Improving Compositional Translation with Comparable Corpora

Hiroyuki Kaji, Takashi Tsunakawa, Yoshihiro Komatsubara
Department of Computer Science, Shizuoka University
3-5-1 Johoku, Naka-ku, Hamamatsu-shi, 432-8011, Japan
E-mail: {kaji, tuna}@inf.shizuoka.ac.jp, gs10017@s.inf.shizuoka.ac.jp

Abstract

We improved the compositional term translation method by using comparable corpora. A bilingual lexicon consisting of pairs of word sequences within terms and their correlations is derived from a bilingual document-aligned corpus. Then, for an input term, compositional translations are produced together with their confidence scores by consulting the corpus-derived bilingual lexicon. Thus, we can select the correct translation for the input term from among as many candidate ones as possible. An experiment with a comparable corpus of Japanese and English scientific-paper abstracts demonstrated that compositional translation using the corpus-derived bilingual lexicon outperforms that using an ordinary bilingual lexicon. Future work includes the incremental improvement of the bilingual lexicon with correlations, the refinement of the confidence score, and the extension of the compositional translation model to allow word order to be changed.

Keywords: term translation, comparable corpus, bilingual lexicon

1. Introduction

Technical term translation is one of the key issues in document translation as well as crosslingual information retrieval. Obviously, no existing bilingual lexicon covers all of the terms in a domain. However, most technical terms are compound words and 88% of Japanese technical terms in some domains have compositional English translations (Tonoike, et al. 2006). Thus, the compositional translation method plays an essential role in translating technical terms.

The performance of the compositional translation method naturally depends on the bilingual lexicon it consults. It cannot produce a correct translation for a term unless the lexicon provides appropriate translations for the constituent words of the term. At the same time, it is difficult to select a correct translation for the term from among many candidate translations produced compositionally when the bilingual lexicon provides as many translations as possible for each of the constituent words. It should be noted that the latter problem may become more serious if we improved the coverage of the bilingual lexicon to overcome the former problem.

We propose improving the compositional translation method by using a bilingual corpus. A wide-coverage bilingual lexicon, which consists of word sequence pairs in two languages together with their correlations, is acquired from a bilingual corpus. Then, a ranked list of translations is produced for an input term by compositionally generating candidate translations together with their confidence scores based on the correlations between the constituent words and their translations. Our contribution is not bilingual lexicon acquisition from a bilingual corpus but an improved compositional translation method with confidence scores.

Our proposed framework is compatible with both parallel and comparable corpora. Parallel corpora generally produce bilingual lexicons with more reliable correlations than comparable corpora (Och and Ney 2003; Koehn et al. 2003). However, there are few domains in which large parallel corpora are available. Therefore, we assume that the input corpus is a comparable corpus, more specifically a document-aligned corpus. Use of weakly comparable corpora, which are much more widely available but may produce bilingual lexicons with less reliable correlations, is beyond the scope of this paper (Fung and Yee 1998; Rapp 1999; Andrade et al. 2010; Ismail and Manandhar 2010; Morin and Prochasson 2011).

There have been many studies on bilingual lexicon acquisition from parallel or comparable corpora, where the task is usually to find translations for terms occurring in the input corpus. Bilingual lexicon acquisition methods have usually been evaluated in terms of recall and precision of target language translations acquired for source language terms occurring in the input corpus (Fung and Yee 1998; Rapp 1999; Cao and Li 2002; Tanaka 2002). In contrast, our task is to translate a term even when it does not occur in the input corpus; therefore, we evaluated our framework in terms of precision of translations produced for a test set of input terms collected independently of the input corpus. This task setting is natural when we assume practical applications of bilingual lexicons such as document translation and crosslingual information retrieval.

2. Problems and our framework

Consider the pair of a Japanese term “光通信<*HIKARI TSUUSHIN*>” and its English translation “optical communication.” We humans can recognize the correspondence between “光<*HIKARI*>” and “optical” as well as that between “通信<*TSUUSHIN*>” and “communication.” In other words, the translation from “光通信” to “optical communication” is compositional. However, few electronic Japanese-English lexicons provide the correspondence between a Japanese noun, e.g.,

"光," and an English adjective, e.g., "optical." Therefore, the automatic compositional translation method usually fails to produce the correct translation "optical communication" for the input term "光通信."

