Mining Parallel Documents across Web Sites

Pham Ngoc Khanh and Ho Tu Bao

Japan Advanced Institute of Science and Technology,
923-1292 Japan, Ishikawa, Nomi, Asahidai 1-1
{khanh,bao}@jaist.ac.jp
http://www.jaist.ac.jp

Abstract. Most methods on building parallel corpora often start from large scale bilingual websites that are not always an available resource for many language pairs. In this paper we present a novel method to mine parallel documents between English and other non-popular languages which are situated on different locations on the Internet. Our method is motivated by the observation that many non-popular language news are translated from popular English news websites. Given a news in a non-popular language, a method is proposed to search for its original English version located on another website using search engines. Experiments with English-Vietnamese show that our method can provide bilingual document pairs in science domain with precision around 90%. Our method is more flexible and scalable than traditional approaches that collect parallel texts from multilingual websites as its starting point is only a set of monolingual news. Furthermore, this method can be applied to mine parallel documents between non-popular languages pairs with scarce resources.

Keywords: Parallel Corpus, Web Information Retrieval.

1 Introduction

Parallel corpora are sets of texts in one language together with their translation in another language. Parallel corpora are indispensable resources for many areas of text and web mining research including statistical machine translation [1], cross-lingual information retrieval [2], automatic bilingual lexical building [3].

While manually complied parallel corpora is costly and time consuming, methods to automatically build the parallel corpora have gained substantial attention in the field of NLP. There are a number of good Web mining systems such as STRAND [4], BITS [5] and PTMiner [6] that automatically acquire parallel documents from bilingual web sites. Those systems first crawl bilingual web sites exhaustedly to collect two sets of web pages in different languages. Then bilingual dictionary and HTML structure information are used to align translation equivalent web pages. Beginning with two sets of large scale monolingual documents, cross-lingual information retrieval models can also be applied to search for noisy parallel documents [7,8]. Using search engine to search for parallel texts is introduced recently in the work of Achim Ruopp and Fei Xia [9].

P.-J. Cheng et al. (Eds.): AIRS 2010, LNCS 6458, pp. 552–563, 2010.

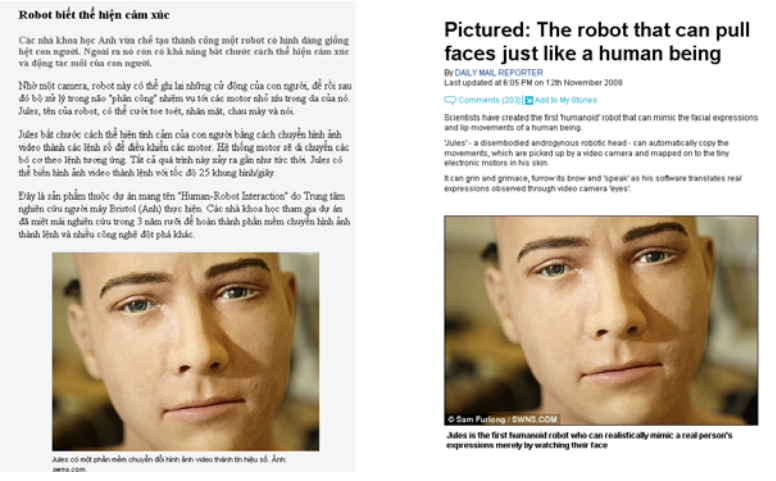
With the help of above-mentioned methods, many parallel corpora with various language pairs [10,11,12] have been built and publicly available for researchers. Unfortunately, parallel corpora are only available for a limited number of popular language pairs like English-French, English-Japanese. For many non popular language pairs like English-Vietnamese, English-Thai such resources are not available for researchers in these countries to conduct their research.

Moreover, the task to build parallel corpora for popular language pairs is much easier than that for non-popular language pairs. Popular language pairs have abundance of online texts database that were well-organized and thus is easy to collect. For instance, in Europarl project[1], authors could easily collect more than one million parallel sentence pairs by just crawling the European parliament website which provides law documents and discussions in 11 European languages[2]. Documents in these websites are well-organized. For example, in Europarl website, documents referring to the same content have the same name and stored in different directories in accordance with their languages.

