# Instant Translation Model Adaptation by Translating Unseen Words in Continuous Vector Space

Shonosuke Ishiwatari[1], Naoki Yoshinaga[2,3], Masashi Toyoda[2], and Masaru Kitsuregawa[2,4]

[1] Graduate School of Information Science and Technology, The University of Tokyo, Japan
[2] Institute of Industrial Science, The University of Tokyo, Japan
[3] National Institute of Information and Communications Technology, Japan
[4] National Institute of Informatics, Japan
{ishiwatari, ynaga, toyoda, kitsure}@tkl.iis.u-tokyo.ac.jp

**Abstract.** In statistical machine translation (SMT), differences between domains of training and test data result in poor translations. Although there have been many studies on domain adaptation of language models and translation models, most require supervised in-domain language resources such as parallel corpora for training and tuning the models. The necessity of supervised data has made such methods difficult to adapt to practical SMT systems. We thus propose a novel method that adapts translation models without in-domain parallel corpora. Our method infers translation candidates of unseen words by nearest-neighbor search after projecting their vector-based semantic representations to the semantic space of the target language. In our experiment of out-of-domain translation from Japanese to English, our method improved BLEU score by 0.5-1.5.

## 1 Introduction

Statistical machine translation (SMT) has been successfully applied to the translation between various language pairs, particularly phrase-based SMT, which is the most common since it can learn a translation model from a sentence-aligned parallel corpus without any linguistic annotations. Although we can improve the quality of translation by using a large language model that can be obtained from easily available monolingual corpora [1], language models capture only the fluency in languages so the quality of translation cannot be improved much if the translation model does not provide correct translation candidates for source-language words and phrases. The quality of translation in SMT is therefore bounded by the size of parallel corpus to train the translation model. Even if a large parallel corpus is available for the pair of languages in question, we often want to translate sentences in a domain that has a different vocabulary from the domain of available parallel corpora, and this inconsistency deteriorates the quality of translation [2, 3].

Researchers have tackled this problem and proposed methods of domain adaptation for SMT that exploits a larger out-of-domain parallel corpus. They have focused on a scenario in which a small or pseudo in-domain parallel corpus is available for training [4]. In actual scenarios when users want to exploit machine translation, the target domains can differ so the domain mismatches between the prepared SMT system and

the target documents are likely to occur. Domain adaptation is thus expected to improve the quality of translation. However, it is unrealistic for most MT users who cannot command the target language to prepare in-domain parallel corpora by themselves. The use of crowdsourcing for preparing in-domain parallel corpora is allowed for a few users who have a large number of documents for translation and are willing to pay money for improving the quality of translation.

In this study, we assume domain adaptation for SMT in a scenario where no sentence-aligned parallel corpus is available for the target domain and propose an instant method of domain adaptation for SMT by using a cross-lingual projection of word semantic representations [5]. Assuming that source- and target-language monolingual corpora are available, we first learn vector-based semantic representations of words in the source and target languages from those monolingual corpora. We next obtain a projection from semantic representations in the source language to those in the target language using a seed dictionary (in general domain) to learn a translation matrix. We then use the translation matrix to obtain translations of unseen (out-of-vocabulary, OOV) words. The translation probabilities are computed by using cosine-similarity between the projected semantic representation of the OOV word and semantic representations of words in the target language.

To evaluate the effectiveness of our method, we apply our method to a translation between English (en) and Japanese (ja) in recipe documents using a translation model learned by phrase-based SMT from Kyoto-related Wikipedia articles. Experimental results confirmed that our method improves BLEU score by 0.5-1.5 and 0.1-0.2 for ja-en and en-ja translations, respectively.

The remainder of this paper is structured as follows. Section 2 explains existing approaches to domain adaptation for SMT without in-domain parallel corpus. Section 3 describes a method of translating word semantic representations. Section 4 proposes a method of adapting SMT to a new domain without a sentence-aligned parallel corpora. Section 5 evaluates the effectiveness of the proposed method on domain adaptation for SMT. Section 6 finally concludes this study and addresses future work.

## 2 Related work

As mentioned in Section 1, most previous approaches to domain adaptation for SMT assume a scenario where a small or pseudo in-domain parallel corpus is available. In this section, we briefly overview a method of domain adaptation for SMT in a setting where no in-domain parallel corpus is available.

