Selection of Entries for a Bilingual Dictionary from Aligned Translation Equivalents using Support Vector Machines

Takeshi KUTSUMI
SHARP Corporation
Yamatokoriyama, Nara, Japan
kutsumi.takeshi@sharp.co.jp

Takehiko YOSHIMI
Ryukoku University
Seta-oe, Otsu, Shiga, Japan

Katsunori KOTANI
National Institute of Information and Communications Technology
Seika-cho, Souraku-gun, Kyoto, Japan

Ichiko SATA
SHARP Corporation
Yamatokoriyama, Nara, Japan

Hitoshi ISAHARA
National Institute of Information and Communications Technology
Seika-cho, Souraku-gun, Kyoto, Japan

Abstract

This paper claims that constructing a dictionary using bilingual pairs obtained from parallel corpora needs not only correct alignment of two noun phrases but also judgment of its appropriateness as an entry. It specifically addresses the latter task, which has been paid little attention. It demonstrates a method of selecting a suitable entry using Support Vector Machines, and proposes to regard as the features the common and the different parts between a current translation and a new translation. Using experiment results, this paper examines how selection performances are affected by the four ways of representing the common and the different parts: morphemes, parts of speech, semantic markers, and upper-level semantic markers. Moreover, we used n-grams of the common and the different parts of above four kinds of features. Experimental result found that representation by morphemes marked the best performance, F-measure of 0.803.

1 Introduction

The acquisition of lexical knowledge from parallel corpora is a promising approach to the extension of bilingual dictionaries for machine translation (MT) systems and cross-lingual information retrieval applications. This paper claims that acquiring appropriate lexical knowledge consists of two processes: one to align two expressions in different languages correctly, and the other to judge their appropriateness as entries for a bilingual dictionary.

The latter process is necessary because not all aligned expressions contribute to improvement of the quality of translation, and therefore expressions which damage the translation quality should be filtered out. Suppose the following three English-Japanese pairs are extracted from a parallel corpus:

Customs and Tariff Bureau	関税局 *kanzeikyoku*
Minshuto and New Komeito	民主党や公明党 *minshuto ya komeito*
Miyagi and Yamagata	宮城，山形両県 *miyagi, yamagata ryoken*

Out of these bilingual pairs, the first pair is suitable for an entry for the bilingual dictionary of our English-to-Japanese MT system, while the rest would not be appropriate. The second pair does not need to be registered in the dictionary because our MT system is able to translate "Minshuto and New Komeito" into "民主党，及び，公明党 *minshuto, oyobi, komeito*," which makes little difference in the translation quality with the Japanese expression of the bilingual pair extracted from the parallel corpus. The inappropriateness of the third pair comes from the fact that it is incorrect in the context in which "Miyagi" or "Yamagata" does not indicate "県 *ken* prefecture."

Another reason why we emphasize the necessity of the selection process is that it is a system-dependent process while the alignment process is system-independent. The correctness of aligning two expressions is independent of a particular MT system. Accordingly, the alignment process can function without assuming any particularity of the system. On the other hand, the selection process must assume a specific system because whether a bilingual pair needs to be registered in the dictionary depends on the specific MT system. In the case of our MT system, for example, the pair of

"Bank for International Settlements" and "国際決済銀行 *kokusai kessai ginko*" should be judged to be a pair which needs registration, because our system is not equipped with a rule for translating "Bank for *ABC*" into "*ABC* 銀行 ABC *ginko*" (a rule for translating a preposition into nil). The judgement should be opposite for a MT system which is equipped with such a rule.

There is no evidence of previous work dealing specifically with the selection process, as contrasted with much attention given to the alignment process (Smadja96, Melamed99, Le00, Mcewan02, Tufis02, Utsuro02, Sadat03, Sato03, Yamamoto03, Ayan04,Izuha04, Sahlgren04).

We propose a method of selecting appropriate entries for a dictionary from aligned expressions. Our method uses Support Vector Machines to construct a selection model. This paper targets complex proper noun phrases in English defined as proper noun phrases with prepositional phrases and/or coordinated phrases (hereafter CPNP) like "National Institute of Information and Communications Technology." In this paper, we introduce n-grams of elements into our previous method (Kutsumi05) to enhance it, and we evaluate the effectiveness of our new method.