These above facts are not true for non-popular language pairs. Resources to build parallel texts for those language pairs are very scarce and building parallel corpora for such language pairs is a really tough challenge for NLP researchers. To our knowledge, research conducted to build bilingual resources for non-popular language pairs until now just focused on only word-level. Sources for extracting bilingual lexicons are either from comparable corpora [13] or via a third pivot language [14].

In this paper, we propose a new method that can be applied to mine parallel documents located on different web sites on the Internet, which can be applied to collect parallel texts for resource scarce language pairs. Our method is based on the observation that many news of non popular languages in domains such as outside world, technology, science, health are collected and translated from other popular English news' websites on the Internet like BBC, VOA and so on. An illustrated example was shown in Figure 1 between a Vietnamese news[3] posted in `http://www.vnexpress.net` website and an English news[4] posted in `http://www.dailymail.co.uk` website. The source websites are cited within the content of translated news due to copyright law. We also noted that between two translated documents there exists contents that do not change during translation process. We defined those data as *Translation Independent Data (TID)*. Given a source news in non popular language, we formulated queries based on TIDs, posted date and news source. By sending those queries to search engines, we obtain some English candidates. Then a document matching process which combines both lexical and length features is implemented to find out the target English news.

[1] `http://www.statmt.org/europarl/`
[2] `http://www.europarl.europa.eu/`
[3] `http://www.vnexpress.net/GL/Khoa-hoc/2008/11/3BA086BE/`
[4] `http://www.dailymail.co.uk/sciencetech/article-1085059/Pictured-The-robot-pull-faces-just-like-human-being.html`



Robot biết thể hiện cảm xúc

Các nhà khoa học Anh vừa chế tạo thành công một robot có hình dáng giống hệt con người. Ngoài ra nó còn có khả năng bắt chước cách thể hiện cảm xúc và động tác mỗi của con người.

Nhờ một camera, robot này có thể ghi lại những cử động của con người, để rồi sau đó bộ xử lý trong não "phần cứng" nhiệm vụ tới các motor nhỏ xíu trong da của nó. Jules, tên của robot, có thể cười toe toét, nhăn mặt, chau mày và nói.

Jules bắt chước cách thể hiện tình cảm của con người bằng cách chuyển hình ảnh video thành các lệnh số để điều khiển các motor. Hệ thống motor sẽ đi chuyển các bộ cơ theo lệnh tương ứng. Tất cả quá trình này xảy ra gần như tức thời. Jules có thể biến hình ảnh video thành lệnh với tốc độ 25 khung hình/giây.

Đây là sản phẩm thuộc dự án mang tên "Human-Robot Interaction" do Trung tâm nghiên cứu người máy Bristol (Anh) thực hiện. Các nhà khoa học tham gia dự án đã miệt mài nghiên cứu trong 3 năm rưỡi để hoàn thành phần mềm chuyển hình ảnh thành lệnh và nhiều công nghệ đột phá khác.

Jules có một phần mềm chuyển đổi hình ảnh video thành tín hiệu số. Ảnh: swns.com.

Pictured: The robot that can pull faces just like a human being

By DAILY MAIL REPORTER
Last updated at 6:05 PM on 12th November 2008

Comments (203) | Add to My Stories

Scientists have created the first 'humanoid' robot that can mimic the facial expressions and lip movements of a human being.

'Jules' - a disembodied androgynous robotic head - can automatically copy the movements, which are picked up by a video camera and mapped on to the tiny electronic motors in his skin.

It can grin and grimace, furrow its brow and 'speak' as his software translates real expressions observed through video camera 'eyes'.

Jules is the first humanoid robot who can realistically mimic a real person's expressions merely by watching their face

Fig. 1. An example of English-Vietnamese cross-website parallel documents

The main contribution of this work includes the followings. First, we identify new clues from the Internet to extract parallel documents. Second, a new method have been proposed to build a set of parallel documents from those clues.

The rest of the paper is organized as follows. In the next section, we introduce our method in details. The experiment and discussion on results are presented in Section 3, 4. In Section 5, we conclude our work and give some directions for future research

Table 1. Steps in the proposed framework

Steps	Description
Clue extraction	Extract clues from t_S that guide the process to find t_E.
Query generation and ranking	Use extracted clues to generate a set of queries combining extracted information with several search constraints. After that queries are sent to search engines with order from high rank to the low rank until we get candidates. The rank of a query is defined based on the constraints put on this query. The more constraints put on a query, the higher the rank it is evaluated.
Filtering	Look for the original English news t_E from the list of candidates. A method that combines both length-based and TID-based features is proposed.

