Acquiring Compound Word Translations Both Automatically and Dynamically

Yujie Zhang
Keihanna Human Info-Communications Research Center,
National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0289, Japan
yujie@nict.go.jp

Hitoshi Isahara
Keihanna Human Info-Communications Research Center,
National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0289, Japan
isahara@nict.go.jp

Abstract

This paper addresses the problem of compound word translation and proposes the approaches to acquiring translations. The proposed approaches focus on exploring web data and utilizing English translations to link words of the source language and the correspondences in the target language. The paper uses Japanese-Chinese language pairs for the sake of illustration and shows initial experimental results. The proposed method is language-independent and therefore can be applied to other language pairs.

1 Introduction

The translation of compound words is an important problem related to the construction of translation dictionaries. On the one hand, the published paper bilingual dictionaries usually collect translations for simple words but not for compound words. On the other hand, new compound words emerge every day. It is reported that most compound words are nominal words. In natural language comprehension, entities or concepts to be recognized are usually described by nouns or compound nouns. Therefore, acquiring translations for compound nouns is extremely important in machine translation and cross-language information retrieval [Cao and Li, 2002; Nakagawa, 2001].

In resolving this problem, we propose a self-learning mechanism that can dynamically and automatically acquire compound words translations from the web and corpora. Here we take Japanese-Chinese language pairs as an instance for illustration and report initial experiment results. The proposed method can also be applied to other language pairs since it is language-independent.

2 Overview of the self-learning mechanism

The self-learning mechanism aims at automatically and dynamically expanding the translation dictionary of compound word. It is implemented in two parts: collecting new compound words (collection part) and then acquiring their translations (acquisition part). In the initialization phase, we design a compound word list for the already known words and extract their composition patterns. When the web data or new corpus data are input into the collection part, the data are matched with the extracted compound word composition patterns. The word strings that matched the patterns are then added to the compound word list if they are not already on the list. When new compound words are added to the list, the acquisition part will be activated to acquire possible translation candidates by using the three approaches described in the next section. The most likely translations will be added to the dictionary. In this way, the translation dictionary of compound words can be expanded automatically and dynamically once new data is available.

In the case of Japanese-Chinese language pairs, we collected compound word list from the EDR Japanese-English dictionary [NICT, 2002] and obtained the composition patterns by using the Japanese

morphological analyzer ChaSen [Matsumoto et al., 2000]. As a result, 178,854 entries were obtained, among which about 75% are nouns and about 14% are verbs. The entries consisting of two words are the most numerous, having 83,779 entries. Most patterns are “noun + noun” for entries of two words. Table 1 lists a part of the composition patterns. Two examples of the obtained compound words are listed below.

Ex.1.

アーバンソシオロジー (urban sociology)
Part-Of-Speech: noun
Composition: ”アーバン(urban) + ソシオロジー (sociology)”
Composition pattern: ”noun + noun”

Ex.2.

アーク電流 (arc current)
Part-Of-Speech: noun
Composition: ”アーク(arc)+ 電流(current)”
Composition pattern: ”noun + noun”

Table 1 Extracted composition patterns within the top 4

Composition Pattern	Count
Noun + Noun	17,481
Noun + Suffix	8089
Noun + Verb	6838
Verb + Verb	6322

3 Approaches to acquiring translations

There are two points underlying our approach: (1) exploring the web [Nagata, et al., 2001] because most new compound words or terminology appear there before they appear in corpora, and (2) utilizing English translations to link the Japanese word and corresponding Chinese translations because there is an abundance of resources for Japanese-English and English-Chinese translation. For a given Japanese compound word, we can try the following approach in turn.

Approach 1

First, get an English translation by looking it up in the Japanese-English dictionary or by searching the Japanese web where the Japanese compound word is followed by an English expression, which might be its English translation. Then use the obtained English translation to search Chinese web to obtain related Chinese texts. Lastly, process the texts to identify the corresponding Chinese translation.

In this approach, English translations are used as intermediates since many domain-specific compound words are usually followed by the corresponding English translations in or without parentheses on Japanese and Chinese websites. Ex. 3 explains the approach.

Ex.3. The English translation of Ex.1, “urban sociology”, is obtained from the EDR Japanese-English dictionary. Through searching the English translation on the Chinese web, we found such text: “中国社会学—名词解释—…都市社会学（urban sociology）社会学的分支学科之一…”. Therefore we obtained the Chinese translation “都市社会学” for the Japanese compounding word “アーバンソシオロジー” .

Approach 2

Compositionally produce Chinese translation candidates by concatenating the Chinese translation of each constituent and then select the one that most frequently appears on the Chinese web. Ex.4 explains the approach.

Ex.4. “アーク電流(arc current)” in Ex.2 has two constituents, “アーク(arc)” and “電流(current)”. The Chinese translations of “アーク” and “電流” are {弧,弧形,弧光,弓形,拱} and {电压}, respectively. The produced candidates are {弧电压,弧形电压,弧光电压,弓形电压,拱电压}. Searching Chinese web returns hits on each candidate as 187, 0, 7, 0, and 0, respectively. The candidate “弧电压”has the largest number of hits and is taken as the most probable translation.

Approach 3

Translate the English translation of the Japanese compound word into Chinese by exploring Japanese-English phrase translations corpora and English-Chinese phrase translations corpora.