Assume that a pair of a Japanese noun "光" and an English adjective "optical" has been registered in a bilingual lexicon. It would provide possible translations, such as "light," "ray," and "beam," as well as "optical" for "光." Likewise, it would provide possible translations, such as "communication," "correspondence," and "report," for "通信." Thus, the compositional translation method may produce many candidate translations including "optical communication," "optical correspondence," "optical report," "light communication," "light correspondence," and others from which it must select the correct one.

As exemplified above, the compositional translation method exhibits two problems, incomplete bilingual lexicons and many candidate translations most of which are spurious. To overcome these problems, we propose a framework consisting of the following two steps: (1) acquiring a bilingual lexicon with correlations from a bilingual corpus, and (2) producing compositional translations together with confidence scores.

(1) Acquiring a bilingual lexicon with correlations from a bilingual corpus

We assume that a comparable corpus consisting of pairs of relevant documents is available and we use the method for calculating pairwise correlations between words in two languages based on co-occurrence statistics in aligned sentences (Matsumoto and Utsuro 2000). This method, which is originally intended for parallel corpora, is applicable to comparable corpora by treating document pairs as sentence pairs (Utsuro et al. 2003). It seems workable as long as the documents are small. It has an advantage in that it does not require a seed bilingual lexicon unlike other methods applicable to comparable corpora.

Our purpose was to construct a wide-coverage bilingual lexicon of term constituents rather than the actual terms. Most of the correspondences between constituents are those between simple words, e.g., "光" and "optical," but some are those between a simple word and a compound word, e.g., "薄膜<*HAKUMAKU*>" and "thin film," and vice versa, e.g., "移動体<*IDOU TAI*>" and "mobile." Therefore, we need to extract not only pairs of simple words but also mixed pairs of simple and compound words. However, it is not necessarily easy to identify compound words. Moreover, from a practical point of view, it is preferable that the bilingual lexicon provides possible translations for any word sequence included in a term; translation pairs of longer word sequences would increase the possibility of correct translations being produced for a term. Therefore, we consider any word sequence included in a term as its constituent and calculate pairwise correlations between those in the source and target languages.

(2) Producing compositional translations together with confidence scores

To select a correct translation from among many candidate translations produced compositionally, we calculate a confidence score for each of the candidate translations. Note that constituent translation pairs have been acquired together with their correlations. We regard the correlations as the confidence scores for the constituent translations and define the confidence score for a compositional translation based on the scores for its constituent translations.

As mentioned in Step 1, the bilingual lexicon provides translations not only for a word but also for a word sequence. However, their correlations or confidence scores are not so reliable. Therefore, we re-evaluate the translations the bilingual lexicon provides for a word sequence: namely, we produce compositional translations for a word sequence even when it is included in the bilingual lexicon and combine the two confidence scores, one provided by the bilingual lexicon and the other calculated compositionally.

The following two sections describe the two steps of our framework in some detail, where the source and target languages are assumed as Japanese and English, respectively. Our framework can be applied to any language pairs with some modifications in language-specific issues such as treatment of morphology.

