International Journal of Engineering & Technology, 7 (2.30) (2018) 15-20

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Review of techniques for extraction of bilingual lexicon from comparable corpora

Manpreet Singh Lehal [1] *, Dr. Ajit Kumar [2], Dr. Vishal Goyal [3]

[1] PhD Student, Punjabi University, Patiala, Punjab India
[2] Assistant Professor, Multani Mal Modi College, Patiala, Punjab India.
[3] Associate Professor, DCS, Punjabi University, Patiala, Punjab India.
**Corresponding author E-mail: mslehal@lkc.ac.in*

Abstract

Bilingual lexicons are important resources for performing a number of bilingual tasks in machine translation (MT) and cross-language information retrieval (CLIR). Since the manual building of bilingual extraction is a tedious affair, researchers have focused upon the automatic extraction of bilingual lexicons from corpora. Another issue is the use of parallel and comparable corpora for extraction. Much success has been achieved in the use of parallel corpora but it is only available for a few language pairs and for limited domains. Therefore, the use of comparable corpora comes as an alternative but a lot need to be done in this field. The paper presents a review of different techniques and methods, which have been used for automatic extraction of bilingual lexicon suggesting that an integrated approach can give better results than using individual approaches. The paper also contains a proposed method for extraction of bilingual method using a combined approach.

Keywords: *Bilingual Lexicon; Comparable Corpora; Machine Translation; Extraction.*

1. Introduction

A bilingual lexicon is constituted of words paired together which are probably translations of one another. [16]. It can be called an extended dictionary which can be deciphered by a computer. Manning and Schutze defined bilingual lexicon as an expandable machine-readable bilingual dictionary consisting of bilingual word pairs. The probable word translations are paired on the basis of similarity scores.

The automatic extraction of bilingual lexicons can be done effectively using statistical methods from a corpora of two different languages in which the content is translation of one another or better called a parallel corpora.[2] [19]. Statistical translation models make use of the correlations and lexical information of two languages. The models usually work upon word-level and use word alignment to estimate the translation probabilities. However this approach requires alignment of sentences as a prerequisite before proceeding towards word-level alignment. It is dependent upon the number of words and efficiency level of the parallel corpora.

Bilingual word lexicons are utilized in a number of NLP tasks, like finding information across languages [6]. The bilingual lexicon is either built manually or automatically extracted from a parallel corpora. Most of the methods of Bilingual lexicon extraction require a seed lexicon (which can again be extracted manually or automatically). But these methods suffer from a great deal of noise. These methods work on assumption that a word in a particular context will have its translation in similar context in the other language. [26]. The researchers focused on the similarity of context and calculated the similarity level. The word pairs with high rate of similarity are selected as probable candidate pairs. These methods are called context-similarity based methods. Seed bilingual lexicons help in calculating the contextual similarities. They map together the contexts of separate languages into one common place. If the word pairs do not occur in seed dictionary they are not selected. As such the context based methods require large size lexicons otherwise there would be little accuracy. As a solution, different methods were attempted which either make no use of seed lexicon [26] or make it size independent [19].

The commonly used methods to measure similarity scores make use of three steps: first of all words are converted to vectors. The context is denoted as a vector. Each dimension is the vector itself and the value stands for occurrence correlation. Then the contexts are brought together into same space using seed bilingual dictionaries. Then similarities are calculated using various methods like city-block metric [25], cosine similarity [14].

The words whose context similarity is above some limit and higher to others is treated as a translation pair. This is totally dependent on the words in a given bilingual lexicon. Hence, if the seed dictionary is smaller, the context vectors get scanty and the performance is affected leading to extraction of incorrect translation pairs. This questions the reliability of current methods when the language pairs are not closely related or when a large seed dictionary is unavailable.

Bilingual lexicon have been extracted using all three approaches i.e. supervised, less supervised as well as unsupervised methods. Supervised learning make use of training data of translated texts and bilingual dictionary. The system is trained to get correct translations of words. So when there is a new word the system can interpret what can be its translation. The second category of less supervised technique use comparable corpora which consists of texts that are not exact translations. Unsupervised learning has no training or precedence and it works upon suitable algorithms. The algorithms vary from one linguistic source to another. [18]. The algorithm used for parallel corpora and un-parallel corpora are not alike.