Since many English translations in the Japanese-English dictionary and Chinese-English dictionary are phrases or short sentences, we can build a Japanese-English phrase translations corpus and an English-Chinese phrase translations corpus. Then we can use the corpus-based machine translation method to translate the English translations into Chinese. The process consists of two steps. In the first step, we use the English translation to search for the most similar English examples in the English-Chinese corpus and obtain the Chinese correspondence. Because of the problem of data sparseness, it is necessary to generalize some words to variables both for the Japanese-English translation pair and for the English-Chinese translation pairs. In the second step, we replace the generalized variable in the Chinese correspondence with the Chinese translation corresponding to the Japanese constituent.

4 Experiment

We conducted initial experiments using the three approaches described above. The experimental results are reported below.

(1) In the experiment of using **Approach 1**, we randomly selected 100 Japanese compound nouns from the EDR Japanese-English dictionary and search their English translations on Chinese web. The search results showed that nine Japanese compound words obtained the texts containing the corresponding Chinese translations, among which five words do not appear in the Japanese-Chinese Dictionary [Kuraishitake and Orishikise, 2001]. This implies that some Japanese words that do not appear in the published bilingual dictionaries can also obtain their translations by this approach. Two examples are listed below.

Ex.5.

アームズコントロー(arms control) 军备控制
タフトハートレー法(Taft-Hartley Act) 塔夫特-哈特利法案

In this paper, we manually extracted Chinese translations from the texts. How to automatically extract the Chinese translations from the texts remains as a task to be finished further.

(2) In the experiment of using **Approach 2**, we randomly selected 50 Japanese compound nouns from the EDR Japanese-English dictionary that consist of two simple words. For each selected Japanese compound noun, the Chinese translations candidates are compositionally produced and are searched on Chinese web. Then the Chinese candidates are sorted in descending order of the numbers of appearances. The evaluation results showed that 27 Japanese nouns obtained their corresponding Chinese translations within Top 3. Some results are listed in Ex. 6, in which the Chinese translations and the numbers of appearances are listed in brace.

Ex. 6.

アーモンドオイル(almond oil) {杏仁油脂,2}
アイスコーヒー(iced coffee) {冰咖啡,488}{雪糕咖啡,4}{钻石咖啡,3)
アースアンカー(earth anchor) {土锚,235}{地锚,92}{地面锚,13}

アーチ橋(arched bridge) {拱形桥,165}{弧桥,20}{弓形桥,19}
アイソトープ電池(isotope battery) {同位素电池,35} {同位素暴行,0}
アイドル歌謡(idol music) {偶像歌,70}{偶像歌曲,29}{偶像诗,5}

We analyzed the remaining 23 Japanese nouns and found that the produced Chinese candidates did not include correct translations.

(3) In the experiment of using **Approach 3**, we built a Japanese-English phrase translation corpus from the EDR Japanese-English dictionary and an English-Chinese phrase translation corpus from the LDC Chinese-English dictionary [LDC 2002]. We randomly selected 100 Japanese words and searched their English translations in the English-Chinese corpus. Here, we used the following formula to compute the similarity between the English translation (E_1) of the given Japanese compound word and the English example (E_2) in the English-Chinese corpus, in which EditDistance means to use edit distance algorithm [Levenshtein, 1965] and $|\cdot|$ denotes the number of the items of the set. The edit unit is English character.

$$Similarity(E_1, E_2) = 1 - \frac{\text{EditDistance}(E_1, E_2)}{\max(|E_1|, |E_2|)} \qquad (1)$$

We selected the English examples whose values of similarity are larger that 0.5 and ranked their Chinese correspondences in descending order of the values. We then examined the results in the top 3 for each Japanese word and found that 59 Japanese words obtained their Chinese correspondences. Some correct translations are listed in Ex.7, in which the Japanese word and its English translation are listed in the left parenthesis and the English example with the largest similarity value and the corresponding Chinese translations are listed in the right parenthesis.

Ex.7.
(明け暮れする devote *oneself to) (devote one‘s efforts to 致力)
(**足ぶみする** come to a standstill) (be at a standstill 停**滞**)
(いきり立つ fly into a rage) (fly into a rage 暴跳如**雷**)
(打ちおとす to knock down) (knockdown 击倒**的**)

In this paper we did not conduct generalization and replacement. We plan to add generalization and replacement operations in our future work.

5 Conclusion

This paper addressed the problem of compound word translation. The approaches to acquiring translations focused on exploring web data and utilizing English translations to link words of the source language and the correspondences in the target language. We used Japanese-Chinese language pairs for the sake of illustration and showed initial experimental results. The proposed method is language-independent and therefore can be applied to other language pairs.

References

Cao, Yunbo and Li, Hang (2002). Base Noun Phrase Translation Using Web Data and the EM Algorithm. In Proc. of COLING-2002, pp.127-133.

Kuraishitake and Orishikise. (2001). Japanese-Chinese Dictionary. IWANAMI SHOTEN.

LDC(2002). English-to-Chinese Wordlist(version 2.0) http://www.ldc.upenn.edu/Projects/Chinese/.

Levenshtein, V.I. (1965). Binary codes capable of correcting deletions, insertions and reversals, Doklady Akademii Nauk SSSR 163(4), pp.845－848.

Matsumoto,Yuji and Akira, Kitauchi, et al.(2000). Japanese Morphological Analysis System ChaSen version 2.2.1. http://chasen.naist.jp/hiki/ChaSen/.

Nagata, M., Saito, T. and Suzuki, K. (2001) Using the Web as a bilingual dictionary. In Proc. of ACL'2001 DD-MT Workshop.

Nakagawa, Hiroshi (2001). Disambiguation of Single Noun Translation Extracted from Bilingual Comparable Corpora. Terminology, 7:1, pp. 63-83.

NICT (National Institute of Communication Technology) 2002. EDR Electronic Dictionary Version 2.0 Technical Guide.