3. Acquiring bilingual lexicon for compositional translation

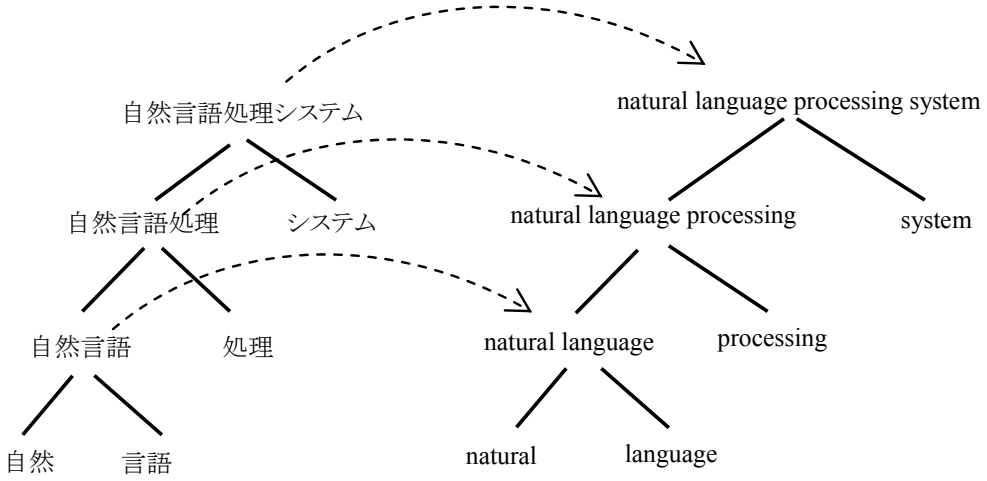
We extract every word sequence included in terms from both Japanese and English documents. Most Japanese terms are Noun+, i.e., sequences of one or more nouns, and most English terms are Adjective*Noun+, i.e., sequences of one or more nouns optionally preceded by one or more adjectives, where adjectives include present participles and past participles of verbs. At present, we do not deal with terms with more complicated structures, e.g., those including prepositional phrases. Therefore, we extract every Japanese word sequence consisting of nouns and every English word sequence consisting of nouns and adjectives.

We define the correlation of a Japanese word sequence J and an English word sequence E by using Dice's coefficient. That is,

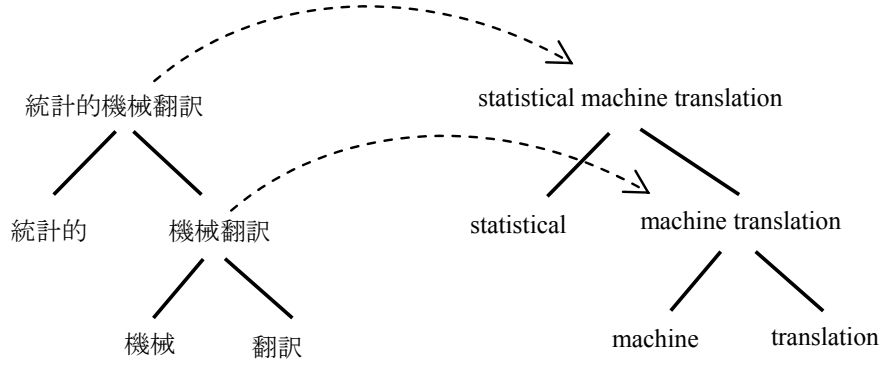
$$C(J,E) = \frac{2 \cdot g(J,E)}{f(J) + f(E)}, \qquad [1]$$

where $f(J)$ and $f(E)$ are the number of Japanese documents and that of English documents in which J and E occur, respectively, and $g(J, E)$ is the number of pairs of Japanese and English documents in which J and E co-occur.

We ignore the frequencies of word sequences occurring in a document. This is because we intend to apply our framework to nonparallel corpora: the frequency of a Japanese word sequence occurring in a Japanese document is not necessarily comparable to that of the corresponding English word sequence occurring in



(a) Example 1



(b) Example 2

Fig. 1: Structure of terms and compositional translation

the English document aligned with the Japanese document. We also ignore the lengths of word sequences because they are not necessarily maintained across languages, as exemplified by the pairs (移動体<*IDOU TAI*>, mobile) and (薄膜<*HAKUMAKU*>, thin film).

It should be noted that we distinguish between maximal word sequences, which are not subsequences of longer word sequences, and non-maximal word sequences. Japanese maximal and non-maximal word sequences tend to correspond to English maximal and non-maximal word sequences, respectively, in a pair of aligned documents. Accordingly, a document pair is counted as 0.5 for a pair of maximal and non-maximal word sequences co-occurring in the document pair, while it is counted as 1.0 for a pair of maximal word sequences co-occurring in it as well as for a pair of non-maximal word sequences co-occurring in it. Assume that "光通信" and "optical communication" co-occur as maximal word sequences in a pair of aligned documents. This document pair is counted as 0.5 for pairs (光, optical communication), (通信, optical communication), (光通信, optical), and (光通信, communication), while it is counted as 1.0 for pairs (光通信, optical communication), (光, optical), and (通信, communication) (Note that it is also counted as 1.0 for incorrect pairs (光, communication) and (通信, optical)). Thus, we reduce the confusion between a compound word and its constituent words.