[2] [3] used a large parallel corpora of English-French Canadian parliamentary proceedings in their formative papers for extraction of translation pairs. Though the initial experiments had many limitations, but as research continued, a number of machine translation systems like Moses, Google Translate were successfully devised.
However parallel corpora is available only for a few languages, so the researchers shifted to the use of unrelated text or comparable corpora as first suggested by [9] and [25]. Comparable corpora are easily available, which solves the problem of obtaining data. For a given number of languages, same number of comparable corpora is sufficient. But in case of parallel corpora, if the translation is not available in many languages the number of required corpora increases four times

$$c = (n^2 - n)/2$$

Still, the task of extraction is much harder in comparable corpora as compared to parallel corpora. Hence despite intensive research (though many International projects are working upon comparable corpora) not much success has been achieved in the area. Research is going on by amending and combining different approaches and some remarkable improvements were suggested (such as use of pivot languages).
The extraction bilingual lexicon from comparable corpora can be done using Correlations, Vectors, Projections, Classifiers, Linguistic knowledge etc.

2. Correlation based extraction

Correlations between word co-occurrence and use of context of words for mapping source word to target word is the most recurrent method utilized for comparable corpora. These methods are used for extraction of bilingual lexicons and not for extracting parallel sentences.
The first and most relevant reference for bilingual lexicon extraction using un-parallel text for single word terms is probably [13] followed by [26].
[4] This approach also called Standard Method utilise the concept of vector formation. A context vector is modelled on the basis of its frequency and co-occurrences in a given window (±k words), in the entire corpus.
Context vectors for source and target language are calculated and using a bilingual dictionary, the context vectors of one language are translated into the other language. The similarity of vectors is calculated which is used to prepare a list of probable translations. The pairs having high probability are placed on the top. The ranking of translation pairs is dependent on the similarity of context vectors. More similar vectors will be probably ranked better. A context vector $\vec{w}$ represents the lexical context of a given word w. Context vectors can be calculated using different methods. The lexical context is represented by all the co-occurrences (sequence of words that occur together) $ct_1, ct_2, \ldots, ct_i$ of the given word w. The co-occurrences $ct_1, ct_2, \ldots, ct_i$ and value assoc(ct) associated to each co-occurrence ct are computed for each word (w) in the corpus. It can be done using Mutual Information [7] and Log Likelihood [8]. This association score also act as normalizing of context vectors.

	CO-OCCURRING TERM	¬ CO-OCCURRING TERM
WORD	a = cooc(word, co-occurring term)	B = COOC(WORD, ¬ CO-OCCURRING TERM)
¬ WORD	C = COOC(¬ WORD, CO-OCCURRING TERM)	D = COOC(¬ WORD, ¬ CO-OCCURRING TERM)

$$\text{Mutual Information}(ct) = \log \frac{a}{(a+b)(a+c)}$$

$$\text{LogLikelihood}(ct) = a\log(a) + b\log(b) + c\log(c) + d\log(d) + (a+b+c+d)\log(a+b+c+d) - (a+b)\log(a+b) - (a+c)\log(a+c) - (b+d)\log(b+d) - (c+d)\log(c+d)$$

Then the tail of context vectors is translated. The context vector word w (head) is not translated. The tail comprise of all the co-occurring terms $ct_1, ct_2, \ldots, ct_i$ of the given word w and their associated value

$$\text{assoc}(ct) = \text{Mutual Information}(ct)$$

or

$$\text{assoc}(ct) = \text{Log Likelihood}(ct)$$

A bilingual source/target dictionary is required to transfer context vectors into a target language. The usual case is that for one co-occurring source term ct_1 to translate there exists one translation $ct_{k_{trans}}$ in the bilingual dictionary. In this case, $\text{assoc}(ct_{k_{trans}}) = \text{assoc}(ct_k)$
If the translation of co-occurring source term ct_k, does not exist in the bilingual dictionary, it is then, not transferred into the translated context vector $\vec{w}_{trans}$. Then the similarity of vectors is calculated for both sides.
After measuring the similarities, the target words are ranked by the vector similarity computed in source side. The best ranked target words are the words that have the highest similarity score, and thus are sorted out as possible candidate translations. However the efficiency will be low if the data is less.

