LE-4

# トピック依存の訳語選択※

# Topic-Dependent Word Selection

パウル ミヒャエル*† 隅田 英一郎* 山本 誠一*†
Michael Paul Eiichiro Sumita Seiichi Yamamoto

## 1. Introduction

A severe problem of machine translation applications is the selection of appropriate word translations in the context of the respective sentences. Depending on the situation in which a sentence is uttered, the translation of a specific source word might vary tremendously. However, conventional translation dictionaries either lack the ability of context-sensitive target selections or require manual adaptation of its general-purpose word translations when shifting to new domains.

In this paper we propose a word selection scheme based on topic-dependent translation dictionaries that are automatically extracted from a bilingual corpus. First, we apply a statistical word alignment method to sentences uttered in the same situation (*topic*) resulting in word translation pairs specific to this topic. In the second step, all topic-dependent word translations of the respective source word are combined with a general translation obtained from the alignment of the complete corpus. The merged translation dictionary is then used for the selection of the most appropriate word translations based on the topic information of the given input sentence.

## 2. Topic Information

The corpus used to test the feasibility of our approach consists of a collection of Japanese (J) utterances and their English (E) translations, which are usually found in phrasebooks for tourists going abroad (Takezawa et al., 2002). The translations were made sentence-by-sentence, resulting in a sentence-aligned corpus. It consists of about 200K bilingual sentences. Moreover, each sentence of the corpus is annotated with a *topic* depending on the situation in which the sentences were uttered. In total, 20 different topics are annotated in this corpus.

We selected the most frequent ones (cf. Table 1) and clustered the respective sentences to topic-specific subsets of the corpus. Each subset exhibits specific characteristics, for example, in respect to its vocabulary that might be exploited to identify topic-specific word translation pairs.

## 3. Topic-Dependent Translation Knowledge

Given a text and its translation into another language, knowledge about translational equivalence can be extracted between expressions that share the same meaning.

However, the meaning of a given expression depends on the context in which it is used and it might change when applied to different situations.

| topic | frequency | vocabulary size | | topic-specific predicates |
|---|---|---|---|---|
| | | J | E | |
| T1: communication | 35.4% | 20755 | 14321 | 慰める,育てる,引退する,… |
| T2: shopping | 12.6% | 6951 | 5171 | 買い取る,品切れする,常緑する,… |
| T3: stay | 11.3% | 5941 | 4583 | 取り次ぐ,短縮する,備わる,… |
| T4: sightseeing | 10.3% | 8610 | 6256 | 開園する,巡航する,登山する,… |
| T5: restaurant | 9.7% | 5292 | 3926 | 飲み込む,加熱する,焼き直す,… |
| T6: contact | 8.3% | 5515 | 4596 | 会談する,欠勤する,送信する,… |
| T7: move | 6.6% | 4214 | 3375 | 遠回りする,乗船する,追突する,… |
| T8: airport | 5.8% | 3880 | 3122 | 出国する,代行する,没収する,… |

Table 1: Topic Characteristics

In our approach we gain knowledge about context-specific translation equivalences by applying a word alignment method only to the text expressions of a specific topic and combine the obtained topic-dependent word translations for each source word. In addition, we extract a general translation from the whole corpus that might be applied when topic-specific information is not available.

### 3.1. Word Alignment

Translation equivalences are extracted from our bilingual corpus utilizing the *competitive linking algorithm* introduced in (Melamed, 2000). This probabilistic method uses a boundary-based model of co-occurrences, which indicates whether a given word pair co-occurs in corresponding regions of the bilingual text and is based on the *one-to-one assumption*, i.e. each word is translated to at most one other word. The algorithm greedily selects the most likely links between co-occurring token pairs and then selects less likely links only if they don't conflict with previous selections.

### 3.2. Topic-Dependent Dictionary

We applied the word alignment algorithm to each topic-specific subset of our corpus and obtained a list of source expressions together with (possibly multiple) translations specific to the given topic (EJ and JE). Moreover, topic-independent translations of each source word are retrieved from the entire corpus, whereby the order of multiple targets depends on the co-occurrence probability obtained by the alignment method. All translations are merged into a topic-dependent dictionary as illustrated in Table 2.

| | | | |
|---|---|---|---|
| *corpus* | 出発する、出る、発つ、忘れる、残す、発車する、置く、預ける、帰る | | |
| *T1* | 出発する、置き忘れる、任せる、置く | *T5* | 任せる |
| *T2* | — | *T6* | 残す |
| *T3* | 発つ、出発する、置き忘れる | *T7* | 出る、出発する、発車する |
| *T4* | 出発する、出る、残る | *T8* | 出発する、出る、置く |

Table 2: Translation Equivalences for (V "leave")

※ This research was supported in part by the Telecommunications Advancement Organization of Japan.

