Combining lexical context with pseudo-alignment for bilingual lexicon extraction from comparable corpora

Bo Li, Qunyan Zhu, Tingting He, Qianjun Chen

The center of the national language tracing and research for the network media,
National Engineering Research Center for E-Learning,
Network Center of Hubei University,
School of Computer, Central China Normal University, Wuhan, 430079, China
liboccnu@126.com,zhuqunyan@yeah.net,
tthe@mail.ccnu.edu.cn,skysky@hubu.edu.cn

Abstract. Only a few studies have made use of alignment information in bilingual lexicon extraction from comparable corpora, in which comparable corpora are necessarily divided into 1-1 aligned document pairs. They have not been able to show extracted lexicons benefit from the embedding of alignment information. Moreover, strict 1-1 alignments do not exist broadly in comparable corpora. We develop in this paper a language-independent approach to lexicon extraction by combining the classic lexical context with pseudo-alignment information. Experiments on the English-French comparable corpus demonstrate that pseudo-alignment in comparable corpora is an essential feature leading to a significant improvement of standard method of lexicon extraction, a perspective that have never been investigated in a similar way by previous studies.

Keywords: comparable corpora, pseudo-alignment, bilingual lexicon extraction, context-vector

1 Introduction

Bilingual dictionaries are bridges between two different languages and important resources for empirical multilingual natural language processing tasks such as cross-lingual information retrieval (CLIR) [1] and statistical machine translation [13]. With the high-speed development of international communication, the bilingual dictionary demand grows highly accordingly. Hand-coded dictionaries are of high quality, but it is expensive to build and researchers have tried, since the end of the 1980s, to automatically extract bilingual lexicons from parallel corpora [7, 10, 11, 17]. Parallel corpora are however difficult and time consuming to get in several domains and the majority of bilingual collections are comparable and not parallel. Due to their more abundant, less expensive and low cost of acquisition via web, various methods have been previously proposed to extract bilingual lexicons from comparable corpora [3, 4, 5, 6, 8, 12, 15, 19].

Most works in bilingual lexicon extraction follow the assumption that words in translation should have similar context in both languages. Based on this assumption, a standard approach usually builds context vectors for each word of the source and

adfa, p. 1, 2011.

target languages. The candidate translations for a particular word are obtained by comparing the translated source context vector with the target context vector using a general bilingual dictionary [8], and it has been proved to get a good performance in previous works.

In addition to context information, heuristics are often used to improve the general accuracy of the context-vector approach, like orthographic similarities between the source and the target terms [16]. Cognate-based techniques are popular in bilingual term spotting, in particular for specific domains. It can be explained by the large amount of transliteration even in unrelated languages. Also, related languages like Latin languages can share similarities between a term and its translation, like identical lemmas.

As far as we could tell, only a few studies have paid attention to the co-occurrence information between words in aligned documents [14, 18]. Those approaches did not take lexical context into account, resulting in poor performance compared to work in the same vain. Moreover, comparable corpora in those studies were divided into strict 1-1 aligned document pairs which, however, did not exist broadly in comparable corpora. For instance, [14] aimed to enhance the performance of lexicon extraction for those rare words. They did make use of the alignment information in a machine learning manner, which however relied on strictly 1-1 alignments and expensive training process. Moreover, by extracting aligned pairs, they had actually reduced a lot the size of original corpora and suffered from great information loss.

However, factors affecting the meaning of a word are far more than this, let alone the task is dealing with two languages of different cultural background. Besides, due to *polysemy* and *homonymy*, it is hard to identify precise translations of a word according to a single type of feature, a conclusion that can be drawn from poor performance in previous studies [4]. A natural solution to this problem is to resort to comprehensive features, especially those reflecting different aspects of a word. Therefore, in this paper, we propose a comprehensive approach considering the lexical context together with the co-occurrence feature in loose aligned document pairs.