2 Training Data

We use as our training data a set of bilingual pairs consisting of a CPNP and its Japanese counterpart, which are compiled as the candidates for entries of the bilingual dictionary of Sharp's English-Japanese MT system[1]. We made the selection of pairs suitable for Sharp's MT system. If we perform it for another MT system, we use the target MT system to obtain current translations. The positive examples in the training data are the bilingual pairs which are judged by evaluators to be added to the dictionary, while the negative examples are those which are judged not to be added.

A positive example consists of the current translation for a CPNP and the new translation for the CPNP:

CPNP	Special Committee on Medical Devices
Current Trans.	医療用具上の特別委員会
New Trans.	医療用具特別部会

The current translation "医療用具上の特別委員会 *iryo yogu jono tokubetsu iinkai*" is acquired by translating the CPNP "Special Committee on Medical Devices" using the current version of the bilingual dictionary. The new translation "医療用具特別部会 *iryo yogu tokubetsu bukai*" is a candidate for a new entry of the dictionary. Comparing the two Japanese noun phrases, evaluators have judged that the quality of translation for the CPNP could be improved by adding the new translation to the dictionary.

A negative example is a bilingual pair which has been judged to potentially cause a decrease in translation quality if it were added to the dictionary:

CPNP	United Kingdom and Scandinavia
Current Trans.	英国、及び、スカンジナビア
New Trans.	英国スカンジナビア経済同盟

The new translation 英国スカンジナビア経済同盟 *eikoku sukanjinabia keizai domei* "United Kingdom Scandinavia economic alliance" is an inappropriate translation because the CPNP does not always mean an economic alliance.

3 Features for Machine Learning

The features our selection method utilizes for machine learning are the common and the different parts between a current and a new translation. The parts are generated by applying the UNIX command **diff** to the two translations.

The **diff** command compares two files line by line (unit by unit). Out of several ways of representing linguistic information as a unit, we examine the selection performances in the following four ways: morphemes, parts of speech, semantic markers, and upper-level semantic markers. Morphemes and parts of speech are obtained by using a morphological analyzer "ChaSen[2]."

After analyzing the current translation "医療用具上の特別委員会 *iryo yogu jono tokubetsu iinkai*" and the new translation "医療用具特別部会 *iryo yogu tokubetsu bukai*" with "ChaSen", we apply the **diff** command to the two files in which each morpheme is given in a line. The application displays the following features:

(a) comm(医療/用具)
diff(上/の, NIL)
comm(特別)
diff(委員/会, 部会)

"diff(*A*, *B*)" means that *A* and *B* differ in the two files, and "comm(*C*)" shows a common part. "NIL" means that no counterpart exists in the other file. A slash separates morphemes.

Comparing the two translations based on the parts of speech gives the following features:

[1] http://www.sharp.co.jp/ej/

[2] http://chasen.naist.jp/

(b) comm(名詞-一般/名詞-一般)
diff(名詞-接尾-副詞可能/助詞-連体化, NIL)
comm(名詞-形容動詞語群/名詞-一般)
diff(名詞-接尾-一般, NIL)

By mapping of morphemes into semantic markers, we can see the common and different parts as follows:

(c) comm(0fe1dd/3cedca)
diff(1eb357/undef, NIL)
comm(2016ed)
diff(3dcaa4/3ceda8, 107777)

The semantic markers such as "0fe1dd" are obtained by consulting the EDR concept dictionary[3]. If we encounter semantic ambiguity, like when there is more than one entry for a morpheme in the concept dictionary, we select one of them randomly.

The EDR concept dictionary has a hierarchical structure. We term upper-level semantic markers semantic markers ranked one level higher than those corresponding to the morphemes discussed above. For example, one of the upper-level semantic markers corresponding to "医療 *iryo* medical" is "30f84f." In cases where semantic markers has multiple upper-level semantic markers, one is chosen at random. By mapping of morphemes into upper-level semantic markers, we can see the common and the different parts as follows:

(d) comm(30f84f /3cfbb9)
diff(4447c6/undef, NIL)
comm(201bb4)
diff(44484c/444549, 444614)

Next we will introduce the N-gram of common and different parts. The introduction enables us to take into account the order of the common and the different parts. The bigrams and the trigrams made from (a) above would be like (e) and (f) respectively:

(e) comm(医療/用具) - diff(上/の, NIL)
diff(上/の, NIL) - comm(特別)
comm(特別) - diff(委員/会, 部会)

(f) comm(医療/用具) - diff(上/の, NIL) - comm(特別)
diff(上/の, NIL) - comm(特別) - diff(委員/会, 部会)

4 Experiment

The data set used in the experiment consists of 10,154 positive examples and 8,878 negative examples. We performed five-fold cross-validation on this data set. We made use of TinySVM[4], and selected first order polynomial as the type of kernel function. We use as avaluation critaria the recall, the precision, and the F-measure. In the formula (1) to determine the F-measure, parameter b, which indicates the weight of precision corresponding to the recall, was set at 0.5.

$$F-measure = \frac{(b^2+1)\times precision \times recall}{b^2 \times precision + recall} \quad (1)$$

4.1 Selection Performance

Table 1 shows the precision, recall, and F-measure for each way of representing linguistic information as a unit. The best F-measure is gained by using a morpheme as a unit of presenting common and different parts.

