Semi-automatic Filtering of Translation Errors in Triangle Corpus

Sung-Kwon Choi, Jong-Hun Shin and Young-Gil Kim
Natural Language Processing Research Section
Electronics and Telecommunications Research Institute, Daejeon, Korea
{choisk, jhsin82, kimyk}@etri.re.kr

Abstract

We are developing a multilingual machine translation system to provide foreign tourists with a multilingual speech translation service in the Winter Olympic Games that will be held in Korea in 2018. For a knowledge learning to make the multilingual expansibility possible, we needed large bilingual corpus. In Korea there were a lot of Korean-English bilingual corpus, but Korean-French bilingual corpus and Korean-Spanish bilingual corpus lacked absolutely. Korean-English-French and Korean-English-Spanish triangle corpus were constructed by crowdsourcing translation using the existing large Korean-English corpus. But we found a lot of translation errors from the triangle corpora. This paper aims at filtering of translation errors in large triangle corpus constructed by crowdsourcing translation to reduce the translation loss of triangle corpus with English as a pivot language. Experiment shows that our method improves +0.34 BLEU points over the baseline system.

1 Introduction

Triangle corpus is the corpus 'source language-pivot language-target language (hereafter, L1-Lp-L2)' where a source language (hereafter, L1) is translated into a pivot language (hereafter, Lp) and then the pivot language is translated into a target language (hereafter, L2). One of methods building large triangle corpus with English as a pivot language is a crowdsourcing translation. The crowdsourcing translation means a distributed model of translation that uses contributors instead of, or combined with, professional translators. In environment of the crowdsourcing translation it is possible to build a lot of bilingual corpora that are both time and cost-effective. In particular, if there is large bilingual corpus of L1-English, we can produce fast translation result of L1-L2 via the crowdsourcing translation of English-L2. Although there is such advantage of crowdsourcing translation, there is also its drawback that translation errors and inconsistency can arise because a large pool of people is going to generate input of differing quality. That is, a translation loss can be produced between L1-English and English-L2.

This paper aims at semi-automatic filtering of translation errors in large triangle corpus constructed by crowdsourcing translation to reduce the translation loss that can occur in crowdsourcing translation of corpus with English as a pivot language. The remainder of this paper is organized as follows. Section 2 presents the related work. In Section 3, we describe large English-French and English-Spanish bilingual corpus constructed by crowdsourcing translation using English as a target language of large Korean-English corpus. Translation errors in the Korean-English-French triangle corpus are manually analyzed by a human translator. In Section 4, we describe how to filter the translation errors caused from the crowdsourcing translation. Section 5 presents the experimental setup and the results.

2 Related Work

There were very little researches to improve the procedural translation loss of L1-English-L2 triangle corpus. Instead, there have been numerous researches in machine translation (hereafter, MT)

29th Pacific Asia Conference on Language, Information and Computation: Posters, pages 72 - 79
Shanghai, China, October 30 - November 1, 2015

using L1-English-L2 corpus as a training set. Such researches can be classified into three methods.

- Transfer Method: the transfer method (Utiyama and Isahara, 2007; Costa-jussà et al., 2011) connects a source-pivot MT system and a pivot-target MT system. The source-pivot MT system translates a source sentence into the pivot language, and the pivot-target MT system translates the pivot sentence into the target sentence. The problem with the transfer method is that the time cost is doubled and the translation error of the source-pivot translation system will be transferred to the pivot-target translation because it needs to decode twice.
- Synthetic Method: the synthetic method creates a synthetic source-target corpus by: (1) translate the pivot part in source-pivot corpus into target language with a pivot-target model; (2) translate the pivot part in pivot-target corpus into source language with a pivot-source model; (3) combine the source sentences with translated target sentences or/and combine the target sentences with translated source sentences (Wu and Wang, 2009). The problem with the synthetic method is that it is difficult to build a high quality translation system with a corpus created by a machine translation system.
- Triangulation Method: the triangulation method obtains source-target phrase table by merging source-pivot and pivot-target phrase table entries with identical pivot language phrases and multiplying corresponding posterior probabilities (Cohn and Lapata, 2007). According to an Arabic-Chinese experiment of Chen et al.(2008), BLEU(Papineni et. al. 2002) of statistical machine translation (hereafter, SMT) based on the triangulation method was better than that of SMT based on L1-L2. The problem of this approach is that the probability space of the source-target phrase pairs is non-uniformity due to the mismatching of the pivot phrase. To resolve this disadvantage, Zuh et al.(2014) proposed the approach to calculate the co-occurrence count of source-pivot and pivot-target phrase pairs.