The rest of the paper is organized as follows: Section 2 reviews related works on some popular methods in lexicon extraction from comparable corpora. Section 3 presents the combined model we proposed: we firstly present the method to calculate similarity using context information and then discuss the principle and method to build word co-occurrence. Section 4 shows the experimental setup and evaluation of our method on three groups of corpora. Conclusions will be drawn in section 5.

2 Related Work

Several research works have been done on the task of extracting bilingual lexicon from comparable corpora. Most approaches are based on the same intuitive assumption of word distribution, that is words appear in context of the same form or semantic are attended to be translation pairs. The starting point of the strategy is as follows: for

each word w, the context information of w is described in a certain way, such as a vector, and then we can get the translation candidates of w by ranking the context similarity between w and any target word. This method was often been conducted with an existing bilingual resources [6] or parser tool [5].

Existing research works are trying to improve performance through two ways: one is trying to find another way to describe the context which contains more context information; the other is to find a more effective way to measure the similarity between word contexts. Such as [3] defines words in a fixed size of window to be the context, and tests several different models in bilingual lexicon extraction from parallel or comparable corpora in specialized domains. It shows that the combination of multi models significantly improves results, and that the use of the thesaurus UMLS/MeSH is of primary importance. But [5] improves the context information by using dependency parsing. It uses contexts derived from head-words linked by dependency trees instead of the immediate adjacent lexical words. With the deep semantic information, it gains significant improvement compared to approaches solely relying on the lexical context. [6] presents a geometric view on bilingual lexicon extraction from comparable corpora, which can interpret all the context vector approaches in a uniform framework. It uses singular value decomposition (SVD) to map the original context vector to another space in which synonyms dictionary entries are close to each other, while polysemous ones still select different neighbors. The precision is proved to be improved. [12] tries to extract French-Japanese terminologies from comparable corpora, considering both single and multi-word terms. They show the fact that the quality of comparable corpora is more important than the quantity and that this ensures the quality of acquired terminological resources. It also concludes that the quality of co-occurrence vectors can be substantially improved by ensuring domain and discourse comparability of the corpora from which co-occurrences are obtained. [18] introduces a new way to align two document collections in different languages, and test the effectiveness of several combined CLIR approaches based on comparable corpora, dictionary-based query translation, and pseudo-relevance feedback. But it did not take the co-occurrence information of words in document pairs into account. [14] incorporates the lexical context similarity with the co-occurrence model between words in strict aligned documents to extract bilingual lexicons, which mainly tries to solve the problem of data sparseness of rare words. Furthermore, comparable corpora in [14, 18] were divided into strict 1-1 aligned document pairs that do not exist broadly in comparable corpora. In this case, the size of original corpora was greatly reduced while retaining the aligned document pairs, leading to great information loss. In addition, a large volume of training data was needed, which makes the method infeasible in general extraction task.

It has been proved in most previous studies that lexical context is important for bilingual lexicon extraction. The pseudo-alignments in comparable corpora are another feature we deem to be important. We thus plan to combine these two features to form a comprehensive model to extract bilingual lexicon from comparable corpora.

3 The combination model

3.1 Lexical context similarity

We implemented the context-vector similarity in a way similar to Pascale Fung [9], using context information to extract bilingual lexicon. It assumes that the words in the source and target language are likely to be mutual translations if their context is similar. Based on the assumption, the standard approach builds a context vector respectively for the source and target word. Then the context vector of the source word is translated to the target language, so that we can compare the source context vector with the target context vector and a similarity between them is also calculated. The context vector is built as follows:

For an English word w_e, we collect its context words in the entire English corpus, then a context vector $\overrightarrow{v_e}$ for w_e is built, the dimension of the $\overrightarrow{v_e}$ is the same with the number of entries the lexicon has, the value of the i-th dimension of $\overrightarrow{v_e}$ is W_{ie} :

$$W_{ie} = TF_{ie} \times IDF_i \tag{1}$$

where TF_{ie} represents the number of times the i-th word in the dictionary appears in the context of w_e.

$$IDF_i = \log\frac{maxn}{n_i} + 1 \tag{2}$$

where maxn is the maximum frequency of any word in the corpus, n_i is the total number of occurrences of word w_e in the corpus.