2.1. Context heterogeneity

Apart from context vectors another method used for extraction of bilingual lexicon from a comparable corpus is [9] context heterogeneity. This “Context Heterogeneity” is a 2-dimensional approach for each word on the right and left side. The context vector uses the unigram frequency and the number of unique words in a particular window. Euclidean distance is used to compute similarity.
The value of context heterogeneity of any function word would be close to one, since it can be followed by a number of words. Whereas the x value of the word which is always followed by a same word is less than one.
To measure the distance between two context heterogeneity vectors, simple Euclidean distance is used.

$$\varepsilon = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

If | ε | is close to zero then similarity is high.

2.2. The K-VEC method

In the K-VEC algorithm developed by Pascale [9] he prepared a list of lexicon candidates on the basis of how similarly the words of source and target language are distributed. The bilingual text is broken in K pieces and K-dimensional binary vectors are constructed for words on both sides. This method uses the mutual information score. It gives little efficiency in case of low frequency words. The t-score [17], [12] sheds the redundant values. If the value of K is very small, the results are inappropriate and in case of high value there will be no processing at all. Hence, [13] suggested that this method is not suitable for extraction of bilingual lexicon.

2.3. Convec

Fung also proposed a method called CONVEC, which uses similarity measured on TF-IDF [14] [10]. The TF-IDF weight (Term Frequency-Inverse Document Frequency) is a statistical technique to determine the importance of a word in an article in a large collection of texts. If the frequency of a word is high in the document the word is deemed important but if the frequency is high in the corpus, the value decreases.

2.4. Relation matrix

Word Relation Matrix identifies words from comparable corpora which can be translations [13]. The correlation vectors are constructed between source word and seed word on one side and between target word and seed word on the other side. For every unknown word in source language, its correlation with every word in the seed word list is identified and relation vectors are made. Similarly for unknown words in target language its correlation with every word in the seed word list is found to make a vector. The vectors are considered translations if their correlation value is high.

2.5. Phrase frequency based methods

This approach is based on the correlation between the co-occurrences of translation pairs of words [26]. It requires a small dictionary, which is to be expanded. The correlation vector is computed from the co-occurrence matrix. The rows of matrix consists of all words of the corpus and columns contains all the target words in the base dictionary. The unknown words are ignored. The positions of vectors are adjusted in relation to the vectors of the target language. With the resulting vector, similarity calculation based on log likelihood is used to calculate the similarity score.

2.6. Iterative extraction

This method extracts bilingual dictionary from comparable corpora [11] by using IBM Model 4. The documents are ranked using similarity scores and translation pairs are identified.
The method is further processed for bootstrapping assuming that there will be more pairs of translations. The iterations are done till convergence. This feature helps to add more parallel sentences from dissimilar documents.

$$Pr(w_s = 1) = \frac{a+b}{a+b+c+d}$$

$$Pr(w_t = 1) = \frac{a+b}{a+b+c+d}$$

$$Pr(w_s = 1 , w_t = 1) = \frac{a}{a+b+c+d}$$

a = number of segments where both words occur

b = number of segments where only w_s occur

c = number of segments where only w_t occur

d = number of segments where neither words occur

2.7. Combination of context and lexical information

Some researchers have also focused on a combination of different methods to get better results. [5] Have used the context vector as a basic component. The translation probabilities are calculated using a multilingual thesaurus. The combination increase the accuracy level as compared to the standard methods.

2.8. Domain specific bilingual dictionary extraction

Yun-Chuang Chiao and Pierre Zweigenbaum used a Medical corpus for extraction of bilingual lexicon.[4]. They used a window of seven words to model context vectors. The source vectors are translated using a small bilingual dictionary. Jaccard and Cosine similarity is used for comparison of all possible vectors.

$$Jaccard(V, W) = \frac{\sum_k v_k w_k}{\sum_k v_k^2 + \sum_l w_l^2 - \sum_m v_m w_m}$$