Wu et al. [6] have proposed domain adaptation for SMT that exploits an in-domain bilingual dictionary. They generate a translation model from the bilingual dictionary and combine it with the translation model learned from out-of-domain parallel corpora. An issue here is how to learn a translation probability between words (or phrases) needed for the translation model, and they resort to probabilities of words in the target language in a monolingual corpus. Although building a bilingual dictionary for the target domain is more effective than developing a parallel corpus to cover rare OOV words, it is still difficult to develop a bilingual dictionary for most MT users who cannot command the target language.

To cope with this problem, several researchers have recently exploited a bilingual lexicon automatically induced from in-domain corpora to generate a translation model for SMT [7–9]. These approaches induce a bilingual lexicon from in-domain comparable corpora prior to the translation and use it to obtain an in-domain translation model.

Marthur et al. [10] exploit parallel corpora in various domains to induce the translation model for the target domain. They used 11 sets of parallel corpora for domains including TED talks, news articles, and software manuals to train the translation model for each domain and then linearly interpolated these translation models to derive a translation model for the target domain. They successfully improved the quality of translation when no parallel corpus was available for the target domain. Yamamoto and Sumita [11] assume various language expressions in translating travel conversations and train several language and translation models from a set of parallel corpora that are split by unsupervised clustering of the entire parallel corpus for travel conversations. The language and translation models for translating a given sentence are chosen in accordance with the similarity between the given sentence and the sentences in each split of the parallel corpus. Although this method is not intended for domain adaptation, it can be used in our setting when we have a parallel corpus for the general domain (and the domain of the target sentence is included in the general domain). These studies, however, implicitly assume in-domain (or related domain) parallel corpora are available, while we assume those resources are unavailable to broaden the applicability of our method.

Among these studies, our method is most closely related to domain adaptation using bilingual lexicon induction [7–9] but is different from these approaches in that it does not need to build a sort of bilingual lexicon prior to the translation to support the translation of OOV words in a given sentence. We use a projection of semantic representations of source-language words to the target-language semantic space to dynamically find translation candidates of found OOV words by computing the similarity of the obtained representations to semantic representations for words in the target language at the time of translation. Also, we empirically show that our approach could even benefit from general-domain non-comparable monolingual corpora instead of in-domain comparable monolingual corpora used in these studies on bilingual lexicon induction.

## 3 Cross-lingual projection of word semantic representations

Our method exploits a projection of semantic representations of OOV words in the source-language onto the target-language semantic space to look for translation candidates for the OOV words. In this section, we first introduce semantic representations of words in a continuous vector space and then describe a method we proposed previously that learns a translation matrix for projecting vector-based representations of words across languages [5].

A vector-based semantic representation of a word, hereinafter *word vector*, represents the meaning of a word with a continuous vector. These representations are based on the distributional hypothesis [12, 13], which states that words that occur in the similar contexts tend to have similar meanings. The word vectors can be obtained from monolingual corpora in an unsupervised manner, such as a count-based approach [14] or prediction-based approaches [15, 16].

The words that have similar meanings tend to have similar vectors [17, 18]. By mapping words into continuous vector space, we can use cosine similarity to compute the similarity of meanings between words. However, the similarity between word vectors across languages is difficult to compute, so these word vectors are difficult to utilize in cross-lingual applications such as machine translation or cross-lingual information retrieval.

To solve this problem, Mikolov et al. [19] proposed a method that learns a cross-lingual projection of word vectors from one language into another. By projecting a word vector into the target-language semantic space, we can compute the semantic similarity between words in different languages. Suppose that we have training data of $n$ examples, $\{(\boldsymbol{x}_1, \boldsymbol{z}_1), (\boldsymbol{x}_2, \boldsymbol{z}_2), \ldots (\boldsymbol{x}_n, \boldsymbol{z}_n)\}$, where $\boldsymbol{x}_i$ is the vector representation of a word in the source language (*e.g.*, "*gato*"), and $\boldsymbol{z}_i$ is the word vector of its translation in the target language (*e.g.*, "*cat*"). Then the translation matrix, $\boldsymbol{W}$, such that $\boldsymbol{W}\boldsymbol{x}_i$ approximates $\boldsymbol{z}_i$, can be obtained by solving the following optimization problem:

$$\boldsymbol{W}^\star = \operatorname*{argmin}_{\boldsymbol{W}} \sum_{i=1}^{n} \|\boldsymbol{W}\boldsymbol{x}_i - \boldsymbol{z}_i\|^2$$

Here, since word vectors are induced from monolingual corpora, vectors of OOV words are easy to obtain by using in-domain or large-scale monolingual corpora.