Despite these three methods, there were still little researches in checking what kind of translation loss the L1-English-L2 triangle corpus has. Furthermore, there were little evaluation about corpus which was constructed by crowdsourcing translation. In this point, this paper aims at semi-automatic filtering of translation errors of large L1-English-L2 triangle corpus constructed by crowdsourcing translation to reduce the translation loss.

3 Human Analysis of Translation Errors in Crowdsourcing Translation

We are developing a multilingual MT system including Korean, English, Chinese, Japanese, French, Spanish, German, and Russian to provide foreign tourists with a multilingual speech translation service in the Winter Olympic Games that will be held in Korea in 2018. The multilingual MT system is characterized as follows:

- Controllability: makes high-quality translation possible through manual correction of knowledge errors by users and obtains the effect of the aforesaid customization.
- Common transfer: makes the addition of new languages easy because many languages share a format of transfer such as universal dependency annotation for multilingual parsing (McDonald et al., 2013)
- Knowledge learning: makes multilingual expansibility and/or domain customization possible because the translation knowledge is automatically learned from training data.

Our multilingual MT system considers in particular a multilingual expansibility as important. For a knowledge learning to make the multilingual expansibility possible, we needed large bilingual corpus. In Korea there were a lot of Korean-English (hereafter, K-E) bilingual corpus, but either Korean-French (hereafter, K-F) bilingual corpus or Korean-Spanish (hereafter, K-S) bilingual corpus lacked absolutely. It was very expensive to construct the K-F and K-S bilingual

corpus by professional translators. We had to think about constructing K-F and K-S corpus by crowdsourcing translation using the existing large K-E bilingual corpus. That is, English of K-E bilingual corpus became a source language and was translated into French and Spanish respectively. Crowdsourcing translation was conducted by Flitto in Korea, a global crowdsourcing translation platform like Amazon's Mechanical Turk (Callison-Burch and Dredze, 2010). K-F and K-S bilingual corpus constructed by crowdsourcing translation were as follows.

	# of sentences	Build-up period
K-E corpus	779,382	
E-F corpus	100,000	1 month
E-S corpus	200,000	1 month

Table 1: E-F and E-S corpus constructed by crowdsourcing translation using K-E corpus

200,000 of English sentences whose word length is in 3<#<23 became candidate sentences for K-F corpus and K-E corpus. Table 1 indicates that E-S corpus had 100,000 more sentences than E-F corpus because Flitto, crowdsourcing translation company held more English-Spanish translators than English-French translators.

To check translation quality in crowdsourcing translation, we extracted randomly 500 K-E-F sentences from 100,000 K-E-F sentences and conducted a human analysis of translation errors. The translation error analysis was based on the translation accuracy, which means conveying correctly the meaning of source sentence to the meaning of target sentence. K-E and E-F sentences were analyzed respectively. Types of translation errors include not only existing error types in machine translation (Fishel et al., 2012; Popovic et al., 2011) but also new error types such as ill-formed source sentence, ungrammatical generation and misunderstanding of situation. The result of analysis was as follows.

Types of translation errors	# of K-E sentences	# of E-F sentences	# of K-F sentences
Missing words- noun	2	3	5
Missing words - pronoun	0	2	2
Missing words - negation	0	1	1
Incorrect words - verb	1	46	47
Incorrect words - noun	6	29	32
Incorrect words – relative pronoun	0	1	1
Incorrect words - article	0	1	1
Incorrect words - adverb	0	5	5
Incorrect words - preposition	0	6	6
Incorrect words - auxilary verb	0	1	1
Incorrect words - adjective	0	1	1
ungrammatical generation - tense	0	5	5
ungrammatical generation - grammar	4	6	10
misunderstanding of situation	16	1	17
ill-formed source sentence	9	12	15
Total	38	120	149
500	7.6%	24.0%	29.8%

Table 2: Translation error analysis in 500 K-E-F sample sentences

In Table 2, the second column indicates the number of translation errors in K-E bilingual corpus constructed by professional translators and shows that 38 of 500 sentences have translation errors. The third column presents the number of translation errors in E-F sentences that were translated from English sentences of K-E bilingual corpus to French sentences by crowdsourcing and shows that 120 of 500 sentences have translation errors. The error analysis of the second and third column was separately conducted. In the fourth column it turns out that the K-F bilingual corpus as a combination between K-E translation and E-F translation has 149 sentences with translation

errors which run to 29.8% of 500 sentences. Through Table 2, we can know that the translation errors in L1-L2 corpus of L1-English-L2 corpus come from a combination of both the translation errors of L1-English and the translation errors of English-L2. The following examples show such cases.