$$Cosine(V, W) = \frac{\sum_k v_k w_k}{\sqrt{\sum_k v_k^2}\sqrt{\sum_l w_l^2}}$$

2.9. Topic model based extraction

[31] The proposed TMBM approach assumes that the words that often occur in similar topics are more likely to be translations of one another. It uses a Bilingual Latent Dirichlet Allocation (BiLDA) topic model. The corpora is aligned document to document. This method is completely unsupervised. Later, [30] built an efficient dictionary combining linguistic knowledge with topic distribution. [21] constructed a parallel corpora from comparable corpora using BiLDA topic models to find translations.
It is difficult to compute similarity between a source word, w_e and a target word w_j directly since they are in different spaces and have different dimensions [5]. The projection based approach maps both source words with V_e dimension and target with V_j dimension into a latent semantic space with S, $z_e = M_e w_e$ and $z_j = M_j w_j$,where matrices M_e (V_e by S) and M_j (V_j by S) can be optimized by an Expectation Maximization learning algorithm for IBM Model 1. The similarity is computed with cosine similarity in semantic space, score (w_e, w_j) = cosine (w_e, w_j) = cosine ($M_e w_e$, $M_j w_j$) = cosine (z_e, z_j).

3. Vector representation

These methods use features other than context and using information retrieval algorithms to extract them. Every sentence is modelled as a vector and using similarity scores the probable candidates of translations are identified

3.1. Geometric interpretation of bilingual text

[15] Proposed a geometric method to extract translation pairs from non-parallel texts. Three new methods were devised to overcome the challenges like polysemy and coverage faced in other techniques. These methods do have the limitations as well but the desired levels of efficiency are achievable.
The source word is given as $s_i, 1 \le i \le p$ and target words is given as $t_j, 1 \le j \le q$. D (bilingual dictionary) is a set of n translation pairs (s_i, t_j), and is represented as a $p \times q$ matrix M, such that

$$\begin{cases} m_{ij} = 1 \ iff \ (s_i, t_j) \in D \\ \quad\ \ = 0 \ otherwise \end{cases}$$

The dot product between $\vec{v}$ and the translation of $\vec{w}$ calculates the level of similarity of vectors.

$$\langle \vec{v} , tr(\vec{w}) \rangle = \sum_{(e,f) \in D} a(v, e) a(w, f)$$

3.2. Text extraction based on signal processing

The work takes motivation from signal processing approach [23]. First of identical sentence pairs are identified by comparing the documents. Further all of the identified sentence pairs are passed through a filter for refinement. The sentences that have less number of translated words are filtered out and sentences which have high number of translated words are selected. Using a smoothing filter, the parallel fragments from sentences are detected. The filter selects those sentence pairs for which the corresponding values are high. The values are calculated using the probabilities and number of words.

4. Projection based approach

The similarity scores for the source words and target words cannot be calculated properly if the words are not brought together into common space. The words occur in different dimensions and spaces which make the calculations difficult. [5]. The projection based approach analyse the words semantically and brings the source words with V_e dimension and target words with V_j dimension into a common space which is latent with S, $z_e = M_e w_e$ and$z_j = M_j w_j$. The matrices M_e (V_e by S) and M_j (V_j by S) are optimized by an Expectation Maximization learning algorithm for IBM Model 1.

5. Classifiers based extraction

A classifier is an NLP tool which classifies the data into different classes. This classification can be used to separate the good sentence pairs from the bad pairs by using feature generation.

5.1. Maximum entropy model

A Maximum Entropy classifier is trained on parallel data using a dictionary [22]. Though the classifiers work upon a number of fetures, The Maximum Entropy classifier use the translation number feature alone. Keeping in view the restraints exercised by the features, the classifier suggest that the optimal parametric form of the model of data is a log linear combination of these functions. The parameter values which enhance accuracy of a given training corpus are calculated using algorithms like GIS or its improved version IIS. The feature function used here takes into account the explicit word alignment, length of the sentences, percentage of the words with translations from both side, percentage of words with no translations, alignment score, length of the longest span etc.

5.2. Support vector machine

C. Brockett and W.B. Dolan worked upon a comparable corpus of news extracted from online sites using Support Vector Machines to find larger monolingual para-phrases. The data is properly labelled and tagged. [1]. It uses many features like diversity of morphology, synonyms and hypernyms, log-likelihood-based word pairings dynamically obtained from baseline sentence alignments and formal string features such as word-based edit distance. This technique improves the alignment rate of the corpora. The heuristic methods which depend on the position of the sentences in the text have more errors.

6. Parallel corpora based approach

Some researchers have converted comparable corpora into parallel corpora for the extraction of bilingual lexicon. The dictionary is constructed through word alignment models. The corpora can be constructed using different methods. [23] Used suffix trees to build parallel corpora. This method worked well on languages having similar word order but not on languages having different word order. [29] Constructed parallel corpora by using seed dictionary features and orthographic features.