We have improved the aforementioned approach by adopting the count-based vectors for words and integrating prior knowledge on translatable context pairs between the dimensions of count-based vectors [5]:

$$\boldsymbol{W}^\star = \operatorname*{argmin}_{\boldsymbol{W}} \sum_{i=1}^{n} \|\boldsymbol{W}\boldsymbol{x}_i - \boldsymbol{z}_i\|^2 + \frac{\lambda}{2}\|\boldsymbol{W}\|^2 - \beta_{train} \sum_{(j,k)\in\mathcal{D}_{train}} w_{jk} - \beta_{sim} \sum_{(j,k)\in\mathcal{D}_{sim}} w_{jk}.$$

The second term is the $L_2$ regularizer, while the third and fourth terms are meant to strengthen $w_{jk}$ when $k$-th dimension in the source language corresponds to $j$-th dimension in the target language. $\mathcal{D}_{train}$ and $\mathcal{D}_{sim}$ are sets of translatable dimension pairs. $\mathcal{D}_{train}$ is obtained from the above training data, while $\mathcal{D}_{sim}$ is obtained by computing the surface-level similarity between the dimensions. $\lambda$, $\beta_{train}$ and $\beta_{sim}$ are corresponding hyperparameters to control the strength of the added terms.

Because our method improved the accuracy of choosing translation candidates for words using the projected semantic representation against [19, 20], we adopt and implement this method again for finding translation candidates of OOV words in our method.

## 4 Method

Our method assumes that monolingual corpora are available for the source and target language (in the target domain, if any) and first induces semantic representation of words from those corpora. It then learns a cross-lingual projection (translation matrix) using a seed dictionary in a general domain as described in Section 3. Note that a seed dictionary for common words is usually available for most pairs of languages or could be constructed assuming English as a pivot language [21].

Having a translation matrix to obtain projections of semantic representations of OOV words in a given sentence, our method instantly constructs a back-off translation model used for enumerating translation candidates for the OOV words in the following way:

**Step 1:** When the translation system accepts a sentence with an OOV word, $f_{\text{OOV}}$, it translates a semantic representation of the word, $\boldsymbol{x}_{\text{OOV}}$ into a semantic representation in the target language $\boldsymbol{x}'_{\text{OOV}}$ using the translation matrix obtained by the method described in Section 3.

**Step 2:** It then computes the cosine similarity between the obtained semantic representations with those in the target languages to enumerate $k$ translation candidates[1] in accordance with the value of cosine similarity. The cosine similarity is also used to obtain $P_{vec}(e|f_{\text{OOV}})$, the direct translation probabilities from the OOV word in the source language, $f_{\text{OOV}}$, to a candidate word in the target language, $e$, by normalizing them to sum up to 1. Although the obtained translation candidates could include wrong translations, the language model can choose one that is more appropriate in the contexts in the next step, unless the contexts are full of OOV words.

**Step 3:** The decoder of phrase-based SMT uses the above translation probabilities as a back-off translation model to perform the translation. More formally, we add new feature function $h_{vec}$ to the log-linear model used in the decoder as following equation:

$$\log P(\boldsymbol{e}|\boldsymbol{f}) = \sum_i \log(h_i(\boldsymbol{e},\boldsymbol{f}))\lambda_i + \log(h_{vec}(\boldsymbol{e},\boldsymbol{f}))\lambda_{vec} \quad (1)$$

The $h_{vec}(\boldsymbol{e},\boldsymbol{f})$ in Eq. (1) is computed with $P_{vec}(e|f_{\text{OOV}})$, only for each OOV word $f_{\text{OOV}}$ in source sentence $\boldsymbol{f}$. An issue here is how to set feature weight $\lambda_{vec}$ since no in-domain training data are available for turning. We simply set $\lambda_{vec}$ to the same value as the weight of direct phrase translation probability of the translation model.