Example 1: Error of K-F translation due to the error of K-E human translation
Korean source sentence: "배수의 진을 쳤다." ("I make a last-ditch fight.")
K-E Human translation : "I was between the devil and the deep blue sea."
E-F Crowdsourcing translation: "J'étais en plein dilemme." ("I was in a dilemma.")

Example 2: Error of K-F translation due to the error of E-F crowdsourcing translation
Korean source sentence: "아무 때라도 좋습니다." ("Anytime is okay.")
K-E Human translation : "Anytime."
E-F Crowdsourcing translation: "Je vous en prie." ("You're welcome")

Example 1 shows a K-F translation error due to the error 'incorrect words –noun' of K-E human translation. The Korean source sentence "배수의 진을 쳤다" that means "I make a last-ditch fight" was wrongly translated into the French sentence "J'étais en plein dilemme" that means "I was in a dilemma" because the Korean source sentence was wrongly translated into the English sentence "I was between the devil and the deep blue sea". Example 2 presents the error of K-F translation due to the error 'misunderstanding of situation' of E-F crowdsourcing translation. The Korean source sentence "아무 때라도 좋습니다" that means 'Anytime is okay" was wrongly translated into the French sentence "Je vous en prie" that means "You are welcome" because the English sentence "Any time" was wrongly translated into the French sentence "Je vous en prie" that means "You're welcome".

4 Assuming Distances in Triangle Corpus

In this section, we show a series of effort to find the sentence pairs including translation errors in crowdsourcing translation. Our goal is to find sentences which have content words that are semantically wrong. A general approach to realize this goal will be to use a bilingual dictionary. But it is difficult to build the bilingual dictionary. Besides, we need the part-of-speech tagger to align the words between source language and target language. To use a comparable corpus for under-resourced languages was also difficult. From this reason, we tried to measure the semantic distance by using L1-Lp-L2 without using a comparable corpus.

A vectorial text representation which is called a distributed word representation is a method to capture semantic and syntactic similarity of words in a monolingual sentence. (Bengio et al., 2003; Mikolov et al., 2013) Previous works on a distributed word representation have been concentrated on a monolingual corpus or have been approach to learn the linguistic regularities which are generalized across languages. (Klementiev et al., 2012; Lauly et al., 2014; Hermann and Blunsom, 2014a, 2014b) Such existing studies are based on the following idea: similar semantic and syntactic properties will be embedded nearby in the embedded vector space. We denote the representation result as a bilingual word embedding. Such representations have been used to achieve an excellent performance on word sense disambiguation, cross-lingual information retrieval, and word alignments. In this paper, we also use the characteristics of bilingual word embedding.

4.1 Motivations and System Structures to Find Translation Errors in Crowdsourcing Translation

When we construct the triangle corpus with English as a pivot, the following problems arise: 1) the translation errors appear due to missing words and grammatical errors, and 2) the meaning difference between L1-Lp sentences and Lp-L2 sentences affects the meaning difference between L1-L2 sentences. In case we implement a SMT system using such triangle corpus, the corpus including translation errors can cause the word alignment mismatching and have a bad influence on the translation quality of the SMT system. To resolve such problems, we tried to measure a sentence distance of L1-Lp-L2 and a sentence distance of L1-L2 respectively to find the semantic or syntactic similarity, since we thought that the similarity might be a clue of translation errors such

as semantic alternation, misprints and missing words. So, we used the bilingual distributed word representation.

Before measuring the sentence distance, the bilingual word embedding was constructed. Given the multilingual parallel corpus consisting of n language pairs including a specific source language, n(n+1)/2 of embedding should be produced. We conducted the word segmentation in Korean. In this paper we measured the distance between embeddings to extract the sentences L1-Lp-L2 that are beyond the threshold.

