웹 번역문서 판별과 병렬 말뭉치 구축

김지형[O] 이일병
연세대학교 컴퓨터과학과
{evankim[O], yblee}@csai.yonsei.ac.kr

Judging Translated Web Document & Constructing Bilingual Corpus

Jee-hyung Kim[0] , Yill-byung Lee
Dept. of Computer Science, Yonsei University

Abstract

People frequently feel the need of a general searching tool that frees from language barrier when they find information through the internet. Therefore, it is necessary to have a multilingual parallel corpus to search with a word that includes a search keyword and has a corresponding word in another language. Multilingual parallel corpus can be built and reused effectively through the several processes which are judgment of the web documents, sentence alignment and word alignment. To build a multilingual parallel corpus, multi-lingual dictionary should be constructed in each language and HTML should be simplified. And by understanding the meaning and the statistics of document structure, judgment on translated web documents will be made and the searched web pages will be aligned in sentence unit.

1. Introduction

This paper experiments with two languages, Korean and English, and makes them possible to apply to any language by building a word dictionary. After abstracting Korean web documents and translated English web documents for the candidate documents from the web, attribute part of meaningless HTML tag is eliminated through a simplification process. Then footnotes, scripts and other parts that are not related to the meaning are removed and then HTML document is reconstructed by classifying tags having meaning. Separator abstracts segments from each tag and applies statistical methods. Compare HTML document structure of two multilingual documents that are built by those processes above. And by using multilingual dictionary, search the words having similar meaning in a ratio of 1:1 or 1:N (polysemy), and find if they have a similar document structure and meaning and judge whether or not the documents are translated. Then the building method of bilingual corpora will be presented by using a sentence alignment.

2. Translated Document on the Web

Web documents have a fixed structure because of using HTML tags. The translated document in one language has similar structure with its original copy; therefore, the tags that decide the documents structure appear mostly in the same order so it makes possible to understand the parallelism of translated web documents. To acquire needed information by using bilingual (parallel) corpora, process of aligning corresponding things between two languages is needed as a precedent step of simplification. This process is called alignment and the alignment is divided into a paragraph alignment, sentence alignment and word alignment according to a corresponding unit.

Study on alignment has carried in various ways such as sentence alignment [1, 3, 6, Wu94], phrase alignment [8, 7], and word alignment [2,4,5]. Generally two languages having the same linguistic family such as English and French are easy to align for their similarity in their words and sentence structures. However, aligning two languages having different linguistic family such as Korean and English is more difficult than the languages having similar linguistic family. Therefore, other approaching method is required. By using Korean-English parallel(bilingual) corpus built by analyzing web document structure, this paper presents a sentence alignment method that applies the frequency of content word use. The frequency of content word use is similar between two languages. This characteristic is found among the languages that have different linguistic family, and the content words have linear correlation. By using the frequency of content word use, this paper presents the sentence alignment among many languages that have different linguistic family.

3. Building bilingual corpus and sentence alignment from the web document

3.1 Building a multilingual dictionary

Korean language has a prepositional word '*Josa*' that helps connect the meaning of sentences and it is classified as a morpheme. However, in English, such prepositional word is not classified as a morpheme but a word is the smallest unit. Therefore, this *Josa* in Korean or a preposition and article in English should be eliminated to compare two languages in the process of word analysis.

Josa
Case ① Subjective ② Nominative ③ Objective ④ Complementary ⑤ Adnominal ⑥ Vocative ⑦ Adverbial
Conjunctional
Supplement

Construction of multilingual dictionary is as below;

Word	Part of speech	Foreign language Pointer
학교 (school)	noun(N)	0
Similar word order		
-1		

Foreign language Pointer is a memory address pointing out a translated word. Synonyms of Korean and English compose a map. One word divides polysemies with tentative blanks such as (), {},, ,:,= ,-. However, note that just comparing words is more effective although separating words sometimes classify words to replace blanks. If there *are no matched words in compared documents, find them* in the multilingual dictionary with the first word. If there' s no words again, find similar words and save the frequency of match. When searching a word, classify content word and function word and eliminate non-existing parts of speech in each language. For example, English prepositional word " of" always accompanies with space of the word in multilingual dictionary, however, it is corresponding to Korean word " *Ui(의)*" therefore it doesn' t require space of the words in Korean. That is to say, it should be treated as a function word.