## 5 Experiments

This section evaluates our method of domain adaptation for SMT, using an out-of-domain parallel corpus and source-language and target-language monolingual corpora.

### 5.1 Settings

First, we prepared two parallel corpora in different domains to carry out an experiment of domain adaptation in the SMT system. One is the "Japanese-English Bilingual Corpus of Wikipedia's Kyoto Articles" (hereinafter KFTT corpus), originally prepared by the National Institute of Information and Communications Technology (NICT) and used as a benchmark in "The Kyoto Free Translation Task"[2][22], a translation task that focuses on Wikipedia articles relates to Kyoto. The other parallel corpus (hereafter RECIPE corpus) is provided by Cookpad Inc.,[3] which is the largest online recipe sharing service

[1] $k$ was set to 10 in the experiments.

[2] `http://www.phontron.com/kftt/`

[3] `http://cookpad.com/`

**Table 1.** Statistics of the dataset.

| Corpus | Japanese | English |
|---|---|---|
| KFTT (training) | 29.5MB (440k sentences) | 30.6MB (440k sentences) |
| RECIPE (test) | 0.8MB (10k sentences) | 0.7MB (10k sentences) |

**Table 2.** Monolingual corpora used to induce semantic representations.

| Corpus | Japanese | English |
|---|---|---|
| Wikipedia (general domain) | 4.4GB | 16GB |
| RECIPE (in-domain) | 12MB | 9.5MB |

in Japan. The KFTT corpus includes many words relates to Japanese history and the temples or shrines in Kyoto. On the other hand, the RECIPE corpus includes many words related to foods and cookware. We randomly sampled 10k pairs of sentences from the RECIPE corpus as test corpus for evaluating our domain adaption method. The language models of the target languages are trained with the concatenation of the KFTT corpus and the remaining portion of the RECIPE corpus, while the translation models are trained with only the KFTT corpus. The sizes of the training data and test data are as detailed in Table 1.

We conducted experiments with Moses [23][4] with the language models trained with SRILM [24][5] and the word alignments predicted by GIZA++ [25].[6] 5-gram language models were trained using SRILM with `interpolate` option and `kndiscount` option. Word alignments were obtained using GIZA++ with `grow-diag-final-and` heuristic. The lexical reordering model was obtained with `msd-bidirectional` setting.

Next, we extracted four sets of count-based word vectors from Wikipedia dumps[7] (general-domain monolingual corpora) and the remaining portion of the RECIPE corpus (in-domain monolingual corpora), for Japanese and English, respectively. We considered context windows of five words to both sides of the target word. The function words are then excluded from the extracted context words following our previous work [5]. Since the count vectors are very high-dimensional and sparse, we selected top-$d$ ($d = 10,000$ for general-domain corpus, $d = 5000$ for in-domain corpus) frequent words as contexts words (in other words, the number of dimensions of the word vectors). We converted the counts into positive point-wise mutual information [26] and normalized the resulting vectors to remove the bias introduced by the difference in the word frequency. The size of the monolingual dataset for inducing semantic representations of words is as detailed in Table 2.

[4] `http://www.statmt.org/moses/`
[5] `http://www.speech.sri.com/projects/srilm/`
[6] `https://github.com/moses-smt/giza-pp`
[7] `http://dumps.wikimedia.org/` (versions of Nov, 4th, 2014 (ja), Oct, 8th,2014 (en).