4.2 Calculating a Sentence Distance of L1-Lp-L2

The distributed word representation presents as a set of fixed-column real valued weights, and each weight can be assumed as a dimension. So we can handle a word of a sentence as a vector point in a hyperspace which can be calculated with a vector distance function.

Suppose we are given set of word pairs and their associated vector representation $\{x_i, y_i\}_{i=1}^{n}$, where $x_i \in \mathbb{R}^{d1}$ is the vector representation of word i in the source language, and $y_i \in \mathbb{R}^{d2}$ is the vector representation of word in target language. We calculate similarity for each word vector in a sentence, by the following n-dimensional cosine distance function:

$$d_1(x,y) = 1 - \cos\theta = 1 - \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \quad (1)$$

And Euclidean distance function considered as alternative to measure sentence distance:

$$\mathbf{d_2}(\mathbf{x},\mathbf{y}) = \sqrt{(\boldsymbol{x_1}-\boldsymbol{y_1})^2 + (\boldsymbol{x_2}-\boldsymbol{y_2})^2 + \cdots + (\boldsymbol{x_n}-\boldsymbol{y_n})^2} = \sqrt{\sum_{i=1}^{n}(\boldsymbol{x_i}-\boldsymbol{y_i})^2} \quad (2)$$

We applied cosine distance functions to set of words in a source-pivot sentence pair and a source-target sentence pair. By looking for a minimum distance to each of the words constituting the given sentence, it will be assist to find improper used vocabulary or absence of core keywords. So, a distance of each sentence is defined as equation (3):

$$\boldsymbol{SDist}(\boldsymbol{S^{d1}}, \boldsymbol{S^{d2}}) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} \boldsymbol{argmin}(\boldsymbol{d_1}(\boldsymbol{a_i},\boldsymbol{b_j}))}{\boldsymbol{n}} \quad (3)$$

where a_i is i-th word of a source sentence S^{d1}, and b_j is j-th word of a target sentence S^{d2}, relatively($a_i \in S^{d1}, b_j \in S^{d2}$). After calculating distance of L1-Lp and Lp-L2 sentence, we need to calculate a complete distance with following equation. given a source language sentence S^{dS}, a pivot language sentence S^{dP}, and a target language sentence S^{dT}, equation (4) is a final 'averaged' distance of $S^{dS} - S^{dT}$:

$$\mathit{A}\text{vgSentDist}(S^{dS}, S^{dP}, S^{dT}) = \sqrt{\left(\boldsymbol{SDist}(\boldsymbol{S^{dS}}, \boldsymbol{S^{dP}}) + \boldsymbol{SDist}(\boldsymbol{S^{dP}}, \boldsymbol{S^{dT}}) - \boldsymbol{SDist}(\boldsymbol{S^{dS}}, \boldsymbol{S^{dT}})\right)^2} \quad (4)$$

We wanted to find whether there is any correlation between the distance and the translation quality, even if we measure the distance of content words in L1-Lp-L2 through the above equation. It was because we had to establish a criterion about how long distance was wrongly translated to find the sentence pairs with translation errors. In our experiment, human translators decided heuristically whether the sentence pairs have a similar meaning in the statistical distribution of a calculated distance.

5 Experimental Result

5.1 Data and parameters

To verify the performance of the proposed methods, we used Korean-English-French corpus consisting of 100,000 parallel sentences. We tokenized and lowercased the English and French sentences, using some useful corpus preprocessing scripts in cdec-decoder. (Dyer et al., 2010) And for Korean we used in-house Korean morphological analyzer to get word tokens instead of using a monotonic whitespace tokenizer. To learn the bilingual word embedding, we used BICVM (Hermann and Blunsom, 2014a). Models were trained for up to 50 iterations. We set a dimensionality of word embedding size to D=128 as a default parameter and set the number of noise elements to 200. The adaptive gradient method (Duchi et al., 2011) was used to update weights of the models.