	Precision	Recall	F-measure
Morpheme	0.857	0.640	0.803
Part of Speech	0.683	0.810	0.705
Semantic Marker	0.815	0.566	0.749
UL Semantic Marker	0.796	0.611	0.750

Table 1 : Experimental Result

Table 2 shows the precision, recall, and F-measure when the features are represented in the combination of each linguistic information (morpheme, part of speech, semantic marker, upper-level semantic marker) and N-grams (unigram, bigram, trigram, unigram+bigram and unigram+bigram+trigram).

This result shows that when morphemes and semantic markers (including the case of upper-level semantic marker) are used as features, there is a tendency for the F value to be higher when the unigram is included in the features. On the contrary, when part of speech is used as a feature, there is a tendency for the F value to be higher when the bigram is included in the features.

Table 2 shows the following points about the precision: (1) when the part of speech is used, there is a tendency for the ratio to be higher when trigrams are included in the features; (2) when morphemes, semantic markers or upper-level semantic markers are used, the ratio tends to be low when bigrams are included in the features.

[3] http://www2.nict.go.jp/kk/e416/EDR/J_index.html

[4] http://chasen.org/~taku/software/TinySVM/

		Precision	Recall	F-measure
Morpheme	unigram	0.857	0.640	0.803
	bigram	0.890	0.192	0.515
	trigram	0.538	0.995	0.592
	u + b	0.864	0.610	0.798
	u + b + t	0.864	0.605	0.796
Part of Speech	unigram	0.683	0.810	0.705
	bigram	0.843	0.638	0.792
	trigram	0.856	0.419	0.708
	u + b	0.739	0.755	0.743
	u + b + t	0.710	0.887	0.739
Semantic Marker	unigram	0.815	0.566	0.749
	bigram	0.858	0.170	0.475
	trigram	0.540	0.996	0.594
	u + b	0.835	0.518	0.744
	u + b + t	0.837	0.515	0.744
Upper-level Semantic Marker	unigram	0.796	0.611	0.750
	bigram	0.864	0.212	0.535
	trigram	0.540	0.995	0.595
	u + b	0.824	0.570	0.757
	u + b + t	0.823	0.566	0.755

Table 2 : Experimental Result (including bigram and trigram)

As for the recall, Table 2 demonstrates the clearly different tendencies when using part of speech as a feature when compared with in the case of morphemes, semantic markers and upper-level semantic markers. Namely it tells us that (1) when the part of speech is used, a low ratio results when trigrams are used as features; (2) when morphemes, semantic markers or upper-level markers are used, an extremely low ratio results when bigrams are used as features while an extremely high ratio results when trigrams are used.

To sum up, the above observations on the precision and the recall indicate that (1) in the case of morphemes, semantic markers and upper-level semantic markers, the use of bigrams results in little noise and many misses, and the use of trigrams results in much noise and few misses; (2) in the case of part of speech, the use of trigrams yields the lower occurance of noise and the higher number of misses.

4.2 Contribution of Features

This subsection investigates the degree of contribution of each feature to the correct selections. Classification of a test example **x** in the first order polynomial function is carried out using the following discriminant function f(x):

$$f(\mathbf{x}) = \mathrm{sgn}(\sum_{i=1}^{n} \alpha_i y_i (\mathbf{x}_i \bullet \mathbf{x} + 1) + b) \qquad (2)$$

where $\mathbf{x}_i$ is the feature vector of training data and y_i is its class ($1 \leqq i \leqq n$), and α_i is the weight obtained by learning. Let $\mathbf{x} = (x^1, x^2, \ldots, x^m)$ and $\mathbf{x}_i = (x_i^1, x_i^2, \ldots, x_i^m)$, then calculating the inner product in Formula (2) gives the following formula:

$$f(\mathbf{x}) = \mathrm{sgn}(\sum_{j=1}^{m} \sum_{i=1}^{n} \alpha_i y_i x_i^j x^j + \sum_{i=1}^{n} \alpha_i y_i + b) \qquad (3)$$

$\sum_{i=1}^{n} \alpha_i y_i x_i^j$ is the weight for the feature x^j. When the weight is a positive number, it means that the feature x^j contributes to the classification of positive examples. The degree of contribution is proportional to the absolute value of the weight.