2 Proposed Method

Given a potential news t_S in non popular and resource scarce language S that is likely to be translated from an English news, our task is to determine its original English version t_E on the Web. To determine whether t_S is a potential news or

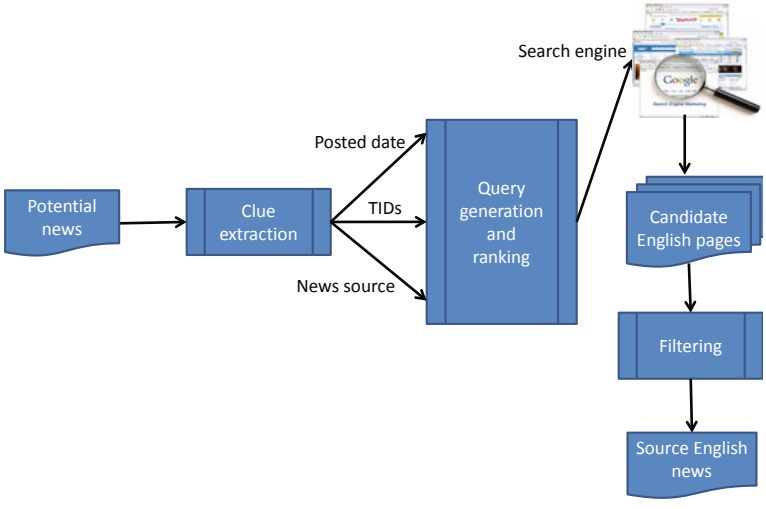


Fig. 2. Framework to find original English news of a potential translated news in non popular language

not, we check its citation. If t_S is cited to an English news website, it is regarded as potential news. Otherwise, t_S is not a potential news. Given a potential news in language S, we propose a framework with 3 main steps to discover its original English news as described in Table 1 and Fig. 2.

2.1 Clues Extraction

Given a potential news t_S, we extract three following clues:

- *Posted date*: the date that the news was put online and available for Internet users to access.
- *URL of news source*: To get the URL of news source, we first extract the name of the cited source which if often put in the bottom of the news using regular expression technique. We then will create a query by appending the word *"news"* to the name and send its to Google search engine. We use language restriction feature (*lr* parameter) to search for only pages in English. The domain name of first URL returned by Google will be chosen as the URL of news website.
- *Translational independent data (TID)*: TID is defined as data that is unchanged during translation process. TIDs can be viewed as common texts in the writing system between two languages. For example, given an English sentence “60 people were killed in Republic of Congo Train Crash" and its equivalent Vietnamese translation “60 người chết trong tai nạn xe lửa ở CHDC Congo", 60 and Congo are considered as numeric and textual TIDs

respectively. TIDs that are close together will be grouped into phrases. The task to extract TIDs can be regarded as a sequence labeling problem and can be solved using well-known techniques such as HMM, CRF. Nevertheless, this approach suffers from a major drawback that we need to build a training data set manually, which is a time-consuming process. We propose two simple unsupervised methods to extract TIDs from the content of potential news t_S.

- The first method is dictionary-based approach. Assuming we have a monolingual dictionary containing all of words that are originated in S. All words that does not appear in this dictionary will be considered as an TID word. We assume that texts that are not originated from language S are kept the same as their appearance in English but not be transliterated into writing system of language S.
- The second method is machine translation-based. This approach comes from the nature of TID. Texts in t_S are translated into English using a translation system. Tokens and phrases that appears in both t_S and its English translated news will be considered as TIDs.

2.2 Query Generation and Ranking

Next step in our method is to create a set of queries from data extracted from previous step. We generate not a single query but a bunch of queries with the different levels of search restriction. Intuitively, the more restriction we set on the query, the smaller number and more accurate the candidates we can get. On the other hand, if we put too strict restriction, it is likely that search engines return no candidates. For those reasons, instead of one query, we generate a set of queries, rank the queries from high to low orders and send them to the search engine until we get some candidates.