7. Linguistic knowledge based extraction

Linguistic knowledge is used to prepare templates, which are then used as filters to identify correct translation. These methods are extremely language dependent.

7.1. Lexico-syntactic methods

This approach extracts translation pairs from comparable corpora without requiring external bilingual resources [24].Lexico-syntactic templates are extracted from parallel texts and their bilingual correspondences are identified to find meaningful bilingual anchors within the corpus. First of all text processing is done followed by extraction of bilingual lexico-syntactic templates from parallel corpora and extraction of word translation from comparable text using these templates. The standard and context based methods relies on the same notion of similarity. The similarity between lemmas is calculated using the following non-weighted Dice coefficient:

$$Dice(l_1, l_2) = \frac{2 \times \sum_i min(f(l_1,t_i),(f(l_2,t_i)}{f(l_1)+f(l_2)}$$

Where $f(l_1, t_i)$ represents the number of times lemma l_1 occurs in a seed template t_i. In addition, $f(l_2, t_i)$ represents the number of times lemma l_2 occurs in a seed template t_i.

Chenhui Chu has proposed a novel approach of bilingual lexicon extraction by combining topic model and context based methods. The system does not rely on any prior knowledge and the performance can be iteratively improved. Experiments conducted on Chinese-English, Japanese-English and Chinese-Japanese Wikipedia data show that the proposed method signicantly outperforms the previous studies. The approach is language independent and can be used for other language pairs as well.

8. Proposed method

Taking cue from it we have proposed a method for extracting bilingual lexicon from English-Punjabi comparable corpora. The Web has been exploited as a source of comparable corpora, and many studies have been conducted for constructing parallel corpora from it. However one common fact among the previous research is use of English, German, Spanish, Arabic, Chinese, Japanese corpora. We have used English- Punjabi corpora with no precedence in the area. A number of different approaches and techniques have been explored for the extraction.

The topic model based method (TMBM) and context based method (CBM) use distributional concept that a word having similar meaning to a word of another language will be distributed in similar ways. TMBM computes the similarity of two words on topical distributions across languages, and CBM is based on similarity of context.

Our idea is to present an extraction system which combines TMBM and CBM. It will utilize topic distribution as well as context similarity thus increasing the accuracy of the bilingual lexicon. The data is aligned in document form. For the purpose of extracting bilingual lexicon we train BiLDA topic model and extract topical bilingual lexicons. The topical bilingual lexicon is used as seed dictionary in context based method to obtain context bilingual lexicons.

Combination

TMBM measures the distributional similarity of two words on cross-lingual topics, while CBM measures the distributional similarity on contexts across languages. A combination of these two methods can exploit both topical and contextual knowledge to measure the distributional similarity, making bilingual lexicon extraction more reliable and accurate. Here we use a linear combination for the two methods to calculate a combined similarity score:

$$Sim_{Comb}(w_i{}^S, w_i{}^T) = \gamma Sim_{Topic}(w_i{}^S, w_i{}^T) + (1 - \gamma) Sim_{Context}(w_i{}^S, w_i{}^T)$$

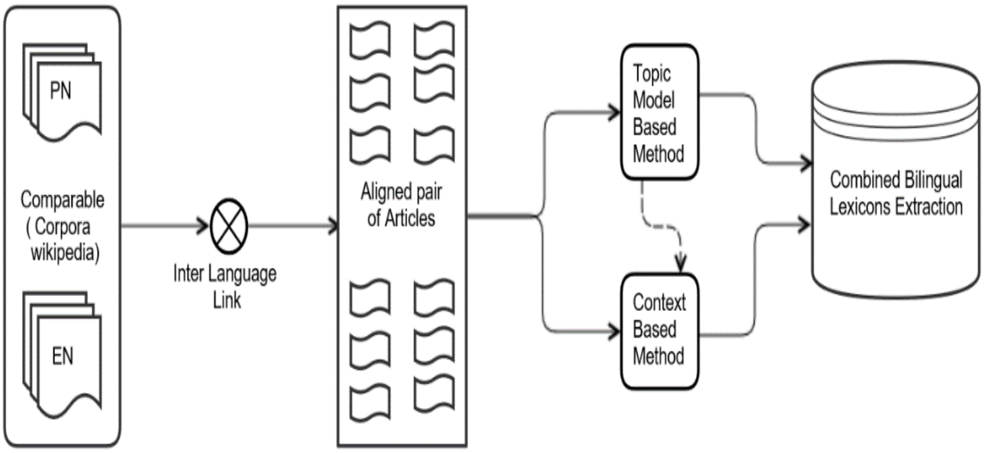

Fig. 1: Integral Approach for Bilingual Lexicon Extraction.