* ATR Spoken Language Translation Laboratories
† Kobe University

In order to get some information about the level of translation ambiguity of the respective entries, we define a *disambiguation score* for each source word as:

$$\text{dscore} = \begin{cases} 0 & , \text{if } |\mathit{target}| = 1 \\ \sum_{\mathit{target}} \frac{1}{|\text{topics}(\mathit{target})|} \Big/ |\mathit{target}| & , \text{otherwise} \end{cases}$$

This measure reflects the average number of topics that share the same translation of the source expression. The larger the score, the more topic-specific translations are defined in the dictionary. Table 3 summarizes the statistics of extracted word pairs for all source entries as well as for nominal (普通名詞,CN) and predicative (本動詞,V) subsets.

| source | Japanese | | | English | | |
|---|---|---|---|---|---|---|
| | all | 普通名詞 | 本動詞 | all | CN | V |
| count | 4373 | 2729 | 714 | 3561 | 1875 | 543 |
| \|*target*\| = 1 | 67.0 | 74.3 | 57.3 | 58.7% | 63.0% | 44.9% |
| \|*target*\| = 2 | 17.4 | 15.6 | 19.3 | 19.4% | 20.9% | 18.3% |
| \|*target*\| > 2 | 15.5 | 10.1 | 23.4 | 21.9% | 16.1% | 36.8% |
| *dscore* avg. | 0.18 | 0.14 | 0.25 | 0.23 | 0.19 | 0.31 |

Table 3: Statistics of Topic-Dependent Dictionary

More source expressions are extracted for Japanese and the percentage of multiple translations indicates a higher translation ambiguity when translating from English to Japanese. Especially for verbal constituents, quite a large number of multiple translations (J: 42.7%, E: 55.1%) are extracted from the corpus. However, the higher *dscore* of the parts-of-speech 本動詞 and V suggest that the incorporation of topic-specific information might help in the word selection task of verbal constituents.

Moreover, the comparison of the number of targets extracted from the entire corpus (general) to those of the topic-specific dictionaries reveals a reduction in the translation ambiguity by an average of 33.8% (J) and 38.5% (E) for 27.8% (J) and 40% (E) of the source predicates. In addition, 12% of the target expressions found in the combination of topic-specific translation equivalences are not contained in the general dictionary, thus increasing the coverage of translation candidates for these source words.

## 4. Word Selection Method

Based on the above analysis of the obtained topic-dependent dictionary, we propose a word selection scheme that chooses topic-specific word translations based on the topic information of the input sentence and uses the general dictionary as a fail-safe strategy, whenever topic-specific information is not available. The baseline method G outputs the target expressions of the general translation dictionary. Method *T* uses the topic information of the test sentence to select the topic-specific translations. The proposed combination *TG* of both methods prefers the topic-specific translation, if such an entry exists, and uses the general target otherwise. In the case that multiple translations candidates are defined in the dictionaries, the first one of the ordered candidate list (cf. marked entries in Figure 2) is selected. If a source word is not covered by the respective method, no translation is available.

For the evaluation of our approach, we used 10,000 utterances of the corpus described in Section 2, that were not used for the creation of the topic-dependent dictionary. Moreover, we used thesauri whose hierarchies are based on the Kadokawa Ruigo-shin-jiten (Ohno and Hamanishi, 1984) for the automatic evaluation of the upper boundary of our approach as described below.

We make use of an automatic evaluation scheme that applies a sentence-based word alignment method to the bilingual input in order to extract the target expressions considered to be the correct ones for the evaluation. The recall of this method is 67.4% and the precision is 87.2% (Imamura, 2001). Therefore, the automatic evaluation is carried out using around 70% of the test data accepting the side-effect of around 10% of noisy data. We calculate (1) a lower and (2) upper boundary for the accuracy of our method based on the agreement between (1) the word and (2) its semantic meaning of the "correct" translation and the output of the three word selection methods. The results of our evaluation are summarized in Table 4.

| source | method | coverage | accuracy | |
|---|---|---|---|---|
| | | | lower boundary | upper boundary |
| 本動詞 | *G* | 95.1 | 47.1 | 65.3 |
| | *T* | 84.5 | 51.2 | 71.6 |
| | *TG* | 95.8 | 55.1 | 78.0 |
| V | *G* | 86.4 | 44.0 | 62.1 |
| | *T* | 87.3 | 54.7 | 70.7 |
| | *TG* | 94.0 | 57.1 | 75.1 |

Table 4: Evaluation Results

The coverage of topic-specific dictionaries is lower than those of general ones, but it still adds some gain, when combined, yielding in translations for 95% of the test data. On the other hand, topic-specific word selections achieve a higher accuracy than general translations. Again, the combination of both dictionaries works best resulting in around 60% "correct" word translations. Taking into account possible word synonyms and paraphrases of the input translations, we might achieve higher accuracy figures when evaluated by humans. However, the upper boundary of our proposed method is around 80%, a relative improvement of up to 13% towards the baseline results using the translations of the general dictionary.

## References

Imamura, K. (2001) Hierarchical Phrase Alignment Harmonized with Parsing. In Proc. of the 6th NLPRS (pp. 377-384).

Melamed, D. (2000) Models of Translation Equivalence among Words. In Computational Linguistics 26-2 (pp. 221-249).

Ohno, S. and Hamanishi, M. (1984), Ruigo-Shin-Jiten, Kadokawa.

Takezawa, T. et al. (2002). Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In Proc. of the 3rd LREC (pp. 147-152), Las Palmas, Spain.