We use only *text* TIDs to generate queries. Numeric TIDs are not included in this step. Regarding to search constraints, we define several kinds of search constraints as follows:

- *Phrase constraint*: Only web pages that contain extract phrase in the query are retrieved.
- *Date constraints*: Only web pages within a range of date will be returned.
- *Site constraint*: Only web pages belonging to a specific web site are returned.
- *Language constraint*: As we want to search for source articles written in English, the language of retrieved pages should be in English.
- *n-gram constraint*: With n-gram constraint, the query will be built from TIDs that have at least n tokens.

After being created, queries are ranked and sent to the search engine from high to low orders. There are many methods proposed to rank words and phrases in a document based on their frequency and co-occurrence statistics [15] that can be applied to rank the queries. In this work, we propose a much more simpler heuristic approach that ranks the query based on the types of constraints put into

the query. We apply language, date constraint and site constraint to all queries. For the remaining constraints including phrase, site and n-gram constraints, we will define their corresponding rank in the following ways:

- The query with more constraints will be rank higher than query with less constraints.
- The rank of site constraint is highest while the rank of the phrase constraint is lowest. That is

$$\text{rank(n-gram constraint)} > \text{rank(phrase constraint)}$$

- Within n-gram constraints, the n-gram with lower n will be assigned higher rank as n-gram constraint with lower n provides query with more specific information .

$$\text{rank(1-gram constraint)} > \text{rank(2-gram constraint)} > \cdots > \text{rank(N-gram)}$$

Queries that contains too few and short TIDs are not dispatched to the search engine. Only query created from at least two words and the total length of words larger than a threshold of, says 10 characters, is qualified.

2.3 Filtering

In this step, after getting a list of candidate web pages returned by Google, we have to select among them the original English news t_E of the given potential news t_S.

First, web pages are preprocessed. Noisy texts in web pages like boilerplates, advertisement and HTML scripts are removed because they are proved to have bad impacts on many text processing tasks. Cleaning web page is a challenging problem. There are many works that employ sophisticated features set and advanced supervised machine learning techniques such as SVM, CRF . In this paper we choose to apply method developed by Jing Li and C. I. Ezeife [16]. This work applied n-gram language models to distinguish between main texts and noisy texts which produces comparable performance in shared task competition of 3rd Web as Corpus workshop in summer 2007 and has been released as the open source project NCleaner[5].

After web pages are preprocessed, we implement a process to determine whether the original English news is among the list of N-best candidates. In common approaches, bilingual lexicons are often used to measure the similarity between two documents in two languages. In this research we develop a simple approach by incorporating length and TID features together. The length-based filtering is based on assumption that a long text should have a long translation and vice versa. The length-based constraint is defined as follows

$$\frac{1}{3} \times \text{length}(t_S) < \text{length(EN-text)} < 3 \times \text{length}(t_S)$$

[5] `sourceforge.net/projects/webascorpus/files/NCleaner/NCleaner-1.0/`

while TID-based matching score is computed by the following formula

$$matchingScore = \frac{\text{Number of TIDs found within English text}}{\text{Total number of TIDs in } t_S}$$

The best matched English document is one that satisfies length-based constraint and have the highest TID-based matching score higher than a pre-defined threshold δ. In our experiment, we choose δ as 0.33 empirically.

3 Evaluation

The main objective of this section is to evaluate effectiveness of our proposed method. The language pair used in our experiment is English-Vietnamese. The finally obtained English documents were checked manually as we do not have any gold data sets for this problem.

3.1 Evaluation Measures

In order to evaluate how good our proposed approach is, we use popular measures: recall, precision and f-measures. In our research, those measures are shortly described as below

$$Precision = \frac{\text{Number of correct English news}}{\text{Number of English news found by our approach}}$$

$$Recall = \frac{\text{Number of correct English news}}{\text{Number of Vietnamese news}}$$

$$F - Score = \frac{2 \times Recall \times Precision}{(Recall + Precision)}$$

3.2 Experiment Setup

To conduct the experiments for parallel documents searching, Vietnamese articles are collected from VNExpress web site[6], which is a very popular news website in Vietnam. We focused on two domains i.e. world and science. In our observation, many articles from those two domains are likely to be gathered and translated from other English web sites. We collected articles in science domain that was posted within the first 4 months of 2009 and articles in world domain in posted within April 2009. We extract only main texts from those news using Java-based HTML Parser[7] library. We compile a list of URLs of more than 20 Vietnamese news agency that are often cited by other news website. We selected only Vietnamese news that is cited to an website that is not in the above list. Totally, we have 211 science articles and 108 world articles for our experiments.

