2016 IEEE 28th International Conference on Tools with Artificial Intelligence

Applying Semantic Similarity to Phrase Pivot Translation

Hai-Long Trieu, Le-Minh Nguyen
Japan Advanced Institute of Science and Technology
Nomi City, Ishikawa, Japan
Email: {trieulh, nguyenml}@jaist.ac.jp

***Abstract*—Pivot methods have shown to be an effective solution to overcome the problem of unavailable large bilingual corpora in statistical machine translation. The representative approach of pivot methods is the phrase pivot translation which is based on common pivot phrases to produce connections between source-pivot and pivot-target phrase tables. Nevertheless, this approach produces insufficient connections behind the phrase tables because pivot phrases still contain the same meaning even when they are not matched to each other. In this work, we propose applying semantic similarity between pivot phrases to phrase pivot translation. In order to extract similar pivot phrases, we used string similarity measures for phrase similarity, and WordNet and Word2Vec were used for word similarity. The experiments show that using semantic similarity is able to extract more informative phrases, which can support for phrase pivot translation.**

I. Introduction

Statistical machine translation (SMT) systems can achieve high translation quality when we provide large bilingual corpora [5]. Nonetheless, such corpora are unavailable for almost language pairs, which leads to a bottleneck for SMT. One solution to overcome this problem is pivot methods which use intermediate languages (pivot languages) to bridge a source language to a target language when bilingual corpora of the pivot languages paired with the source and the target languages are available ([1], [3], [4], [7], [18], [19]).

In various approaches of pivot methods, triangulation is the representative approach which multiplies scores of phrase tables via common pivot phrases ([3], [18], [19]). Nevertheless, one weakness of the conventional triangulation approach is that it does not generate sufficient source-target phrase pairs. For instance, a source-target phrase pair should be generated, but it may be lost when each of these phrases is not connected to a common pivot phrase.

In this work, we propose using semantic similarity between pivot phrases to extract more sufficient information for phrase pivot translation. This is based on the fact that phrases may still contain the same meaning even when they are not matched to each other. We divided pivot phrases into two sets based on the length of phrases. For phrases that contain more than one word, we used string similarity measures to score the similarity between pivot phrases P_s and P_t of the two phrase table: $S - P_s$ and $P_t - T$. For pivot phrases that contain only one word, we applied two well-known techniques of word similarity: WordNet [14] and Word2Vec [13]. We conducted evaluations on some language pairs: Czech-Russian, German-French, and Japanese-Vietnamese.

II. Related Work

There are three main approaches in pivot methods: cascade, synthetic, and phrase pivot translation (or triangulation). The cascade approach translates source sentences to pivot then to target ([4], [18]). The second approach, synthetic [4], uses source-pivot or pivot-target translation model to produce a synthetic source-target corpus. For example, the pivot side of pivot-target corpus is translated using source-pivot translation model. The third approach, phrase pivot translation ([3], [18], [19]), estimates translation probabilities of the source-target phrase table by multiplying scores of the source-pivot and pivot-target phrase tables via common pivot phrases. This approach has shown to be the most effective in pivot methods ([6], [18]).

In order to overcome the problem of lost information in triangulation method, Zhu et al., 2013 [20] used random walk. In our work, we propose using semantic similarity of pivot phrases to extract more sufficient information for the triangulation.

III. Method

A. Triangulation

Given a source phrase s and target phrase t of the source-pivot phrase table $S - P_s$ and the pivot-target phrase table $P_t - T$, the phrase translation probability is estimated via common pivot phrases p based on the following feature function.

$$\phi(t|s) = \sum_{p \in (S-P_s) \cap (P_t - T)} \phi(p|s)\phi(t|p) \qquad (1)$$

The conventional triangulation approaches ([3], [18], [19]) are based on the common pivot phrases to produce connections between s and t. Nevertheless, using matched pivot phrases only obtains a part of potential connections between source and target phrases because some pivot phrases p_s and p_t may contain the same meaning even when they are not matched to each other. We propose using semantic similarity to overcome this problem, which is shown in the following.

2375-0197/16 $31.00 © 2016 IEEE
DOI 10.1109/ICTAI.2016.157

B. Bridge via Similar Pivot Phrases

The formula of translation probability via similar pivot phrases is shown in Equation 2.

$$\phi(t|s) = \sum_{p_s \in P_s, p_t in P_t} \phi(p_s|s)\phi(t|p_t)\Theta(p_s, p_t) \qquad (2)$$

where $0 \leq \Theta(x, y) \leq 1$ denotes the similarity between the two phrases x and y.

The procedure of the method can be described in more detail in Algorithm 1.