**Table 3.** BLEU on RECIPE corpus. * indicates statistically significant improvements in BLEU over the respective baseline systems in accordance with bootstrap resampling [27] at $p < 0.05$.

| Method | All | | OOV sentences | |
|---|---|---|---|---|
| | ja-en | en-ja | ja-en | en-ja |
| **Baseline (no adaptation)** | 5.58 | 3.37 | 5.36 | 3.16 |
| **Proposed (general-domain)** | 6.05* | 3.48* | 5.87* | 3.42* |
| **Proposed (in-domain)** | **7.08*** | **3.57*** | **7.00*** | **3.63*** |
| **Parallel Corpus** | 20.88 | 16.69 | 20.72 | 17.01 |

**Table 4.** Statistics of the OOV words in test data (the 10k sentences in the RECIPE corpus).

| | ja-en | en-ja |
|---|---|---|
| The number of OOV words (types) | 3,464 | 1,613 |
| The number of OOV words (tokens) | 21,218 | 4,639 |
| The number of sentences with OOV words | 8,742 | 3,636 |

Finally, we used Open Multilingual WordNet[8] to train the translation matrices as in [5]. The hyperparameters were tuned on the development set as follows: $\lambda = 0.1$, $\beta_{train} = 5$, $\beta_{sim} = 5$ for (ja-en, general-domain). $\lambda = 1$, $\beta_{train} = 0.1$, $\beta_{sim} = 0.2$ for (ja-en, in-domain). $\lambda = 0.1$, $\beta_{train} = 5$, $\beta_{sim} = 5$ for (en-ja, general-domain). $\lambda = 0.5$, $\beta_{train} = 1$, $\beta_{sim} = 2$ for (en-ja, in-domain).

### 5.2 Results

We performed domain adaptation as described in Section 4 and evaluated the effectiveness of our method through BLEU score [28]. Table 3 shows results of the translations of the 10k sentences in the RECIPE corpus between Japanese and English. **All** and **OOV sentences** in Table 3 show the BLEU scores measured in the whole test set and the scores measured only in the sentences that include OOV words, respectively. Statistics of the OOV words are shown in Table 4.

All four methods shown in Table 3 use translation models that were trained with the KFTT corpus and are tested with the RECIPE corpus. **Proposed (general)** uses the word vectors extracted from Wikipedia corpus, while **Proposed (in-domain)** uses the vectors extracted from the remaining portion of the RECIPE corpus. In both these methods, we performed domain adaptation by automatically constructing back-off translation models for OOV words. **Parallel Corpus** in Table 3 uses the remaining portion of the RECIPE corpus as a parallel corpus to learn the translation models, resources of which are assumed to be unavailable in this study. Thus, **Parallel Corpus** is the upper-bound for the task. The low BLEU score for en-ja translation is explained by the direction of the translation being different from the direction when the corpus was built (ja-en) [29].

[8] http://compling.hss.ntu.edu.sg/omw/

**Table 5.** Hand-picked examples of the translations for the 10k sentences in the RECIPE corpus from Japanese to English. Text in bold denotes OOV words in the input sentences and their translations. The subscripts of the translation of the OOV words refer to a manual word alignment of the OOV words.

| | |
|---|---|
| Input | 混ぜながら弱火で煮る。 |
| Ref | simmer over low heat while mixing . |
| **Baseline** | **煮る**$_1$ at low heat while mixing . |
| **Proposed (general)** | **boil**$_1$ over a low heat while mixing . |
| **Proposed (in-domain)** | **simmer**$_1$ over a low heat while mixing . |
| **Parallel Corpus** | **simmer**$_1$ over low heat while stirring . |
| Input | 玉ねぎ、ニンニクをみじん切りに。 |
| Ref | finely chop the onion and garlic . |
| **Baseline** | **みじん切り**$_1$ in the onion and garlic . |
| **Proposed (general)** | the garlic and onion in **butter**$_1$ . |
| **Proposed (in-domain)** | **mince**$_1$ the onion and garlic . |
| **Parallel Corpus** | **finely**$_1$**chop**$_1$ the onion and garlic . |
| Input | オーブントースターで焦げ目がつくまで焼く。 |
| Ref | bake until browned in a toaster oven . |
| **Baseline** | in **トースター**$_1$ oven until **焦げ目**$_2$ made **焼く**$_3$ . |
| **Proposed (general)** | oven in the **refrigerator**$_1$ until **fenbuconazole**$_2$ made **bread**$_3$ . |
| **Proposed (in-domain)** | in a **toaster**$_1$ oven , **bake**$_3$ until the **end**$_2$ . |
| **Parallel Corpus** | **bake**$_3$ in a **toaster**$_1$ oven until **golden**$_2$**brown**$_2$ . |
| Input | しっとりした食感の素朴なケーキです。 |
| Ref | a simple cake with a moist texture . |
| **Baseline** | **しっとり**$_1$ food of a simple cake **です**$_2$ . |
| **Proposed (general)** | the food **texture**$_1$ as a simple cake thing . |
| **Proposed (in-domain)** | the **moist**$_1$ food that 's simple cake . |
| **Parallel Corpus** | a **moist**$_1$**texture**$_1$ of the simple cake . |
| Input | 火を消し、ごま油を入れ混ぜる。 |
| Ref | turn off the heat , and stir in the sesame oil . |
| **Baseline** | **消し**$_1$ fire , and put **ごま油**$_2$ **混ぜる**$_3$ . |
| **Proposed (general)** | heat **butter**$_1$ completely , add the **milk**$_2$ . |
| **Proposed (in-domain)** | fire , add **coconut**$_2$ , and **mix**$_3$ . |
| **Parallel Corpus** | **turn**$_1$**off**$_1$ the heat , add the **sesame**$_2$**oil**$_2$ and **mix**$_3$ . |