We calculated the weights by using a morpheme as a unit where the best F-measure is gained. The top ten features with the highest degree of contribution are shown in Table 3. Table 3 indicates that most of the features with the highest weight concern the different parts between the current and the new translations.

Rank	Feature	Weight
1	comm(及び)	2.647463
2	diff(の下院議員, 共和国)	1.543886
3	diff(ためのアジェンダ, 課題)	1.424892
4	comm(仮釈放)	1.398082
5	diff(法, 法則)	1.391991
6	diff(の部門, 省)	1.385298
7	diff(社産業の財務諸表, 法人企業統計)	1.344135
8	diff(貴重な頂上、アンズのオーダ, 勲五等宝冠章)	1.344131
9	diff(利便の同盟, 一時的便宜協定)	1.344130
10	diff(補助に関するグローバルなプログラム, エイズ対策特別計画)	1.344124

Table 3 : Top 10 Features with the Highest Weight (positive examples)

We investigated examples which are composed of the features in Table 3. The investigation revealed that most of the features concerned the improvement of the translation quality of the constituent nouns of CPNPs. The following is a positive example which includes the feature with the highest weight "diff(の下院議員, 共和国):"

CPNP	Rep. of Afghanistan
Current Trans.	アフガニスタンの下院議員
New Trans.	アフガニスタン共和国

The feature "diff(の下院議員, 共和国)" has frequently appeared in positive examples where higher translation quality would be achieved by interpreting the abbreviation "Rep." as "Republic (共和国 *kyowakoku*)," not as "Representative (下院議員 *kain giin*)," when the expression "*ABC*" in "Rep. of *ABC*" represents a country's name or a part of it. The feature "diff(のための部門, 省)"

means that "*ABC*省 *ABC sho*" is often better than "*ABC*のための部門 *ABC notameno bumon*" as a translation of "Department of *ABC*."

Features such as "diff(の下院議員, 共和国)," "diff(の部門, 省)" and "diff(のためのアジェンダ, 課題)" indicate that the new translations are better than current ones in terms of translation quality of prepositions as well as nouns. These features mean that better translations are gained by translating the prepositions "of" and "for" into nil, not into "の *no*" and "のための *notameno*."

The top ten features with the highest degree of contribution the correct selections of the negative examples are shown in Table 4. We investigated examples which are composed of the features in Table 4, and were able to divide the examples into two categories.

Rank	Feature	Weight
1	diff(NIL, 、)	-2.931983
2	diff(NIL, の)	-2.247715
3	diff(NIL, ；（社）)	-2.001176
4	diff(国家の貿易, 全米統一通商)	-1.962333
5	diff(NIL, 、財団法人)	-1.832608
6	diff(農業、林学、及び、水産業のための行政上の, 農林水産)	-1.764492
7	diff(のための世界首脳会議, サミット)	-1.717183
8	diff(改善改革, NIL)	-1.693394
9	diff(NIL, 、及び、)	-1.680185
10	diff(保障、及び、基本的自由のための会議, 条約)	-1.616857

Table 4 : Top 10 Features with the Highest Weight (negative examples)

One category shows that current translations are more appropriate than new ones. For example, "diff(国家の貿易，全米統一)," the feature with the forth degree of contribution in Table 4, appears in the following example:

CPNP	Committee for a National Trade Policy
Current Trans.	国家の貿易政策のための委員会
New Trans.	全米統一通商

Since the new translation "全米統一通商" means the committee of the United States, it is inappropriate in every case CPNP doesn't refer to the committee of the United States.

Another example, "diff(NIL, の)," the feature with the second degree of contribution in Table 4, appears in the following example:

CPNP	The Japan Society of Clinical Hematology
Current Trans.	日本臨床血液学会
New Trans.	臨床の血液学の日本社会

While the current translation "日本臨床血液学会" corresponds to an actual organization name, the new translation "臨床の血液学の日本社会" is inappropreate. Therefore this example is not suitable for entry for a dictionary.