Since the correlations are unreliable for a word sequence infrequently occurring in the input corpus, we set a threshold θ_f for the number of documents in which a word sequence occurs. We calculate correlations for every pair of Japanese and English word sequences both of which occur in θ_f or more documents. Since we intend to translate Japanese terms into English, we select the top N_1 English word sequences in descending order of correlation for each Japanese word sequence (In the experiment described in Sec. 5, we set θ_f and N_1 to 10 and 20, respectively.).

4. Compositional translation with confidence score

Note that a term can be represented with a binary tree according to its head-modifier relations, as exemplified in Fig. 1. We assume that Japanese term J can be compositionally translated into English term E if and only if J and E are isomorphic or represented with the same

Input term: $jw_1^N (= jw_1 jw_2 \cdots jw_N)$

Bilingual lexicon: L

Triangular matrix: $A(i,j)$ (i=1,…,N; j=i,…,N)

Algorithm:

1) For i:=1 to N do

For j:=i to N do

$A(i,j) \leftarrow \phi$.

2) For i:=1 to N do

$A(i,i) \leftarrow \{\langle ew, S(jw_i, ew)\rangle \mid (jw_i, ew) \in L\}$

3) For r:=1 to N–1 do

For i:=1 to N–r do

For k:=i to i+r–1 do

$$A(i,j) \leftarrow A(i,j) \cup \left\{ \langle e_1 e_2, S(jw_i^j, e_1 e_2)\rangle \mid \begin{matrix} \langle e_1, S(jw_i^k, e_1)\rangle \in A(i,k), \\ \langle e_2, S(jw_{k+1}^j, e_2)\rangle \in A(k+1, j) \end{matrix} \right\}$$

Fig. 2: Compositional translation algorithm

binary tree. Based on this assumption, we define the confidence score $S(J,E)$ for the compositional translation from a Japanese term or word sequence J to an English word sequence E as follows:

$$S(J,E) = \begin{cases} \lambda \cdot S'(J,E) + (1-\lambda) \cdot C(J,E) \\ \qquad (|J| \geq 2, |E| \geq 2) \\ C(J,E) \\ \qquad (\min\{|J|, |E|\} = 1) \end{cases}, \quad [2]$$

where $S'(J,E)$ is a confidence score based on compositionality, $C(J,E)$ is the correlation based on co-occurrence in pairs of aligned documents, λ is a parameter adjusting the weights for $S'(J,E)$ and $C(J,E)$, and $|J|$ and $|E|$ denote the lengths of word sequences J and E, respectively.

We define the confidence score based on compositionality as follows:

$$S'(J,E) = \max_{\substack{1 \leq i < p \\ 1 \leq j < q}} \frac{2 \cdot S(jw_1^i, ew_1^j) \cdot S(jw_{i+1}^p, ew_{j+1}^q)}{S(jw_1^i, ew_1^j) + S(jw_{i+1}^p, ew_{j+1}^q)}, \quad [3]$$

where $J = jw_1 jw_2 \cdots jw_p (= jw_1^p)$ and $E = ew_1 ew_2 \cdots ew_q (= ew_1^q)$. This formula is based on the following idea. We define the confidence score based on compositionality as the harmonic mean of the confidence scores for the two constituent translations. However, we do not know the correct structures for J and for E. Therefore, we calculate the confidence score for every combination of possible decompositions of J and E and select the maximum confidence scores based on the assumption that the combination of correct structures maximizes the confidence score.

Formula [3] shows that we assume the coincidence of word order between a Japanese term and its English translation. This is generally not the case. It is not difficult to modify the formula to deal with the change in word order. Moreover, this formula does not contain the factor representing the compatibility of the two constituent translations. It should be noted that correlation $C(J,E)$ reflects to some extent the compatibility of the constituent translations.