Algorithm 1: Phrase pivot translation through similar pivot phrases

Input : phrase tables: $S - P_s, P_t - T$
Output: phrase table: $\bar{S} - \bar{T}$ through similar pivot phrases $(\bar{P}_s, \bar{P}_t)$ where $\bar{P}_s \in P_s, \bar{P}_t \in P_t, \bar{P}_s \neq \bar{P}_t$

begin
  # Extract similarity between pivot phrases
  $P' = P_s \cap P_t$ // common pivot phrases
  $\bar{S} = S \setminus S'$ where phrase pairs $(S' - P') \in S - P_s$
  Extract $\bar{P}_s$ where phrase pairs $(\bar{S} - \bar{P}_s) \in S - P_s$ and $\bar{P}_s \in P_s \setminus P'$
  for $\bar{p_s}$ *in* $\bar{P}_s$ **do**
    for $\bar{p_t}$ *in* P_t **do**
      compute: $\Theta(\bar{p_s}, \bar{p_t})$
  # Pivot through similar phrases
  for $\bar{s} \in \bar{S}$ *and* $(\bar{s} - \bar{p_s}) \in (\bar{S} - \bar{P}_s)$ **do**
    for $t \in T$ *and* $(\bar{p_t} - t) \in (P_t - T)$ **do**
      if *(* $\Theta(\bar{p_s}, \bar{p_t}) > threshold$*)* **then**
        $\phi(\bar{s}, t) =$
        $\phi(\bar{s}, t) + \phi(\bar{s}, p_s) * \phi(p_t, t) * \Theta(\bar{p_s}, \bar{p_t})$

Pivoting through similar pivot phrases can be seen as a graph (Figure 1). Phrases are denoted by notes, and arcs represent the relation between phrases. As shown in Figure 1, there are two sides: source-pivot and pivot-target representing phrases in the two phrase tables. The black nodes denote matched pivot phrases, whereas the arrows mean similarity between pivot phrases.

In order to compute the similarity between pivot phrases, $\Theta(x, y)$, phrases were divided into two sets: phrases containing only one word, and phrases containing more than one words. String similarity measures were then used to compute similarity between phrases containing more than one words. For phrases containing only one word, we used two techniques: WordNet and Word2Vec. This is because string similarity measures may not strong for word similarity, so we decided to use powerful techniques for word similarity

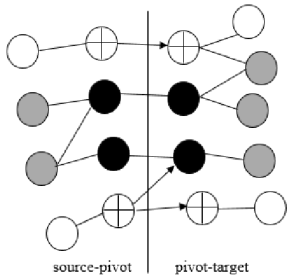

Figure 1. Semantic similarity for triangulation as a graph.

like WordNet or Word2Vec. The following presents these strategies.

1) String Similarity Measures:

Cosine Similarity: Given two string s_1 and s_2, the similarity between these two strings can be computed using cosine similarity which is the cosine of the angle between these two vectors representation of s_1 and s_2.

$$cosine(s_1, s_2) = \frac{v_1 * v_2}{|v_1| * |v_2|} \qquad (3)$$

where v_1 and v_2 denote the two vectors representing the two string s_1 and s_2, respectively.

Levenshtein Distance: The similarity of two strings s_1 and s_2 can also be computed using Levenshtein distance.

$$lev_{s_1,s_2}(|s_1|, |s_2|) = \begin{cases} max(i, j), & \text{(if min(i,j)=0)} \\ min \begin{cases} lev_{s_1,s_2}(i-1, j) + 1 \\ lev_{s_1,s_2}(i, j-1) + 1, \\ lev_{s_1,s_2}(i-1, j-1) + 1_{(s_{1_i} \neq s_{2_j})} \end{cases} & \text{(otherwise)} \end{cases}$$

where $|s_1|$ and $|s_2|$ denotes the length of s_1 and s_2, respectively. The function $1_{(s_{1_i} \neq s_{2_j})}$ is the indicator function equal to 1 when $s_{1_i} \neq s_{2_j}$, and equal to 0 otherwise. $lev_{s_1,s_2}(i, j)$ is the distance between the first i characters of s_1 and the first j characters of s_2.

Longest Common String: The string similarity measure of strings s_1 and s_2 based on longest common string is defined as

$$d(s_1, s_2) = 1 - \frac{f(s_1, s_2)}{M(s_1, s_2)} \qquad (4)$$

where $f(s_1, s_2)$ is the length of the longest common subsequence(s) of s_1 and s_2, and $M(s_1, s_2)$ is the length of the longest string of s_1 and s_2.