5.2 Filtering Experiment by Sentence Distance

We now present the calculated distance results by using our methodology. Test sentences were 300, which are in low order of calculated distances. The error analysis was as follows:

Incorrect Translations	Correct Translations	Total
114	186	300

Table 3: Error analysis of 300 sample sentences

If a Korean sentence was ambiguous, but a French sentence was correctly translated from its English sentence, we considered the Korean-French sentence pair as a correctly translated sentence pair. We analyzed the sentences that were incorrectly translated. They were 114 sentences which consisted of error types such as missing target word, irrelevant translation, incomplete sentences, and meaning change. Detailed error types were as follows:

Missing target word	Irrelevant translation	Incomplete sentences	Meaning changes	Tot-al
14	8	16	76	114

Table 4: Error types of incorrect translation

Most errors of "missing target word" were error type "missing noun word" (11 of 14 sentences, 78%). The meaning changes due to the literal translation occurred in French sentences with narrowish meaning via the ambiguous predicates in English sentences (35 of 76 sentences, 46%). The examples of sentences with incorrect translation are shown in below table:

1	KO	습관성 턱 관절 탈골이예요.
	EN	He is tendency temporomandibular dislocation.
	FR	Je ne comprends pas cette phrase, désolé.
2	KO	그 은행이 계좌를 개설하면 고작 금반지를 나눠준대.
	EN	The bank only gives away foil when you open an account.
	FR	La banque ne donne que.
3	KO	정리를 해 주세요.
	EN	Please take care of it.
	FR	S'il vous plaît occupez - vous en.
4	KO	저는 개를 좋아합니다.
	EN	I am a dog person.
	FR	J'aime les chiens.
5	KO	더 보고 싶으신 건 없나요?
	EN	Is there anything else you want?
	FR	Avec ceci?

Table 5: Examples of Translation Errors

In the case of first sentence example, the French sentence "Je ne comprends pas cette phrase, désolé" means "I cannot understand that phrase, I'm sorry…". We guess that a crowdsourcing participant translated the French sentence so because he/she did not understand the meaning of a medical term 'temporomandibular dislocation'. In the second example, French sentence that means "the bank only gives away" was not completed unlike Korean and English sentence. In the third example, Korean sentence means "Please clean up" or "Please arrange it". But it was incorrectly translated into "Take care of it" in English and "S'il vous plaît occupez - vous en" in French that means "Please take care of you". And the fourth Korean sentence was correctly translated into both English sentence and French sentence in the point of view of common speech (or slang). The last example is considered as a bad translation because the French sentence means "with this?" literally, even if it has same meaning as "is there anything else?" in French cultural area. Like this, translated sentences are dependent on cultural differences and slang/common speeches.

5.3 Verifying Experiment of Sentence Distance using Phrase-based SMT

To compare a performance of a filtered Korean-English-French corpus with a performance of an original Korean-English-French corpus, we trained a phrase-based SMT (Koehn et al., 2007). 90,000 sentences were a training set and the remaining 10,000 sentences were an evaluation set in order to train a SMT model. To make a filtered corpus, we removed the farthest distance of 1,000 sentences from the calculated sentence distance list, which would be assumed the incorrectly translated sentences. The sentences removed from training set were 919 sentences. So, sentences to train a filtered SMT model became 89,081. 87 sentences

were removed from the evaluation set, so we used 9,913 sentences for a performance evaluation. The evaluation metric of SMT model was BLEU (Papineni et al., 2002). Along with this evaluation set, we conducted an additional automatic evaluation using in-house Korean-French corpus which contains 3,000 parallel sentences with 1 reference. This evaluation set has same tourist/dialog domains as crowdsourcing translation corpus. Total number of Korean words were 12,284 and a sentence consisted of 4 words in average, while total number of French words were 20,346 and a sentence consisted of 6 words in average. The evaluation results are illustrated with below table:

	BLEU (Original)	BLEU (Filtered)
10k samples(pivot)	8.45	8.44
9.9k samples(pivot)	8.46	8.47
3k evalset(pivot)	14.13	14.47

Table 6: Original (=Non-filtered) / Filtered BLEU evaluation score result. 10k samples and 9.9k samples denote an evaluation corpus size, which is non-filtered original and filtered evaluation set respectively. And 3k evalset denotes our in-house Korean-French BLEU evaluation set.

In table 6, the 'pivot' denotes the transfer method (Wu and Wang, 2007), that is, Korean-English SMT results were used to get the translation results of the English-French SMT system. Despite of the simplicity of proposed method, the amount of the total training corpus was decreased, but we could see a slight performance improvement. From the above results, we could discover that removing the sentences which have a weak semantic similarity is helpful for improving translation corpus quality.