Next, we describe a dynamic programming algorithm for compositionally producing translations. It is similar to the CKY parsing algorithm for context-free grammars, as shown in Fig. 2. It constructs a triangular matrix $A(i,j)$ consisting of cells each of which corresponds to a subsequence jw_i^j in the input term and contains translation candidates and their confidence scores for its corresponding subsequence. To prevent combinatorial explosion, we restrict candidate translations contained in each cell to those with N_2 highest confidence scores (In the experiment described in Sec. 5, we set N_2 to 100.).

5. Experiment

5.1 Experimental settings

本報告では，振幅スペクトルからの音声合成法を利用した音質劣化の少ないピッチ変換法を提案する。本方式ではまず，サンプリング変換を用いて周波数スケーリングを行なうことでピッチ周波数を変更する。その後，サンプリング変換した音声に対して，振幅スペクトル系列のスペクトル包絡特性を原音声の特性に復元する。最後に，変更された振幅スペクトル系列から，原音声の音声速度と等しくなるように音声を合成する。本方式で得られるピッチ変換音声は，変換倍率が約0．8倍から2．0倍の範囲では，かなり原音声の音質を保存している。	This paper proposes a method of pitch modification using the speech synthesis method from short‐time Fourier transform (STFT) magnitude. The method modifies first the pitch frequency by frequency scaling using sampling rate conversion. For the speech whose sampling rate is converted, spectral envelopes of STFTs magnitude are restored to the ones of original speech. Finally, a speech is synthesized from the modified STFT magnitude but the frame shift rate for synthesis is set so that the synthesis speech rate equals to the original one. The resulting pitch modified speech can preserve very well the quality of original speech over the range 0.8‐2.0 of rate conversion.

(a) Almost parallel

鳥沢のHPSGパージングアルゴリズムは，HPSGの辞書項目からコンパイルされたCFG（文脈自由文法）を用いるフェーズ1と，それだけではカバーしきれない制約を素性構造を用いて計算するフェーズ2からなる。本稿ではフェーズ1の並列化アルゴリズムを提案した。超並列計算機AP1000＋上で並列オブジェクト指向言語ABCL／fを用いて実装した。新聞を例題として50語以下の文（平均19語）をパージングし，構文木をすべて数え上げるのに要した時間は一文当たり98ミリ秒であった。	This paper describes an attempt to develop a parallel parsing algorithm for Torisawa's parsing algorithm for HPSG. Torisawa's algorithm consists of two phases. At Phase 1, a parser enumerates possible parse trees using CFG rules compiled from lexical entries in HPSG. The constraints uncovered by the CFG are solved at Phase 2, using feature structures and a variant of unification, partial unification. We realized a parallel parsing algorithm for Phase 1, on a highly parallel computer AP1000+ (256 Super Sparc 50Mhz) with concurrent object‐oriented programming language ABCL/f. The average parsing time for the sentences consisting of less than 50 words was 98msec.

(b) Totally comparable but organized differently

1台のカメラとターンテーブルを用い，さまざまな角度から撮影した物体の2次元画像から3次元形状を構築する手法を開発した。ターンテーブルの分割角度 θ 毎に仰角 ϕ で対象の2次元画像を撮影し，3次元モデルにより3次元形状モデルを構築する。モデルから復元した2次元画像と元の2次元画像を比較し，復元精度によって3次元モデルの評価解析を行った結果，各種誤差要因のほか形状の複雑さの影響が判明した。形状の複雑度を定義し，複雑度に基づいて精度指標を修正すことで，複雑さの影響を減少した。	Using one CCD camera and the turn table, we propose a method to construct three dimensional object shape from two dimensional images. By comparing the two dimensional image obtained from three dimensional object shape constructed by our proposed method, and original image, we find that three dimensional object shape is restored precisely.

(c) Partially comparable

Fig. 3: Example pairs of Japanese and English abstracts

We conducted an experiment using the Japan Science and Technology Agency (JST) corpus of Japanese and English scientific-paper abstracts. It consists of pairs of Japanese and English abstracts with varying comparability, as exemplified in Fig. 3. The lengths of the Japanese abstracts range from 200 to 500 characters and those of the English abstracts range from 50 to 300 words. We used 107,979 pairs of abstracts in the field of information engineering, published in 1980 through 2004, to derive a bilingual lexicon with correlations. We used a Japanese morphological analyzer Mecab [1] and a language independent part-of-speech tagger TreeTagger [2] to segment the Japanese and English texts into words, respectively.