2) Word Similarity: For similarity between pivot phrases that contain only one word, the task becomes word similarity. We applied two powerful techniques: WordNet and Word2Vec. WordNet is a valuable linguistic resource built by annotators that contains relationship between words including synonym, sets of words that share the same meaning. Nevertheless, this resource is built just for some languages

so far. Fortunately, the resource is available for English that is usually used as the pivot language to bridge language pairs in pivot translation. In the case that WordNet is unavailable for the pivot language, we propose using another technique: Word2Vec, an unsupervised learning algorithm on monolingual corpora.

WordNet: WordNet[14] provides synonym in a term namely synset. We take advantage of this resource for similarity between words. The similarity between words e_1 and e_2 is computed as follows.

$$strength(e_1, e_2) = [\{e|e \in synset(e_1) \cap synset(e_2)\}] \quad (5)$$

The function $strength(e_1, e_2)$ denotes the number of common words between the two synset of e_1 and e_2. We computed the values of $strength(e_i, e_j)$; we then normalized using maximum likelihood to obtain the similarity between e_1 and e_2 whose values belong to [0..1].

Word2Vec: We applied Word2Vec [13], a powerful tool of distributed vector representation, to compute word similarity. There are two techniques of Word2Vec including skip-gram and continuous bag of word (CBOW), which use context to predict a word or phrase. We adapted the CBOW model for word similarity.

The similarity between words e_1 and e_2 is computed as:

$$\Theta(e_1, e_2) = cosine(v_1, v_2) \quad (6)$$

where v_1 and v_2 are vector representations of e_1 and e_2 trained by the Word2Vec model.

IV. Experiments

We evaluated the proposed method on some language pairs. We chose Japanese-Vietnamese, a low-resource Asian language pair which contains different linguistic structures. Another language pair is Czech-Russian, a less common language pair in SMT for European. We also conducted experiments on more common languages: translation from German to French.

A. Setup

We used English as the pivot languages for all experiments. For Japanese-Vietnamese, we used the multilingual Bibble corpus. The Japanese-English and English-Vietnamese corpora include 29K for each corpus. For the Czech-Russian experiment, we used two bilingual corpora: Czech-English included 160K sentence pairs from Common Crawl Corpus[1] and English-Russian from the corpus News Commentary 8[1] which includes 146K sentence pairs. For German-French translation, we used a small part of the Europarl corpus [10] for German-English which includes 400K sentence pairs. We used 400K English-French sentence pairs extracted from the shared task WMT14[2]. We used 200K German-French corpus extracted from the multilingual JRC-Acquis corpus [17] for direct translation. These corpora are described in Tables I-III.

Table I
CZECH-RUSSIAN: DATASETS

Setup	Training	Tuning	Testing
Czech-English	145K	2K	–
English-Russian	150K	3K	–
Czech-Russian	93K	2K	1K

Table II
GERMAN-FRENCH: DATASETS

Setup	Training	Tuning	Testing
German-English	400K	2K	–
English-French	400K	2K	–
German-French	200K	2K	3K

Table III
JAPANESE-VIETNAMESE: DATASETS

Setup	Training	Tuning	Testing
Japanese-English	29K	400	–
English-Vietnamese	29K	400	–
Japanese-Vietnamese	29K	400	500

All experiments are conducted based on the Moses toolkit [11] with default configurations. Word alignment is trained based on GIZA++ [15]. We tuned model using Batch Mira [2]. We used Ken LM [8] to trained 5-gram language model on the target languages for both language pairs. We used BLEU[16] for evaluation metric.

For the triangulation model, we implemented the method in [19] using java.

B. Results

1) Phrase Similarity: For training Word2Vec, we used the model CBOW with configurations size=300 and window=5. The English side of the parallel corpora were combined with the English monolingual corpora from WMT-2015 to train the model.

Similarity between pivot phrases are described in Table IV, Table V, and Table VI. These tables show that using strategies like string similarity measures or WordNet can extract informative phrases which the conventional phrase pivot translation cannot capture based on common pivot phrases. The informative characteristics include morphology (*counseller* and *counsellers*) and synonym phrases that contain few different words or punctuation but almost the same meaning (*to the solemn feasts* and *the solemn feast .*) or (*his kingdom ;* and *! his kingdom*). In using WordNet, the result also produces interesting similar word pairs: (*10* and *ten*), (*maybe* and *possibly*). Using Word2Vec also obtains useful