In addition, the smaller number of OOV tokens in en-ja than in ja-en also causes the smaller improvement in BLEU score.

Table 3 shows that our methods perform well for the translation task. We found that it was better to use the in-domain monolingual corpora rather than general-domain monolingual corpora to obtain the word vectors. This conforms to our expectation because the contextual information included in the word vectors strongly correlates with the target domains. The **Parallel Corpus** has much higher BLEU than all other methods. This result shows that the domain adaptation task we performed was intrinsically difficult because of the significant differences between the two domains.

We show hand-picked examples of the translations in Table 5 to analyze the methods in more detail. The first two examples show that **Proposed (in-domain)** provides

more accurate translations than **Proposed (general)**. Despite our method being able to improve the translations of OOV words, the third and the fourth examples indicate that it is not good at improving the translations of **Baseline** that have wrong syntax. The last example shows that some OOV words tend to be translated into their related words, mainly because of their similarity in the semantic space.

The examples show that the OOV words such as "煮る" (simmer), "トースター" (toaster), and "焼く" (bake) could successfully be translated with **Proposed (in-domain)**. These words almost never appear in the KFTT corpus, since they do not have any relation with Japanese history or the temples in Kyoto. By comparing **Proposed (in-domain)** and **Proposed (general)**, we see that the latter method translated many OOV words into related words (e.g., "トースター" (toaster) to "refrigerator", or "煮る" (simmer) to "boil") by mistake. This result also indicates that the word vectors extracted from the in-domain corpus will work better than the vectors extracted from the general-domain corpus.

## 6 Conclusions

A cross-lingual projection of word semantic representations has been leveraged to obtain a translation model for unseen (out-of-vocabulary, OOV) words in domain adaptation for SMT. Assuming monolingual corpora for the source and target languages, we induce vector-based semantic representations of words and obtain a projection (translation matrix) from source-language semantic representations into the target-language semantic space. We use this projection to find translation candidates of OOV words and use the cosine similarity to induce the translation probability. Experimental results on domain adaptation from a Kyoto-related domain to a recipe domain confirmed that our method improved BLEU by 0.5-1.5 and 0.1-0.2 for en-ja and ja-en translations, respectively.

In the future, we plan to i) assign better translation probabilities for non-OOV words that exist in the translation model learned from an out-of-domain parallel corpus, ii) extend our method to obtain a translation between phrases as in [30], and iii) combine our method with the existing approaches to domain adaptation for SMT that assumes no bilingual corpus in the target domain.

## Acknowledgments

The authors thank Nobuhiro Kaji and the anonymous reviewers for their valuable comments and suggestions. We also thank Jun Harashima for providing us the Cookpad recipe corpus. This work was partially supported by JSPS KAKENHI Grant Number 25280111.