[6] `http://www.vnexpress.net`
[7] `http://htmlparser.sourceforge.net/`

The dictionary for Vietnamese are taken from training corpus for word segmentation task compiled by Lê Hồng Phương et al, a Vietnamese group doing research on word segmentation. This Vietnamese dictionary data was provided together with their open source Vietnamese word segmentation tools namely vn-tokenizer[8]. However, we do not need all data in the dictionary. We only need a list of Vietnamese words that are composed from English alphabet. Words that contain special Vietnamese characters such as 'á', 'â', 'ê', 'ớ' are considered to be Vietnamese and are filtered out from the original dictionary. We then revised the list by hand and finally get a list of our Vietnamese words contains only 420 words.

To translate texts from Vietnamese into English, we use Google translation system.

3.3 Experiment Results and Discussion

We conducted three experiments. The first experiments is to evaluate the effectiveness of the proposed method and to compare the performance in two cases: dictionary-based and machine translation-based TID extractions. In this experiment, we set the threshold δ for document matching score is 0.33 and only top 8 web pages are examined. Due to the different nature of world and science news (world news need to be updated at a short time while science news is not required to be delivered to audience in a short time), we set the date range for them as 90 and 1 respectively. Experiment results are shown in Table 2. In this table, DIC and MT are represented for dictionary-based and machine-translation-based methods respectively. Experimental results show that the machine translation based TID extraction has better performance in comparison with dictionary-based method. However, the difference is not too big. With language pairs that a machine translation is not available, simple dictionary-based approach can be a good alternative choice.

Table 2. Performance with world and science data

Settings	Docs	Found	True	Precision	Recall	F-Score
Science						
MT	211	136	124	**91.18%**	58.77%	71.47%
DIC	211	129	117	**90.76%**	55.45%	68.82%
World						
MT	108	77	55	71.43%	50.93%	59.46%
DIC	108	67	43	64.17%	39.81%	49.14%

The experimental results show that our method has promising performance with precision approximately 90% on science news while that with world news is around 70%. This indicates that TIDs extracted from science news has more discriminative power than TIDs extracted from world news. In a world news,

[8] `http://www.loria.fr/~lehong/tools/vnTokenizer.php`

many of the TIDs extracted are popular names and have their appearance in many other news. News in world domain can have many developments, which are different phases of a same story. Moreover, a Vietnamese news in world domain is often translated and integrated not only from one news source but multiple news agencies. This fact makes the task of finding the original news in English become much more difficult for world domain data.

We also do further investigation and figure out some reasons why English news could not be found as follows: (1) The English news are not posted within the date range we set. This often happens with Vietnamese news in science domain which has its original English news posted later than 90 days. (2) Wrong typings: There are terms and named entities that are mis-spelled. As a result, our method could not find the right original English news. (3) News that are integrated from many news sources. In world domain, there are many news that are integrated from not only one news website but some news websites to provide different developments of a hot news to readers. For those news, the search engine often returns no candidates. (4) The news do not contain text TIDs. There is a fact that not all data has TIDs within its contents. For such kind of data, our method fails to gather enough and reliable clues to construct the queries and consequently fails to find the original English news.