[1]http://www.statmt.org/wmt16/translation-task.html

[2]http://www.statmt.org/wmt14/translation-task.html

Table IV
AN EXAMPLE OF SIMILAR PHRASES USING COSINE SIMILARITY

$\bar{p_s}$	p_t	$similarity(\bar{p_s}, p_t)$
's counseller	counsellers	0.804030
to the solemn feasts	the solemn feast .	0.824958
his kingdom ;	! his kingdom	0.818182

Table V
AN EXAMPLE OF WORD SIMILARITY USING SYNONYM OF WORDNET

$\bar{p_s}$	p_t	$similarity(\bar{p_s}, p_t)$
10	ten	0.11111
10	X	0.11111
maybe	perhaps	0.25
maybe	possibly	0.25

word pairs like (*stream* and *flow*), (*New_York* and *NY*). These informative phrases can support for the conventional phrase pivot translation which is based solely on common pivot phrases.

2) Results on Czech-Russian Translation: We evaluated our method using the test set of UMC-01 corpus [9]. We used the existed training data of UMC-01 as the direct model. We applied back-off interpolation. The experimental result is described in Table VII. The setup **Single** means that we used only one phrase table in decoding: *direct* (training translation model using the Czech-Russian corpus), or *triangulation*. **Backoff direct+pivot** means that we used *direct* as the first phrase table, and when phrases were not contained in the first phrase table, they can be looked up in the pivot phrase table (*triangulation*). **Backoff +similarity** means that we combined the *direct* or the *triangulation* phrase table with the phrase table extracted by similar pivot phrases using our method.

From the results presented in Table VII, using only pivot phrase table in decoding achieves a low performance: 4.13 BLEU score. The vocabulary in the pivot phrase table may cover just a small ratio the vocabulary of the test set. Meanwhile, the direct phrase table can obtain the highest result: 10.71 BLEU score. This is consistent with results shown in previous pivot researches, whereas direct translation usually obtains better results, and pivot translation is used to improve the direct translation. In using similarity of pivot phrases, it can slightly improve the direct translation: 10.71 to 10.78 BLEU scores. Although the improvement is still small, this can imply that using similarity between pivot phrases can contribute to pivot translation. This should be further investigated.

Table VI
AN EXAMPLE OF WORD SIMILARITY USING WORD2VEC

$\bar{p_s}$	p_t	$similarity(\bar{p_s}, p_t)$
stream	streams	0.73877
stream	flow	0.55566
stream	torrent	0.48497
New_York	NY	0.85850

Table VII
EXPERIMENTAL RESULTS ON LANGUAGE PAIRS: CZECH-RUSSIAN (CS-RU), GERMAN-FRENCH (DE-FR), JAPANESE-VIETNAMESE (JA-VI)

Setup	**Single**	**Backoff direct+pivot**	**Backoff pivot+direct**	**Backoff +similarity**
Direct (CS-RU)	10.71	–	–	10.78
Pivot (CS-RU)	4.13	10.73	8.41	4.12
Direct (DE-FR)	8.25	–	–	8.28
Pivot (DE-FR)	3.39	8.70	7.37	3.41
Direct (JA-VI)	12.26	–	–	12.28
Pivot (JA-VI)	8.80	12.25	8.83	8.85

3) Results on German-French Translation: We used the the *newstest2013* of the WMT2013[3] to evaluate the performance of the method. As experimental results described in Table VII for German-French translation, the direct translation achieves the highest BLEU score: 8.25. Meanwhile, using pivot translation can improve the direct translation: 8.25 to 8.70 BLEU scores. Using similarity can improve the pivot translation: 3.39 to 3.41 BLEU score. Since the improvement is quite small, it is necessary to conduct more experiments and analyses.

4) Results on Japanese-Vietnamese Translation: We conducted experiments on Japanese-Vietnamese translation. This is a challenge task in SMT when Japanese-Vietnamese is not only a low-resource language pair but also different structure in linguistics. One of the differences between Japanese and Vietnamese is the sentence order: subject-verb-object vs subject-object-verb. Japanese also tends to dropout pronouns in sentences.

We divided the Bibble corpus[4] into tuples; each tuple contains three sentences: Japanese, English, and Vietnamese. We used 29K tuples for training, 400 tuples for tuning. For testing, we used Japanese-Vietnamese sentence pairs of 500 tuples for testing.