3.2 Abstracting candidate pair of translated web documents

Firstly, we got samples from the web site providing search engine to search translated web documents. Except for the translated part for machines, most of them were introduction of the company. We followed the links manually to take samples that were translated by human beings only and saved HTML source. And we named " korean.html" ," english.html each and saved other totally different web sites to compare.

3.3 HTML Simplification

Simplify the web documents that will be compared. Actual translated web document contains more content than its corresponding translated document so that there can be a part that are not corresponding to each other in the document. And the contents of the document are the same but the document structure can be different by using different HTML tags although they are translated web documents. Moreover, the number of tag in the pages that are composed of the same content can be not corresponding.

Figure 1. English web document

```
<font class="font14link"><img src=/img/1.gif>&nbsp;Seoul
Metropolitan Ministry of Education</font>
```

In an actual translated document, location of the image and size and the link can be different. In addition, Javascript and footnotes can appear in the middle and the size or the color of words can be expressed differently. These can disturb analyzing sentence structure. Therefore, following is the result of simplification by eliminating irrelevant parts to the content.

Figure 2. Preprocessed English web document

```
<font><img>Seoul Metropolitan Ministry of Education</font>
```

Figure 2 simplifies HTML tags. To improve accuracy, update tools for judging functional tags and important tags by using experiential statistics. Now character of each tag, footnotes, function, script, and so forth are eliminated.

3.4 Searching for word distinction and content word

To judge words, sentences composed of only one language are hardly used but using mathematical signs or mixing with other languages are frequently preferred.

Figure 3 Mixed sentence with language and sign

```
log1010 = 1 , log525 => log5(5^2) => 2 log2(2n) 혹은
3log10(10n+3)+66
for ( b=1 ; b<2n+20 ; b++ ) {
  printf("%d %dn",a,b); }
```

The mathematical signs and functions are found in Figure 3. Eliminate blank words such as mathematical formula and function before dividing into words and check whether or not the sentence corresponds with the whole sentence unit. If it does not correspond, put the word distinction as a first priority and then divide it into letters.

Korean sentence is n while foreign sentence is one (n:1)
Korean sentence is one while foreign sentence is n (1:n)
Korean sentence is m while foreign sentence is m (m:n)

1:0 and 0:1 are frequently found when the *sentence order is changed or some other sentence is* inserted so that makes unmatched. n:1, 1:n, and m:n are found when sentence order is linguistically different or when a different sentence form causes division into several sentences.

For example, take a look 1:n alignment as follows;

1.
나는 어제 학교에 갔지만 그를 찾을 수 없었다.
2.
I went to the school in yesterday.
3.
But I couldn' t find him.

Word location in the sentence cannot be changed. However, Korean sentence 1 has 8 content words and sentence 2 has 4 content words excluding preposition, article and conjunction. Sentence 3 has 4 content words. Therefore the number of content words in Korean sentence 1 becomes similar when the content words in sentence 2 and 3 are added. Then by analyzing meanings, check if the number of content word and their actual meaning are corresponded to each other. Sentence alignment is proportioned to the number of content word regardless to the number of sentence when the total number of Korean content word is n and the total number of English content word is m. Each language will be characterized according *to the statistics of distance by content word order. When* through the simplification process, however, the searching words, eliminate prefix and suffix in a morpheme unit to improve probability of search in a word dictionary. Search algorism will be presented in detail in this paper when aligning sentences or searching words.

3.5 Comparison of document structure by statistical methods

Examine whether or not two documents are translated while comparing the number of tags and content words and their order through the whole document. Save content words and their number while counting the number of content words and there will be some words repeatedly showing up. Both the content word that shows up once and the duplicated words have 1 weight equally. Because Korean frequently omits a subject otherwise English uses subject very often. Document structure and the relationship in meaning are primarily judged by these three factors.