6 Conclusion

The crowdsourcing translation is an excellent method to reduce the translation cost and the translation period to construct large bilingual corpus. In case the corpus by the crowdsourcing translation is very large, the assessment of translation quality about the corpus should depend on the random sampling. Such random sampling could not resolve the translation loss caused by crowdsourcing translation.

This paper aimed at no random sampling, but the total crowdsourcing translation to be examined. Through word distance and sentence distance, we could extract high-quality translations of L1-L2 without translation loss from total crowdsourcing translation of L1-Lp-L2. Furthermore, our approach has the advantage to make efficient management of high quality multilingual corpus possible because it can reduce a translation loss due to triangulation translation and intensify L1-Lp-L2 due to a combination among languages.

Acknowledgments

This work was supported by the ICT R&D program of MSIP/IITP. [10041807, Development of Original Software Technology for Automatic Speech Translation with Performance 90% for Tour/International Event focused on Multilingual Expansibility and based on Knowledge Learning]

References

Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003) A neural probabilistic language model. The Journal of Machine Learning Research, 3, 1137-1155.

Callison-Burch, C. and Dredze, M. (2010) Creating Speech and Language Data With Amazon's Mechanical Turk. In Proceedings NAACL.

Chen, Y., Eisele, A., and Kay, M. (2008) Improving Statistical Machine Translation Efficiency by Triangulation. In Proceedings of the Sixth International Language Resources and Evaluation, 2875-2880.

Cohn, T. and Lapata, M. (2007). Machine Translation by Triangulation: Make Effective Use of Multi-Parallel Corpora. In Proceedings of 45th Annual Meeting of the Association for Computational Linguistics, 828-735.

Costa-jussà, M.R., Henríquez, C., and Banchs, R.E. (2011). Enhancing Scarce-Resource Language Translation through Pivot Combinations. In Proceedings of the 5th International Joint Conference on Natural Language Processing, 1361-1365.

Duchi, J., Hazan, E., and Singer, Y. (2011) Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research, 12, 2121-2159.

Dyer, C., Lopez, A., Ganitkevitch, J., Weese, J., Ture, F., Blunsom, P., Setiawan, H., Eidelman, V., and Resnik, P. (2010) A decoder, alignment, and learning

framework for finite-state and context-free translation models. In Proceedings of ACL System Demonstrations, 7-12.

Fishel, M., Bojar, O., and Popovi´c, M. (2012) Terra: a Collection of Translation Error-Annotated Corpora. In Proceedings of Language Resources and Evaluation Conference, 7-14

Hermann, K. M., and Blunsom, P. (2014a) The Role of Syntax in Vector Space Models of Compositional Semantics. In Association for Computational Linguistics, 894-904.

Hermann, K. M., and Blunsom, P. (2014b) Multilingual models for compositional distributed semantics. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 58 – 68.

Klementiev, A., Titov, I., and Bhattarai, B. (2012). Inducing crosslingual distributed representations of words. In Proceedings of COLING, 1459-1474.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., and Herbst, E. (2007) Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions, 177-180.

Lauly, S., Boulanger, A., and Larochelle, H. (2014) Learning multilingual word representations using a bag-of-words autoencoder. arXiv preprint arXiv:1401.1803.

McDonald, R., Nivre, J., Quirmbach-Brundage, Y., and Goldberg, Y. (2013) Universal Dependency Annotation for Multilingual Parsing. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, 92-97.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013) Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of NIPS.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. (2002) BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics, 311-318.

Popovi´c, M. and Ney, N. (2011) Towards Automatic Error Analysis of Machine Translation Output. Computational Linguistics. Vol.37, Number 4, 657-688.

Utiyama, M. and Isahara, H. (2007). A Comparison of Pivot Methods for Phrase-Based Statistical Machine Translation. In Proceedings of Human Language Technology: the Conference of the North American Chapter of the Association for Computational Linguistics, 484-491.

Wu, H., & Wang, H. (2007). Pivot language approach for phrase-based statistical machine translation. Machine Translation, 21(3), 165-181.

Wu, H. and Wang, H. (2009). Revisiting Pivot Language Approach for Machine Translation. In Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th IJCNLP of the AFNLP, 154-162.

Zuh, X., He, Z., Wu, H., Zhu, C., Wang, H., and Zhao, T. (2014) Improving Pivot-based Statistical Machine Translation by Pivoting the Co-occurrence Count of Phrase Pairs. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1665–1675.