From the results shown in Table VII for Japanese-Vietnamese translation, the direct translation is higher than the pivot translation: 12.26 vs 8.80. The interesting observation drawn from the result is pivot phrase table cannot improve the direct translation when we combine direct and pivot phrase tables in decoding. The pivot translation obtains the lower performance than the direct translation, this may be because the pivot translation cannot cover a high vocabulary ratio of the test set. Therefore, we tried to use direct translation to improve the pivot translation in order that we can observe how the pivot translation can be improved. Unfortunately, the improvement is very small: 8.80 to 8.83 BLEU score. This implies that even when we use the direct translation, the highest performance in the experiments, the pivot translation still cannot be improved much. As a result, the similarity of pivot phrases slightly improve the pivot

[3]http://statmt.org/wmt13/
[4]http://homepages.inf.ed.ac.uk/s0787820/bible/

translation: 8.80 to 8.85 BLEU score. There is another challenge in pivot translation that can be drawn from this experiment, which is not only the vocabulary coverage but also the problem of selecting translation candidates and reordering related to the languages used in pivot methods.

V. Conclusion

This work presents applying semantic similarity between pivot phrases in phrase pivot translation. Conventional phrase pivot translation is based solely on common pivot phrases, which still lacks informative source-target phrase pairs. We used some strategies for similarity between pivot phrases: string similarity measures for phrases containing more than one word; WordNet and Word2Vec for word similarity. Experiments show that using these methods can extract more informative phrases for pivot translation, which may contribute to improve pivot translation. However, it is needed to conduct more investigations to show the contribution of this method more clearly, and we plan to perform experiments and analyses on this method in future researches.

Acknowledgment

This work was supported by JSPS KAKENHI Grant number 3050941 and JAIST's research grant for DRF.

References

[1] M. Cettolo, N. Bertoldi, M. Federico, and F.-F. B. Kessler, "Bootstrapping arabic-italian smt through comparable texts and pivot translation," in *Proc. of EAMT*, 2011, pp. 249–256.

[2] C. Cherry and G. Foster, "Batch tuning strategies for statistical machine translation," in *Proc. of HLT/NAACL*. Association for Computational Linguistics, 2012, pp. 427–436.

[3] T. Cohn and M. Lapata, "Machine translation by triangulation: Making effective use of multi-parallel corpora," in *Proc. of ACL*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 728–735. [Online]. Available: http://www.aclweb.org/anthology/P/P07/P07-1092

[4] A. De Gispert and J. B. Marino, "Catalan-english statistical machine translation without parallel corpus: bridging through spanish," in *Proc. of LREC*. Citeseer, 2006, pp. 65–68.

[5] C. Dyer, A. Cordova, A. Mont, and J. Lin, "Fast, easy, and cheap: Construction of statistical machine translation models with mapreduce," in *Proc. of the Third Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 2008, pp. 199–207.

[6] A. El Kholy, N. Habash, G. Leusch, E. Matusov, and H. Sawaf, "Language independent connectivity strength features for phrase pivot statistical machine translation," in *Proc. of ACL*, 2013, pp. 412–418.

[7] N. Habash and J. Hu, "Improving arabic-chinese statistical machine translation using english as pivot language," in *Proc. of the Fourth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 2009, pp. 173–181.

[8] K. Heafield, "Kenlm: Faster and smaller language model queries," in *Proc. of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 2011, pp. 187–197.

[9] N. Klyueva and O. Bojar, "Umc 0.1: Czech-russian-english multilingual corpus," in *Proc. of CILC*, 2008, pp. 188–195.

[10] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *MT summit*, vol. 5, 2005, pp. 79–86.

[11] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens *et al.*, "Moses: Open source toolkit for statistical machine translation," in *Proc. of ACL*. Association for Computational Linguistics, 2007, pp. 177–180.

[12] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," in *Proc. of HLT/NAACL*. Association for Computational Linguistics, 2003, pp. 48–54.

[13] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[14] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[15] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[16] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proc. of ACL*. Association for Computational Linguistics, 2002, pp. 311–318.

[17] R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufis, and D. Varga, "The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages," *arXiv preprint cs/0609058*, 2006.

[18] M. Utiyama and H. Isahara, "A comparison of pivot methods for phrase-based statistical machine translation," in *Proc. of HLT/NAACL*. Rochester, New York: Association for Computational Linguistics, April 2007, pp. 484–491. [Online]. Available: http://www.aclweb.org/anthology/N/N07/N07-1061

[19] H. Wu and H. Wang, "Pivot language approach for phrase-based statistical machine translation," in *Proc. of ACL*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 856–863. [Online]. Available: http://www.aclweb.org/anthology/P/P07/P07-1108

[20] X. Zhu, Z. He, H. Wu, H. Wang, C. Zhu, and T. Zhao, "Improving pivot-based statistical machine translation using random walk," in *Proc. of EMNLP*. Seattle, Washington, USA: Association for Computational Linguistics, October 2013, pp. 524–534. [Online]. Available: http://www.aclweb.org/anthology/D13-1050