3.6 Sentence alignment and searching multilingual dictionary

Align sentences that their meanings are found in the process of searching words. Actual order of sentence can be changed and the sentence can be 1:1 and 1:N correspondence. However, this paper only considers following cases;

One Korean sentence but no corresponding sentences in foreign language (1:0)

No Korean sentence but one sentence in foreign language (0:1)

Too broad so the multilingual dictionary for Korean and English builds with only sample documents

3.7 Building Corpora

Corpora can be built by the methods discussed above and the alignment of content words appeared in the building process helps find appropriate and the most frequently used words among polysemies so that it will be useful when translating later.

Word corpus is built as the word is in the sentence containing morphemes.

4. Conclusion

We abstracted and used various Korean and English web documents for a candidate pair of translated documents and the documents are totally different from the original translated document on the web. Accuracy of original Korean document is low to compare to other web documents in foreign language because Korean document contains many irrelevant items such as various Javascript, windows, banners and so forth. The web document in native language is usually more complicated in its structure than the translated document in foreign languages.

Document complication = the number of no meaning tag/total number of tag

Rate of function word = the number of function word/total number of tag

Rate of translating = the number of matched content word/total number of content word

Generally the complication rate of translated document is 5% while native language is more than 10%. Sometimes inaccurate tag grammar is found in the simplification process. We eliminated actual position and matching size of the image and word color and its size to correspond rate decreased because there' s no information to judge when tags didn' t appear in the same place of the document but appeared in another line. When there is a sentence containing words that are found in tags, abstract it and add on a sentence list.

Sentence number is a line number in the document of HTML simplification. And when comparing to the translated document, sentence alignment confirmation is a flag that shows whether or not matched sentences are found in the multilingual dictionary. Corresponding sentence is a matched sentence line number that is found in the translated document. Analysis of the parts of speech is also important and it is required when actually looking up in a multilingual dictionary.

As a result of examining 100 web documents, judgment rate of translating was about 91%. Among the documents that were judged as a translated document, sentence corpora and word corpora were each 87% and 84% successfully built. Actual vocabulary change in morpheme unit and content decreased search rate, in addition, partial correspondence of the actual content of Korean and English web document were also frequently found and it was another important factor that decreased search rate.

5. References

[1] Brown,P.F., Jenniger C. Lai, and Robert L. Mercer, "Aligning Sentences in Parallel Corpora", *In Proceedings of the 29th Annual Meeting of the Association for Computational linguistics*, 1991

[2] Brown, P. F., Stephen A. Della Pietra, Vincent J. Della Pietra, Robert L. Mercer, "The mathematics of statistical machine translation: paramater estimation", Computational Linguistics, Vol 2, No 19, pp. 263-311,1993

[3] Chen, Stanley F., "Aligning sentences in bilingual corpora using lexical information", *In Proceedings of the 31th Annual Meeting of the Association for Computational linguistics*, pp.9-16, 1993

[4] Dagan, Ido, Kenneth W. Church, and William A. Gale,"Robust bilingual word alignment for machine aided translation", *In Proceedings of the 17th Annual workshop on Very Large Corpora: Academic and Insustrial Perspectives*, pp. 1-8,1993

[5] Fung, Pascale, "A pattern matching method for finding noun and proper noun translation from noisy parallel corpora", *In Proceedings of the 33th Annual Meeting of the Association for Computational linguistics. for Computational Linguistics*, pp. 236-243, 1995

[6] Gale, W. A., and Church, K W., "A Program for Aligning Sentences in Bilingual Corpora", Using Large Copora, MIT Press, 1994

[7] Kitamura, Akira and Hideki Hirakawa, "Automatic exctraction of word sequence correspondences in parallel copora", *In Proceedings of the 4th Annual Workshop on Very Large Coprpora*, pp. 79-87, 1996

[8] Kupiec, Jilian. "An algorithm fo finding noun phrase correspondences in biligual corpora", *In Proceedings of the 31th Annual Meeting of Association for Computational Linguistics*, pp. 17-22, 1993