## References


1. Brants, T., Popat, A.C., Xu, P., Och, F.J., Dean, J.: Large language models in machine translation. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). (2007) 858–867
2. Irvine, A., Morgan, J., Carpuat, M., III, H.D., Munteanu, D.: Measuring machine translation errors in new domains. Transactions of the Association for Computational Linguistics **1** (2013) 429–440
3. Costa-Jussà, M.R.: Domain adaptation strategies in statistical machine translation: a brief overview. The Knowledge Engineering Review **30** (2015) 514–520
4. Mansour, S., Ney, H.: Unsupervised adaptation for statistical machine translation. In: Proceedings of the Ninth Workshop on Statistical Machine Translation. (2014) 457–465
5. Ishiwatari, S., Kaji, N., Yoshinaga, N., Toyoda, M., Kitsuregawa, M.: Accurate cross-lingual projection between count-based word vectors by exploiting translatable context pairs. In: Proceedings of the 19th Conference on Computational Natural Language Learning (CoNLL). (2015) 300–304
6. Wu, H., Wang, H., Zong, C.: Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In: Proceedings of the 22nd International Conference on Computational Linguistics (COLING). (2008) 993–1000
7. Daume III, H., Jagarlamudi, J.: Domain adaptation for machine translation by mining unseen words. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT). (2011) 407–412
8. Irvine, A., Quirk, C., Daumé III, H.: Monolingual marginal matching for translation model adaptation. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP). (2013) 1077–1088
9. Razmara, M., Siahbani, M., Haffari, R., Sarkar, A.: Graph propagation for paraphrasing out-of-vocabulary words in statistical machine translation. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL). (2013) 1105–1115
10. Mathur, P., Keseler, F.B., Venkatapathy, S., Cancedda, N.: Fast domain adaptation of SMT models without in-domain parallel data. In: Proceedings of the 25th International Conference on Computational Linguistics (COLING). (2014) 1114–1123
11. Yamamoto, H., Sumita, E.: Bilingual cluster based models for statistical machine translation. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). (2007) 514–523
12. Harris, Z.S.: Distributional structure. Word **10** (1954) 146–162
13. Firth, J.R.: A synopsis of linguistic theory. Studies in Linguistic Analysis (1957) 1–32
14. Lund, K., Burgess, C.: Producing high-dimensional semantic spaces from lexical co-occurrence. Behavior Research Methods, Instruments, & Computers **28** (1996) 203–208
15. Bengio, Y., Ducharme, R., Vincent, P., Janvin, C.: A neural probabilistic language model. Journal of Machine Learning Research **3** (2003) 1137–1155
16. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Proceedings of Workshop at International Conference on Learning Representations (ICLR). (2013)
17. Turney, P.D., Pantel, P., et al.: From frequency to meaning: Vector space models of semantics. Journal of Artificial Intelligence Research (JAIR) **37** (2010) 141–188
18. Erk, K.: Vector space models of word meaning and phrase meaning: A survey. Language and Linguistics Compass **6** (2012) 635–653

19. Mikolov, T., Le, Q.V., Sutskever, I.: Exploiting similarities among languages for machine translation. arXiv preprint (2013)
20. Fung, P.: A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora. In: Proceedings of the Third Conference of the Association for Machine Translation in the Americas (AMTA). (1998) 1–17
21. Tsunakawa, T., Okazaki, N., Liu, X., Tsujii, J.: A Chinese-Japanese lexical machine translation through a pivot language. ACM Transactions on Asian Language Information Processing (TALIP) **8** (2009) 9:1–9:21
22. Neubig, G.: The Kyoto free translation task. http://www.phontron.com/kftt (2011)
23. Koehn, P., Knight, K.: Learning a translation lexicon from monolingual corpora. In: Proceedings of ACL Workshop on Unsupervised lexical acquisition. (2002) 9–16
24. Stolcke, A., et al.: SRILM-an extensible language modeling toolkit. In: Proceedings of the seventh International Conference on Spoken Language Processing (ICSLP). (2002) 901–904
25. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Computational linguistics **29** (2003) 19–51
26. Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. Computational linguistics **16** (1990) 22–29
27. Koehn, P.: Statistical significance tests for machine translation evaluation. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP). (2004) 388–395
28. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). (2002) 311–318
29. Lembersky, G., Ordan, N., Wintner, S.: Adapting translation models to translationese improves SMT. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL). (2012) 255–265
30. Schwenk, H.: Continuous space translation models for phrase-based statistical machine translation. In: Proceedings of 24th International Conference on Computational Linguistics (COLING): Posters. (2012) 1071–1080