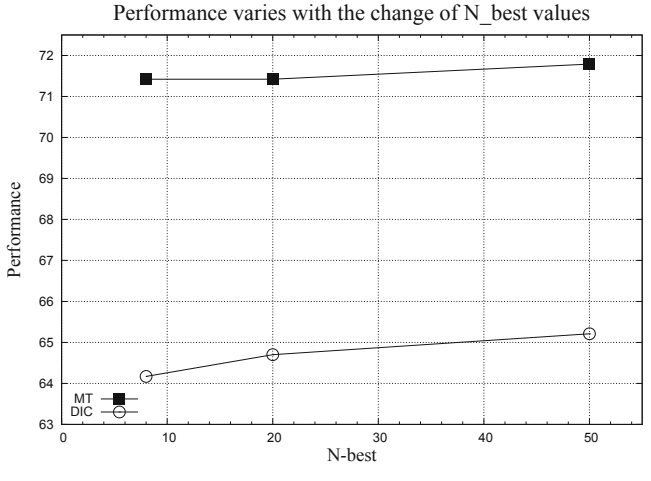


Fig. 3. Performance with different values of N

Moreover, we also implemented the other two experiments with an aim to see how performance of our approach varies when we change the value of matching threshold parameter δ and the number of documents N need to be checked. We did experiments with various values of δ and N. In these experiments, we did the test on world data only. We can see from Fig. 4 that δ having values around 0.375-0.4 produced the highest F-score while still achieving high precision. We also noted that increase the value δ does not always lead to the increase of precision. This may be due to the fact that the TID-based matching score is

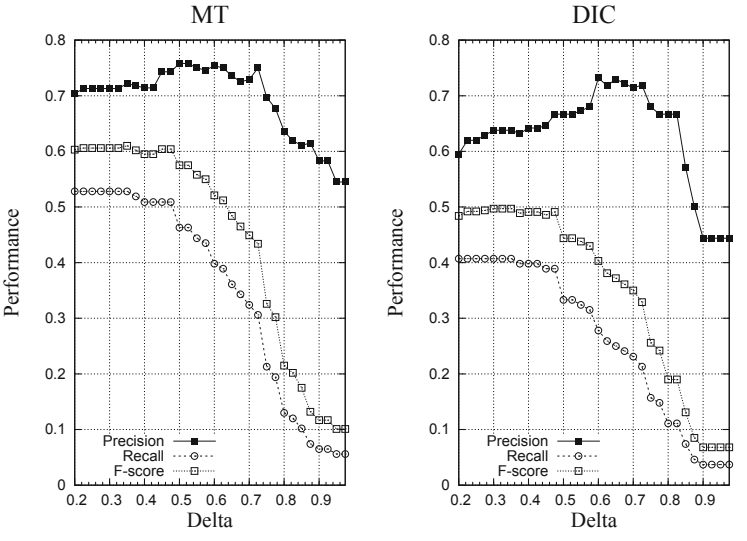


Fig. 4. Performance varies with different values of δ

not strong enough to distinguish between relevant English news and the original English news as their contents often contain the same set of TIDs. This weakness need to be overcome in future research. Meanwhile, increasing the number of examined candidates N from 8 to 20 or 50 does not lead to much improvements of precision as seen in Fig. 3. This suggests that the best candidates can be found in the top 8 candidates. We thus need to consider only first 8 pages returned by the search engine, which helps to reduce a lot of processing time.

4 Discussion

Our method is more flexible than previous works that relied only on collecting data from large scale bilingual websites [4,6,17] or collection of domain data [18]. Our starting point is monolingual data that is abundant and straightforward to collect.

Munteanu et al. [8] built a system to extract parallel sentences from two sets of millions monolingual documents that refer to some common topics. However, their focus is recall rather than precision. Besides, their approach suffer from high computational cost that requires an expensive clustering system to implement. Whereas, this proposed method has inexpensive computational cost by using Google's powerful system to conduct the search process.

The work closely related to this study is that of Achim Ruopp and Fei Xia [9]. This work also use commercial search engines to find translated document pairs. Author applies query expansion techniques to form sampling queries in both

languages using a set of bilingual lexicons. However, their approach suffers a very large search spaces because for each query pair there are substantial amount of web pages returned by search engines in both languages. In this work, as we only search for documents in target languages with time and domain restriction, the search space is greatly reduced.

The proposed method allows people to incrementally mine parallel documents which is not easy to implement with approaches starting from a bilingual source [4,5]. One can build a system to collect updates from news websites using XML-based RSS feed technology. Whenever there are new contents available, they will be input into monolingual news database. Accordingly, the system will search for new pairs of parallel documents and add them to current database. Therefore, it is possible to incrementally upgrade the scale of the corpus.