For a French word w_f , we acquire a context word vector $\overrightarrow{v_f}$ in a similar manner as the English word w_e.

Once the context vector of each word has been built, the problem is to measure the similarity between two words. In the cross-language settings, one needs to compare two vectors in different languages, i.e. one vector in the source language needs to be mapped to a vector in the target language. In this paper, $\overrightarrow{\mathrm{v_f}}$ is mapped to $\overrightarrow{\mathrm{v_e}}$ by accumulating the contributions from words in $\overrightarrow{\mathrm{v_f}}$. Here we calculate the contributions by adding up the weights of words in $\overrightarrow{\mathrm{v_f}}$ with the identical translations in $\overrightarrow{\mathrm{v_e}}$. We used the Cosine similarity for context-vector comparisons, which has often been shown to achieve superior results in comparative studies. With the cross-language vector mapping and the cosine formula, the similarity between words w_e and w_f is computed as:

$$S_c(w_e, w_f) = \cos(\overrightarrow{v_e}, \overrightarrow{v_f}) = \frac{<\overrightarrow{v_e}, \overrightarrow{v_f}>}{\sqrt{\sum_{w_e \in \overrightarrow{v_e}} (f(w_e)^2 + (\sum_{w_f \in T_{w_e} \cap \overrightarrow{v_f}} f(w_f))^2)}} \tag{3}$$

where T_{we} is the translation set of w_e in the bilingual dictionary.

3.2 Pseudo-alignment similarity

According to the characteristics of the comparable corpus, Comparable documents are not strictly parallel, but have overlapping information. Therefore, words occur in a

pair of comparable documents tends to have more probability of being translation candidates of each other. So in this work, first, we get a loose alignment between the two corpora, which has great differences with strict 1-1 alignment in other works and it is more suitable for the reality. Then, we propose a quantity which is large if w_e and w_f appear in many document pairs with a high comparability score, and small otherwise.

In order to establish loose alignments among documents, one in fact needs to measure the similarity of each document pair consisting of documents in two languages. We directly use here the measure $\mathrm{M}(\mathrm{C}_e, C_f)$ proposed by [9]. The measure is light-weighted and does not depend on complex resources like the machine translation system. Let us assume we have an English document C_e and a French document C_f, then $\mathrm{M}(\mathrm{C}_e, C_f)$ measures the proportion of the English and French words for which a translation can be found in the document pair, that is:

$$\mathrm{M}(C_e, C_f) = \frac{\sum_{w \in W_e^C \cap D_e^v} \mu(\mathrm{w}, W_f^C) + \sum_{w \in W_f^C \cap D_f^v} \mu(\mathrm{w}, W_e^C)}{|W_e^C \cap D_e^v| + |W_f^C \cap D_f^v|} \tag{4}$$

where W_e^C (resp. W_f^C) is the set of all words which appear in the English(resp. French) corpus . D_e^V (resp. D_f^V) is the English (resp. French) part of a given, independent bilingual dictionary. μ is a function indicating whether a translation from the translation set T_w of w is found in the French word set, that is:

$$\mu(\mathrm{w}, W^C) = \begin{cases} 1 & iff \ \ T_w \cap W^C \neq \emptyset \\ 0 & else \end{cases} \tag{5}$$