In this way, we extract lexicon from comparable English-Punjabi corpus. This system can however be improved upon further and accuracy can be increased. It currently deals only with simple noun words and cannot handle compound words and rare words.

9. Conclusion

Hence a wide variety of methods and approaches have been developed which use different algorithms for the extraction of bilingual extraction. However these approaches have their limitations like high levels of complexity, sensitivity to parameters, size of corpus, low accuracy etc. The effective way to overcome the problem can be to use a combined approach as proposed above i.e. to combine two or three methods for better performance so that the limitation of one method can be tackled with other.

References

[1] Brockett, C. (2005). Support vector machines for paraphrase identification and corpus construction. Proceedings of the third International Workshop on Paraphrasing (IWP), 1–8. Retrieved from http://acl.ldc.upenn.edu/I/I05/I05-5001.pdf

[2] Brown, P. F., Cocke, J., Pietra, S. a Della, Pietra, V. J. Della, Jelinek, F., Lafferty, J. D., Watson, T. J. (1990). Statistical approach to machine translation. Computational Linguistics, 16(2), 79–85. https://doi.org/10.3115991365.991407

[3] Brown, P. F., Pietra, S. A. Della, Pietra, V. J. Della, & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics, 19, 263–311. Retrieved from http://dl.acm.org/citation.cfm?id=972474

[4] Chiao, Y., Zweigenbaum, P., Dsi, S., Publique, A., Paris, H. De, Biomathématiques, D. De, & Paris, U. (2002). Looking for candidate translational equivalents in specialized, comparable corpora. Proceedings of the 19th International Conference on Computational Linguistics, 3–7. https://doi.org/10.3115/1071884.1071904

[5] Déjean H, Gaussier É, S. F. (2002). An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. Proceedings of the 19th International Conference on Computational Linguistics (COLING), 1–7. https://doi.org/http://dx.doi.org/10.3115/1072228.1072394

[6] Deng, Y. (2005). Bitext Alignment for Statistical Machine Translation. PhD Thesis, 29–51.

[7] Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. Computational Linguistics, 19, 61–74. Retrieved from http://portal.acm.org/citation.cfm?id=972454

[8] Fano, R. M., & Hawkins, D. (1961). Transmission of Information: A Statistical Theory of Communications. American Journal of Physics, 29(11), 793–794. https://doi.org/10.1119/1.1937609

[9] Fung, P. (1995). Compiling Bilingual Lexicon Entries from a Non-Parallel English-Chinese Corpus A Non-parallel Corpus of Chinese and English. Proceedings of the Third Workshop on Very Large Corpora, 173–183.

[10] Fung, P. (1998). A statistical view on bilingual lexicon extraction: from parallel corpora to non-parallel corpora. Computer Vision and Mathematical Methods in Medical and Biomedical Image Analysis, 1529, 1–17. Retrieved from papers2://publication/uuid/8A778A29-6509-4FF8-95F5-D283E5D5AC76

[11] Fung, P., & Cheung, P. (2004). Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus, 1051. https://doi.org/10.3115/1220355.1220506

[12] Fung, P., & Church, K. W. (1994). K-vec: A new approach for aligning parallel texts. Proceedings of the 15th Conference on Computational Linguistics: Volume 2, 2, 1096–1102. https://doi.org/10.3115/991250.991328

[13] Fung, P., & McKeown, K. (1994). Aligning Noisy Parallel Corpora across Language Groups : Word Pair Feature Matching by Dynamic Time Warping. Proceedings of AMTA94 Association of Machine Translation in the Americas, 8. Retrieved from http://arxiv.org/abs/cmp-lg/9409011

