Rather than predicting word senses for ambiguous words, the reformulated WSD directly predicts target translations for source words with context information. Our sense-based translation model also predicts target translations for SMT. The significant difference is that we predict word senses automatically learned from data and incorporate these predicted senses into SMT. Our experiments show that such word senses are able to improve translation quality.

Topic-based WSI Topic-based WSI can be considered as the foundation of our work as we use it to obtain broad-coverage word senses to annotate our large-scale training data. Brody and Lapata (2009)'s work is the first attempt to approach WSI via topic modeling. They adapt LDA to word sense induction by building one topic model per word type. According to them, there are 3 significant differences between topic-based WSI and generic topic modeling.

- First, the goal of topic-based WSI is to divide contexts of a word type into different categories, each representing a sense cluster. However generic topic models aim at topic distributions of documents.
- Second, generic topic modeling explores whole documents for topic inference while topic-based WSI uses much smaller units in a document (e.g., surrounding words of a target word) for word sense induction.
- Finally, the number of induced word senses in WSI is usually less than 10 while the number of inferred topics in generic topic modeling is tens or hundreds.

As LDA-based WSI needs to manually specify the number of word senses, Yao and Durme (2011) propose HDP-based WSI that is capable of

determining the number of senses for each word type according to training data. Lau et al. (2012) adopt the HDP-based WSI for novel sense detection and empirically show that the HDP-based WSI is better than the LDA-based WSI. We follow them to set the hyperparameters of HDP for training and incorporate automatically induced word senses into SMT in our work.

Topic model for SMT Generic topic models are also explored for SMT. Zhao and Xing (2007) propose a bilingual topic model and integrate a topic-specific lexicon translation model into SMT. Tam et al. (2007) also explore a bilingual topic model for translation and language model adaptation. Foster and Kunh (2007) introduce a mixture model approach for translation model adaptation. Xiao et al. (2012) propose a topic-based similarity model for rule selection in hierarchical phrase-based translation. Xiong and Zhang (2013) employ a sentence-level topic model to capture coherence for document-level machine translation. The difference between our work and these previous studies on topic model for SMT lies in that we adopt topic-based WSI to obtain word senses rather than generic topics and integrate induced word senses into machine translation.

Lexical selection Our work is also related to lexical selection in SMT where appropriate target lexical items for source words are selected by a statistical model with context information (Bangalore et al., 2007; Mauser et al., 2009). The reformulated WSD discussed before can also be considered as a lexical selection model. The significant difference from these studies is that we perform lexical selection using automatically induced word senses by the HDP on the source side.

7 Conclusion

We have presented a sense-based translation model that integrates word senses into machine translation. We capitalize on the broad-coverage word sense induction system that is built on the nonparametric Bayesian HDP to learn sense clusters for words in the source language. We generate pseudo documents for word tokens in the training/test data for the HDP-based WSI system to infer topics. The most probable topic inferred for a pseudo document is taken as the sense of the corresponding word token. We incorporate these learned word senses as translation evidences into maximum entropy classifiers which form the foundation of the proposed sense-based translation model.

We carried out a series of experiments to validate the effectiveness of the sense-based translation by comparing the model against the baseline and the previous reformulated WSD. Our experiment results show that

- The sense-based translation model is able to substantially improve translation quality in terms of both BLEU and NIST.
- The sense-based translation model is also better than the previous reformulated WSD for SMT.
- Word senses automatically induced by the HDP-based WSI on large-scale training data are very useful for machine translation. To the best of our knowledge, this is the first attempt to empirically verify the positive impact of word senses on translation quality.

Comparing with macro topics of documents inferred by LDA with bag of words from the whole documents, word senses inferred by the HDP-based WSI can be considered as micro topics. In the future, we would like to explore both the micro and macro topics for machine translation. Additionally, we also want to induce sense clusters for words in the target language so that we can build sense-based language model and integrate it into SMT. We would like to investigate whether automatically learned senses of proceeding words are helpful for predicting succeeding words.

Acknowledgement

The work was sponsored by the National Natural Science Foundation of China under projects 61373095 and 61333018. We would like to thank three anonymous reviewers for their insightful comments.

References

Srinivas Bangalore, Patrick Haffner, and Stephan Kanthak. 2007. Statistical Machine Translation through Global Lexical Selection and Sentence Reconstruction. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 152–159, Prague, Czech Republic, June. Association for Computational Linguistics.

Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. 1996. A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, 22(1):39–71.

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

Samuel Brody and Mirella Lapata. 2009. Bayesian Word Sense Induction. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 103–111, Athens, Greece, March. Association for Computational Linguistics.

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 (ACL'05)*, pages 387–394, Ann Arbor, Michigan, June. 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 (EMNLP-CoNLL)*, 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 of Computational Linguistics*, pages 33–40, Prague, Czech Republic, June. Association for Computational Linguistics.

Stanley F. Chen and Joshua Goodman. 1996. An Empirical Study of Smoothing Techniques for Language Modeling. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, ACL '96, pages 310–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

David Chiang. 2005. A Hierarchical Phrase-Based Model for Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 263–270, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better Hypothesis Testing for Statistical Machine Translation: Controlling for Optimizer Instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 176–181, Portland, Oregon, USA, June.

George Doddington. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, pages 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

George Foster and Roland Kuhn. 2007. Mixture-Model Adaptation for SMT. In *Proc. of the Second Workshop on Statistical Machine Translation*, pages 128–135, Prague, Czech Republic, June.

Philipp Koehn, Franz Joseph Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 58–54, Edmonton, Canada, May-June.

Jey Han Lau, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin. 2012. Word Sense Induction for Novel Sense Detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 591–601, Avignon, France, April. Association for Computational Linguistics.

Arne Mauser, Saša Hasan, and Hermann Ney. 2009. Extending Statistical Machine Translation with Discriminative and Trigger-Based Lexicon Models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 210–218, Singapore, August. Association for Computational Linguistics.

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*, pages 160–167, Sapporo, Japan, July. 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 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July.

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

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2004. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101.

David Vickrey, Luke Biewald, Marc Teyssier, and Daphne Koller. 2005. Word-Sense Disambiguation for Machine Translation. In *HLT/EMNLP*. The Association for Computational Linguistics.

C. Wang and D. M. Blei. 2012. A Split-Merge MCMC Algorithm for the Hierarchical Dirichlet Process. *ArXiv e-prints*, January.

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 (Volume 1: Long Papers)*, pages 750–758, Jeju Island, Korea, July. Association for Computational Linguistics.

Deyi Xiong and Min Zhang. 2013. A Topic-Based Coherence Model for Statistical Machine Translation. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence (AAAI-13)*, Bellevue, Washington, USA, July.

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 44th Annual Meeting of the Association for Computational Linguistics*, pages 521–528, Sydney, Australia, July.

Xuchen Yao and Benjamin Van Durme. 2011. Nonparametric Bayesian Word Sense Induction. In *Proceedings of TextGraphs-6: Graph-based Methods for Natural Language Processing*, pages 10–14, Portland, Oregon, June. Association for Computational Linguistics.

Bin Zhao and Eric P. Xing. 2007. HM-BiTAM: Bilingual Topic Exploration, Word Alignment, and Translation. In *Proc. NIPS 2007*.