In order to measure the co-occurrence feature of w_e and w_f, first, we define the joint probability of w_e and w_f as (6), which is in direct proportion to the number of comparable document pairs they occur in, that is:

$$\mathrm{p}(w_e, w_f) \propto \sum_{d_e \in D_e, d_f \in D_f} \delta(d_e, d_f) \tag{6}$$

where D_e (resp.D_f) is the set of documents containing word w_e (resp. w_f). $\delta(d_e, d_f)$ is defined as:

$$\delta(d_e, d_f) = \begin{cases} 1 & if(M(d_{e,} d_f) \geq \eta) \\ 0 & else \end{cases} \tag{7}$$

where η is the threshold to judge the comparability of two corpus or documents. Here we can conclude that the co-occurrence probability of w_e and w_f is in direct proportion to the number of comparable document pairs they occur in. According to equation (6), the marginal probability of w_e is:

$$\mathrm{P}(w_e) = \sum_{w_f \in W_f^C} p(w_{e,} w_f)$$

$$= \sum_{w_f \in W_f^C} \sum_{d_e \in D_e, d_f \in D_f} \delta(d_e, d_f)$$

$$= \sum_{d_e \in D_e} \sum_{d_f \in D_f^C} |d_f| \cdot \delta(d_e, d_f) \tag{8}$$

where D_f^C is the set of all documents in French corpus. Our corpora is large enough to assume that all d_f in D_f^C have almost the same vocabulary size and all d_e have the same number of comparable counterparts in D_f^C, then $\mathrm{p}(w_e) \propto |D_e|$. Point-wise mutual information (PMI) is a measure of association widely used in information theory and statistics. PMI is first proposed by [2] as equation (9), Here we use the $\mathrm{PMI}(w_e, w_f)$ to judge how relevant the source word w_e is to the target wordw_f:

$$\mathrm{PMI}(w_e, w_f) = \log \frac{p(w_e, w_f)}{p(w_e) \cdot p(w_f)} \tag{9}$$

Then, we proposed a quantity $\varphi(w_e, w_f)$, which is in direct proportion to the PMI of w_e and w_f:

$$\varphi(w_e, w_f) = \frac{\sum_{d_e \in D_e} \sum_{d_f \in D_f} \delta(d_e, d_f)}{|D_e| \cdot |D_f|} \tag{10}$$

which is easy to compute and has the desired property: it is reasonable to measure the co-occurrence feature of w_e and w_f.

The lexical context is a classic feature that has been proved to be efficient in lexicon extraction. The pseudo-alignment in comparable corpora is another feature we deem to be important. Those two features can be combined to form a comprehensive measure and we thus obtain:

$$S_{coo}(w_e, w_f) = S_c(w_e, w_f)(1 + \varphi(w_e, w_f)) \tag{11}$$

This is simply the product of two similarities from different aspects.

4 Experiments and Results

4.1 Experimental setup

We perform our lexicon extraction on English and French comparable corpora which were used in the multilingual track of CLEF (http://www.clefcampaign.org), including the Los Angeles Times (LAT94, English), Glasgow Herald (GH95, English), Le Monde (MON94, French), SDA French 94(SDA94, French) and SDA French 95(SDA95, French). To gain the diversity of the corpora, two monolingual corpora from the Wikipedia dump[1] were built. English articles are retained below the root category *Society* and French articles are extracted from the *Société* category. A dump contains text, but also some special data and syntax (images, internal links, etc.) which are not interesting for our experiments. We remove all tags in the collection. A stop-word list is used to filter the textual content of the articles. The English corpus consists of 533k documents and the French one consists of 508k documents. Table 1

[1] http://download.wikimedia.org

contains the details about the comparable corpora in which documents containing less than 30 words have been deleted.

Standard preprocessing steps: tokenization, POS-tagging by Tree tagger and lemmatization are performed on all the linguistic resources. We will directly work on lemmatized forms of content words (nouns, verbs, adjectives, adverbs).

Table 1. The size of the corpora

	language	Docs
CLEF corpora	English	165 K
	French	130 K
Wikipedia Corpora	English	368 K
	French	378 K

The seed lexicon used in our experiments is constructed from an online dictionary. It consists of 33k distinct English words and 28k distinct French words, which constitutes 76k translation pairs.