[14] Fung, P., & Yee, L. Y. (1998). An {IR} Approach for Translating New Words from Nonparallel, Comparable Texts. Proceedings of the 36th Annual Meeting of the ACL and 17th International Conference on Computational Linguistics: COLING/ACL-98, 414–420. https://doi.org/10.3115/980845.980916

[15] Gaussier, E., Renders, J.-M., Matveeva, I., Goutte, C., & Dejean, H. (2004). A Geometric view on bilingual lexicon extraction from comparable corpora. Association for Computational Linguistics, 1529, 1–17. https://doi.org/10.3115/1218955.1219022

[16] Haghighi, A., Liang, P., Berg-Kirkpatrick, T., & Klein, D. (2008). Learning Bilingual Lexicons from Monolingual Corpora. In Proceedings of ACL-08: HLT, 2008(June), 771–779. Retrieved from http://www.researchgate.net/publication/220873349_Learning_Bilingual_Lexicons_from_Monolingual_Corpora/file/3deec52254895b9903.pdf

[17] Kenneth, Erlbaum, N. J. L., Church, Gale, W., Hanks, P., & Hindle, D. (1991). Lexical Acquisition : Exploiting On-Line Resources to Build a Lexicon. Association for Computational Linguistics, 214–216.

[18] Koehn, Philipp; Knight, K. (2001). Knowledge sources for word-level translation models. Proceedings of the Conference on Empirical Method in Natural Language Processing.

[19] Koehn, P., & Knight, K. (2002). Learning a translation lexicon from monolingual corpora. Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition, 9(July), 9–16. https://doi.org/10.3115/1118627.1118629

[20] Levow, G. A., Oard, D. W., & Resnik, P. (2005). Dictionary-based techniques for cross-language information retrieval. Information Processing and Management, 41(3), 523–547. https://doi.org/10.1016/j.ipm.2004.06.012

[21] Liu, X., Duh, K., & Matsumoto, Y. (2013). Topic Models+ Word Alignment= A Flexible Framework for Extracting Bilingual Dictionary from Comparable Corpus. Proceedings of the 17th Conference on Computational Natural Language Learning (CoNLL '13), 212–221. Retrieved from https://www.aclweb.org/anthology-new/W/W13/W13-35.pdf#page=224

[22] Munteanu, D., & Fraser, A. (2004). Improved machine translation performance via parallel sentence extraction from comparable corpora. HLT-NAACL, Main Proce(Boston, Massachusetts, USA, May. Association for Computational Linguistics), 265–272. Retrieved from http://acl.ldc.upenn.edu/hlt-naacl2004/main/pdf/93_Paper.pdf

[23] Munteanu, D. S., & Marcu, D. (2006). Extracting parallel sub-sentential fragments from non-parallel corpora. Proceedings of ACL, 44(July), 81–88. https://doi.org/10.3115/1220175.1220186

[24] Otero, P. G. (2007). Learning Bilingual Lexicons from Comparable English and Spanish Corpora. Proceedings of MT Summit XI, 191--198.

[25] Rapp, R. (1995). Identifying Word Translations in Non-Parallel Texts. Proceedings of the 33rd ACL, Cambridge, MA, 320–322.

[26] Rapp, R. (1999). Automatic identification of word translations from unrelated English and German Corpora. Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, 519–526. https://doi.org/10.3115/1034678.1034756

[27] Sadat, F., Yoshikawa, M., & Uemura, S. (2003). Bilingual terminology acquisition from comparable corpora and phrasal translation to cross-language information retrieval. Association for Computational Linguistics, 141–144. https://doi.org/10.3115/1075178.1075201

[28] Shao, L., & Ng, H. T. (2004). Mining new word translations from comparable corpora. Proceedings of Coling 2004, 618–es. https://doi.org/10.3115/1220355.1220444

[29] Smith, J. R. J. R., Quirk, C., & Toutanova, K. (2010). Extracting parallel sentences from comparable corpora using document level alignment. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, (June), 403–411. Retrieved from http://dl.acm.org/citation.cfm?id=1858062

[30] Vulić, I., & Moens, M.-F. (2012). Detecting highly confident word translations from comparable corpora without any prior knowledge. Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, 449–459.

[31] Vulić, I., Smet, W. De, & Moens, M. (2011). Identifying word translations from comparable corpora using latent topic models. Proceedings of the 49th Annual Meeting …, 479–484. Retrieved from http://dl.acm.org/citation.cfm?id=2002832