Another category shows that new translations include so-called "garbage." The first-degree feature "diff(NIL, 、)" indicates that since the new translation has a pause mark "、" at the top of the new translation in the following example, this translation can be regarded as inappropriate for direct entry to a dictionary:

CPNP	Committee to Protect Journalists
Current Trans.	ジャーナリストを保護するための委員会
New Trans.	、ジャーナリスト保護委員会

The third-degree feature "diff(NIL, ；(社))" suggests the inappropriateness of the new translation in the following which includes the comment "; (社):"

CPNP	Japan Society of Corrosion Engineering
Current Trans.	日本腐食防食協会
New Trans.	腐食防食協会；(社)

5 Conclusion

This paper claims that constructing a dictionary using bilingual pairs obtained from parallel corpora needs not only correct alignment of two noun phrases but also judgment of its appropriateness as an entry. It specifically addresses the latter task, which has been paid little attention. It demonstrates a method of selecting a suitable entry using Support Vector Machines, and proposes to regard as the features the common and the different parts between a current translation and a new translation. Using experiment results, this paper examines how selection performances are affected by the three ways of representing the common and the different parts: morphemes, parts of speech, semantic markers, and their N-grams. Moreover, we tested n-grams of the common and the different parts of above four kinds of features. Experimental result found that representation by morphemes marked the best performance, F-measure of 0.803.

References

[Ayan04] Ayan, N., Dorr, B. and Habash, N.: Multi-Align: Combining Linguistic and Statistical Techniques to Improve Alignments for Adaptable MT, *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas*, pp.17-26, 2004.

[Izuha04] Izuha, T.: Machine Translation Using Bilingual Term Entries Extracted From Parallel Texts, *The Transactions of IEICE*, Vol.J87-D-II, No.6, pp.1244-1251, 2004.

[Le00] Le, S., Youbing, J. and Yufang, S.: Word Alignment of English-Chinese Bilingual Corpus based on Chunks, *Proccedings of the Joint SIGDAT Conference on Empirical Methods in Natural, Language Processing and Very Large Corpora*, pp.110-116, 2000.

[Mcewan02] McEwan, C., Ounis, I. and Ruthven, I.: Building bilingual dictionaries from parallel web documents, *Proceedings of the 24th European Colloquium on Information Retrieval Research*, pp.303-323, 2002.

[Melamed99] Melamed, I.: Bitext Maps and Alignment via Pattern Recognition, *Computational Linguistics*, Vol.25, No.1, pp.107-130, 1999.

[Sadat03] Sadat, F., Yoshikawa, M. and Uemura, S.: Bilingual Terminology Acquisition from Comparable Corpora and Phrasal Translation to Cross-Language Information Retrieval, *Proceedings of the Companion Volume to the Proceedings of 41st Annual Meeting of the Association for Computational Linguistics*, pp.141-144, 2003.

[Sahlgren04] Sahlgren, M.: Automatic Bilingual Lexicon Acquisition Using Random Indexing of Aligned Bilingual Data, *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pp.1289-1292, 2004.

[Sato03] Sato, K. and Saito, H.: Extracting Word Sequence Correspondences Based on Support Vector Machines, *Journal of Natural Language Processing*, Vol.10, No.4, pp.109-124, 2003.

[Smadja96] Smadja, F., Hatzivassiloglou, V. and McKeown, K.: Translating Collocations for Bilingual Lexicons: A Statistical Approach, *Computational Linguistics*, Vol.22, No.1, pp.1-38, 1996.

[Tufis02] Tufis, D.: A Cheap and Fast Way to Build Useful Translation Lexicons, *Proceedings of the 19th International Conference on Computational Linguistics*, pp.1030-1036, 2002.

[Utsuro02] Utsuro, T., Horiuchi, T., Chiba, Y.and Hamamoto, T. : Semi-automatic Compilation of Bilingual Lexicon Entries from Cross-Lingually Relevant News Articles on WWW News Sites, *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas*, pp.165-176, 2002.

[Yamamoto03] Yamamoto, Y. and Sakamoto, M.: Extraction of Technical Term Bilingual Dictionary from Bilingual Corpus, *IPSJ SIG Notes* NL 94-12, Information Processing Society of Japan, pp.85-92, 1993.

[Kutsumi05] Kutsumi, T., Yoshimi, T., Kotani, K., Sata, I. and Isahara, H.: Selection of Entries for a Bilingual Dictionary from Aligned Translation Equivalents using Support Vector Machines, *Proceedings of Pacific Association for Computational Linguistics 2005*, 2005.