In order to measure the performance of the bilingual lexicon extraction method presented above, we divided the original seed dictionary into 2 parts: 10% of the English words (about 3k words) together with their translations are randomly chosen and used as the evaluation set, the remaining words being used to compute context vectors and similarity between them.

4.2 Experiment groups

While getting the POS information of the corpus, we divided the entire corpus to three parts which were the three groups of corpora using in the experiments. Then we randomly pick some files from them to do the experiments, documents in each language have the same number. The source and size of the final corpus is listed in Table 2.

Table 2. The source and size of the corpora in the experiments

group	Data	Language	Source	Size
1	CLEF Base corpus	English	GH95	19K
		French	SDA95	19K
2	CLEF Base corpus + CLEF extend corpus	English	GH95, LAT94	46 K
		French	SDA95, MON94,SDA94	46 K
3	CLEF Base corpus+ Wikipedia corpus	English	GH95, Layer less than 4 under Society category,	54 K
		French	SDA95, Layer less than 7 under Société category	54 K

While building the context vector of the source and target corpora, syntactic contexts are considered to be less ambiguous and more sense-sensitive than contexts

defined as windows of size N for the reason of no structural damage to a complete sentence. But in our work, the Wikipedia texts are not so standard in grammar because of their voluntary editor. Therefore, we use the period concluding a sentence combined with a context widow of size 10 to define the range of context.

4.3 Results and analysis

After calculating similarities based on the baseline method and the one developed in this paper, for each word in the test set, we list their French translation candidates which are ranked by the two methods of calculating similarity respectively. In order to test the performance of the words in the evaluation lists, we get the top N ranked translate candidates of each word, then measure the precision rate and recall rate. In addition, several studies have proved that it is easier to find the correct translations for frequent words than for infrequent words, to take this fact into account, we distinguished different frequency ranges to assess the validity of our approach for all frequency ranges, Words with frequency less than 100 are defined as low-frequency words (W_L), while words with frequency larger than 400 are high-frequency words (W_H), and words with frequency in between are medium-frequency words (W_M). The results obtained when using the lexical context information alone (baseline) and the refined combined model proposed in this paper (new) were displayed in Table 3. The relative improvement is further shown in Table 4. Gi' (i=1, 2, 3) stands for the result of group Gi' while using the improved method.

Table 3. Precision and recall rate of two approaches on three groups

	Precision		Recall	
	baseline	new	baseline	New
G1	0.250	0.251	0.106	0.106
G2	0.277	0.293	0.122	0.127
G3	0.325	0.345	0.145	0.153

Table 4. Improvement of presicion for words in different frequcney ranges. G2' and G3' denotes performance of our approach on test group G2 and G3 respectively.

	G2	G3	G2'	> G2	G3'	> G3
W_L	0.167	0.206	0.175	4.8%	0.221	7.3%
W_M	0.338	0.390	0.358	5.9%	0.420	7.7%
W_H	0.561	0.632	0.599	6.8%	0.657	4.0%
All	0.277	0.325	0.293	**5.8%**	0.345	**6.2%**

Table 3 shows the overall results on three groups of test corpora obtained with our approach as well as the baseline method. One can find that performance of lexicon extraction, in terms of both precision and recall, has been enhanced on all groups, although the improvement is more remarkable on G2 and G3. It is a tough task to improve the performance of lexicon extraction when considering target words distributed in all frequency ranges[6, 9], compare to those studies only focusing on words of

high frequency [4, 12]. According to the experiment results obtained in previous work for the same task [9], our results here should be considered as important, although the increase in terms of precision is not that much. We also notice from table 3 that the performance does not change much on group G1 with our approach. The reason is that corpora in group G1 are of small size where context information and alignments of high quality do not exist in large scale. We can draw a conclusion out of these results: the size of corpus influences the quality of bilingual lexicons extracted with the method proposed in this paper.