5 Conclusion and Future Work

We have presented a novel and flexible way to discover parallel documents from the Internet. Different from other approaches that often focus on only one large scale bilingual website or collection, our method can identify parallel documents that are located separately on two different websites using available search engines. Clues to identify those parallel pairs are TID data, date information and the news source cited within the content of source news. Based on those clues, we propose a method to generate and rank the queries, which is based on search restrictions put on them. We develop a parallel document matching algorithm that combines both TID and length features. Our method performs well with data in science domain (precision around 90%) while the precision for data in world domain is still not high but acceptable (around 70%). This method can be applied to collect parallel resources for non-popular languages.

This work is just our initial investigation and there are still room for improvements. Our future works will explore the following directions: (1) Develop a new method to extract and rank TIDs more efficiently. (2) Develop a new method to find source English news corresponding to potential news that does not share any TIDs. (3) Develop a new method to extract the parallel sentences or sub-sentences from obtained parallel documents.

Acknowledgments

This work is supported by KC.01.01.05/06-10 national project. The first author has been supported by Japanese Government Scholarship (Monbukagakusho) to study in Japan. We also want to thank the four anonymous reviewers for their invaluable comments.

References

1. Brown, P.F., Cocke, J., Pietra, S.A.D., Pietra, V.J.D., Jelinek, F., Lafferty, J.D., Mercer, R.L., Roossin, P.S.: A statistical approach to machine translation. Computational Linguistics 16(2), 79–85 (1990)

2. Gey, F.C., Kando, N., Peters, C.: Cross-language information retrieval: the way ahead. Inf. Process. Manage 41(3), 415–431 (2005)
3. Kumano, A. and Hirakawa, H.: Building an MT dictionary from parallel texts based on linguistic and statistical information. In: Proceedings of the 15th conference on Computational Linguistics, pp. 76–81 (1994).
4. Philip, R., Smith, N.A.: The web as a parallel corpus. Computational Linguistics 29(3), 349–380 (2003)
5. Ma, X., Liberman, D.Y.: BITS: A method for bilingual text search over the web. In: Proceedings of Machine Translation Summit VII, pp. 538–542 (1999)
6. Chen, J., Nie, J.Y.: Automatic construction of parallel English-Chinese corpus for cross-language information retrieval. In: Proceedings of the Sixth Conference on Applied Natural Language Processing, pp. 21–28 (2000)
7. Colleier, N., Hirakawa, H., Kumano, A.: Creating a noisy parallel corpus from newswire articles using cross-language information retrieval. Transactions of Information Procession Society of Japan 40(1), 351–361 (1999)
8. Munteanu, D.S., Marcu, D.: Improving machine translation performance by exploiting non-parallel corpora. Computational Linguistics 31(4), 477–504 (2005)
9. Ruopp, A., Xia, F.: Finding parallel texts on the web using cross-language information retrieval. In: Proceedings of the Second International Workshop On Cross Lingual Information Access Addressing the Information Need of Multilingual Societies, pp. 18–25 (2008)
10. Jorg, T., Nygaard, L.: The OPUS corpus - parallel and free. In: Proceedings of LREC 2004, pp. 1183–1186 (2004)
11. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: Proceedings of MT Summit, pp. 79–86 (2005)
12. Ralf, S., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., Varga, D.: The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In: Proceedings of the 5th International Conference on Language Resources and Evaluation, pp. 2142–2147 (2006)
13. Yu, K., Tsujii, J.: Extracting bilingual dictionary from comparable corpora with dependency heterogeneity. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, pp. 121–124 (2009)
14. István, V., Shoichi, Y.: Bilingual dictionary generation for low-resourced language pairs. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pp. 862–870 (2009)
15. Matsuo, Y., Ishizuka, M.: Keyword extraction from a single document using word co-occurrence statistical information. International Journal on Artificial Intelligence Tools 13, 157–170 (2004)
16. Li, J., Ezeife, C.I.: Cleaning web pages for effective web content mining. In: Bressan, S., Küng, J., Wagner, R. (eds.) DEXA 2006. LNCS, vol. 4080, pp. 560–571. Springer, Heidelberg (2006)
17. Zhang, Y., Wu, K., Gao, J., Vines, P.: Automatic acquisition of Chinese-English parallel corpus from the web. In: Proceedings of the 28th European Conference on Information Retrieval, pp. 420–431 (2006)
18. Utiyama, M., Isahara, H.: A Japanese-English patent parallel corpus. In: Proceedings of Machine Translation Summit XI, pp. 475–482 (2007)