We will give here more comments on translation performance for words distributed in different frequency ranges, since it is much easier to translate words of higher frequency [4, 12]. The improvement on group G1 is not that significant and we only focus on those on groups G2 and G3 and list detailed results in table 4. One could find a consistent improvement for words in all the frequency ranges. When using the improved approach, the precision of low-frequency words is strongly improved from 0.167 to 0.175 (corresponding to a relative increase of 4.8%) on group G2, from 0.206 to 0.221 (corresponding to a relative increase of 7.3%) on group G3. For middle frequency words, the precision is relatively increased by 5.9% on group G2 and 7.7% on group G3. Lastly, for high frequency words, the performance is also significantly improved: from 56.1% to 59.9% (corresponding to a relative increase of 6.8%) on group G2 and from 63.2% to 65.7 % (corresponding to a relative increase of 4%) on group G3. We can thus conclude that our approach performs consistently on all frequency ranges. Especially for those low frequency words of which the performance is rather difficult to improve, our approach has shown a consistent and satisfactory enhancement.

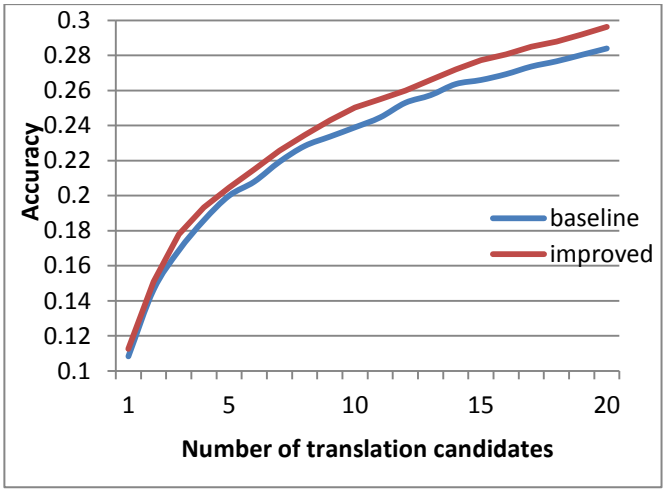

Fig.1. Comparison of the average precision for the top 1 to 20 candidates with the baseline approach and our improved approach on group G2.

The average precision of three groups on different approaches is displayed in Fig 1. From it, one can see that the overall average precision is further improved by 1.3% compared with the baseline. For N from 1 to 20, there is always a significant im-

provement in the precision and recall rate. The most fastest rising appears in top 1 to top 5, this is mainly because correct translations can be easily found in top 5 with only little random error, with the increase of number of candidates, the accuracy does not increase so sharply means the candidates among top 10 to 20 are usually words relates to the true translations, so one can refine the model eventually to get better precision.

5 Conclusion

We have proposed in this paper a combined model to improve the efficiency of bilingual lexicon extraction from comparable corpora. This model combines the traditional lexical context information with the word co-occurrence in loosely aligned documents to compute the similarity between two words. It has been proved to be effective mainly because the novel model has taken into account various characteristics that could reflect word meaning from a comprehensive perspective. We have first established in our approach loosely aligned document pairs relying on a light-weighted comparability measure proposed in [9]. Contrary to previous studies, the pseudo-alignment in our work does not need to be strictly 1-1 style, which is compliant with the real case in practice. The co-occurrence information of words in aligned documents is then incorporated into the traditional model solely relying on lexical context similarity to form a comprehensive model. Experiments have shown that translation precision can be improved significantly on all test groups, i.e. a relative improvement of 6.2% on group G3 from 32.5% to 34.5%. In addition, we have noticed much more improvement on those corpora of larger size where alignment of higher quality could be found easier. In future work, we will try to discover and incorporate more influential factors to measure the similarity of words in two languages.

Acknowledgement

This work was supported by the Major Project of National Social Science Fund (No. 12&2D223), the Natural Science Foundation of China (No. 61300144), the Natural Science Foundation of Hubei Province (No.2011CDA034), the Major Project of State Language Commission in the Twelfth Five-year Plan Period (No.ZDI125-1), the Project in the National Science &Technology Pillar Program in the Twelfth Five-year Plan Period (No.2012BAK24B01), the Program of Introducing Talents of Discipline to Universities (No.B07042), and the self-determined research funds of CCNU from the colleges' basic research and operation of MOE(No. CCNU13A05014, No. CCNU13C01001, No. CCNU13F010).

References

1. Ballesteros, L., Croft, W.B.: Phrasal translation and query expansion techniques for cross-language information retrieval. In: SIGIR, pp. 84–91 (1997)
2. Church, K.W.; Hanks, P.: Word Association Norms, Mutual Information, and Lexicography. Computational Linguistics, 16(1):22–29(1990)

3. Déjean, H., Gaussier, E., Sadat, F.: An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In Proceedings of the 19th international conference on computational linguistics (COLING'02). Taipei, Taiwan, pp. 218–224(2002)
4. Fung, P., Yee, L.Y.: An IR approach for translating new words from nonparallel, comparable texts. In: 17th international conference on computational linguistics. Montreal, Quebec, Canada, pp. 414–420(1998)
5. Garera, N., Callison-Burch, C., Yarowsky, D.: Improving translation lexicon induction from monolingual corpora via dependency contexts and part-of-speech equivalences. In: Proceedings of the 13th Conference on Computational Natural Language Learning, pp. 129–137(2009)
6. Gaussier, E., Renders, J.-M., Matveeva, I., Goutte, C., Déjean, H.: A geometric view on bilingual lexicon extraction from comparable corpora. In: 42nd annual meeting of the Association for Computational Linguistics. Barcelona, Spain, pp. 526–533(2004)
7. Kay, M. and Röscheisen, M.: Text-Translation Alignment. Computational Linguistics 19,121–142(1993)
8. Laroche, A., Langlais, P.: Revisiting Context-based Projection Methods for Term translation spotting in Comparable Corpora. In: Proceedings of the 23rd Coling Conference, Beijing, China, pp617–625 (2010)
9. Li, B., Gaussier, E.: Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010). Beijing (2010)
10. Melamed, I. D.: A portable algorithm for mapping bitext correspondence. In: Proceedings of the 35th Annual Meeting of the ACL (1997)
11. Melamed, I. D.: A word-to-word model of translational equivalence. In: Proceedings of the 35th Conference of the Association for Computational Linguistics (ACL'97), Madrid, 490–497(1997)
12. Morin, E., Daille, B., Takeuchi, K., and Kageura, K.: Bilingual terminology mining-using brain, not brawn comparable corpora. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, pp. 664–671(2007)
13. Och, F.J., Ney, H.: A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics, 29(1):19–51(2003)
14. Prochasson, E., Fung, P.: Rare word translation extraction from aligned comparable documents. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA, pp. 1327–1335 (2011)
15. Robitaille, X., Sasaki, Y., Tonoike, M., Sato, S., Utsuro, T.: Compiling French-Japanese terminologies from the web. In: Proceedings of the 11th conference of the European chapter of the association for computational linguistics (EACL'06). Trento, Italy, pp. 225–232 (2006)
16. Shao, L., Ng, H.: Mining New Word Translations from Comparable Corpora. In: Proceedings of the 20th ACL Conference, pp. 618 (2004)
17. Stanley, F. Chen: Aligning sentences in bilingual corpora using lexical information. In: Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, Columbus, Ohio, pp. 9–16 (1993)
18. Talvensaari, T., Laurikkala, J., Järvelin, K., Juhola, M.: Creating and exploiting a comparable corpus in cross-language information retrieval. TOIS 25(4) (2007)
19. Yu, K., Tsujii, J.: Extracting bilingual dictionary from comparable corpora with dependency heterogeneity. In: Proceedings of HLT-NAACL, Boulder, Colorado, pp. 121–124(2009)