We prepared two test sets; AI test set consisting of 1,094 Japanese terms with reference English translations from the Japanese-English Index in the Encyclopedia of Artificial Intelligence (JSAI 2008) and NLP test set consisting of 1,661 Japanese terms with reference English translations from the Japanese-English Index in the Encyclopedia of Natural Language Processing (ANLP 2010).

We used the compositional translation method with each of the following three bilingual lexicons to produce a ranked list of English translations for a Japanese term in the two test sets.

(1) Corpus-derived lexicon + ordinary lexicon

The bilingual lexicon derived from the JST corpus was

[1] http://mecab.sourceforge.net/

[2] http://www.ims.stuttgart.de/projekte/corplex/TreeTagger/

merged with the EDR[3] Japanese-English, EDICT[4] Japanese-English, and Eijiro[5] English-Japanese Dictionaries. Since these ordinary lexicons do not contain correlations, a uniform correlation value of 0.1 was given to all pairs of Japanese and English words in them, and the maximum of the two values was selected for a pair of Japanese and English words contained in both the corpus-derived lexicon and the ordinary lexicons.

(2) Corpus-derived lexicon

The bilingual lexicon derived from the JST corpus only

(3) Ordinary lexicon

The EDR Japanese-English, EDICT Japanese-English, and Eijiro English-Japanese Dictionaries were merged into one and, then, augmented so that a ranked list of translation candidates could be output for an input term; namely, each pair of Japanese and English words was given a correlation value proportional to the number of pairs of aligned documents in which they co-occur.

For each bilingual lexicon, λ was adjusted using another set of Japanese terms and their English translations from the Japanese-English Index in the Encyclopedia of Artificial Intelligence. This set was disjoint with the above-mentioned AI test set. The value of λ was 0.40, 0.43, and 0.33 for (1) corpus-derived lexicon + ordinary lexicon, (2) corpus-derived lexicon, and (3) ordinary lexicon, respectively.

5.2 Experimental results

Table 1 lists the mean reciprocal rank (MRR) of the correct translations and Top *k* precision (*k*=1, 3, and 10), i.e., the percentage of input terms whose correct translations were included in those with *k* highest confidence scores, for the compositional translation with each of the three bilingual lexicons, where we judged only the reference translations as correct. The data suggest that the proposed framework is promising; not only the corpus-derived lexicon + ordinary lexicon but also the corpus-derived lexicon outperformed the ordinary lexicon. In Table 1, correct translations are broken down into two categories: translations the bilingual lexicon provides and translations produced compositionally. When the corpus-derived lexicon + ordinary lexicon and the corpus-derived lexicon were used, about 30% of the correct translations were those produced compositionally. This demonstrates the necessity and effectiveness of on-the-fly compositional translation.

Top *k* precisions of at most 50% were very low compared with those reported in previous literature on bilingual lexicon acquisition from parallel or comparable corpora. One of the reasons for the low precision is the test sets prepared independently of the corpus from which the bilingual lexicon derived. In fact, 11% of the Japanese terms in the AI test set and 12% of those in the NLP test set included word sequences not covered by the corpus-derived lexicon. Most of such terms were unpopular transliterated ones, e.g., "タクタイルボコーダ <*TAKUTAIRU BOKOODA*>" (*tactile vocoder*), those including proper nouns, e.g., "ボールドウィン効果 <*BOORUDOUIN KOOKA*>" (*Boldwin effect*), and scarcely used terms, e.g., "ブラーフミ文字 <*BURAAHUMI MOJI*>" (*Brahmi script*).

The data in Table 1 is rather singular; for almost 80% of the test terms whose correct translations were in top 10, the top ranked ones were actually correct. We can say that the proposed method is reliable for a term occurring rather frequently in the corpus, while it is unreliable for a term occurring infrequently in the corpus. The performance for the NLP test set was much worse than that for the AI test set. This is probably because the JST corpus contains a relatively small number of paper abstracts on natural language processing.

Table 2 lists the results of compositional translation with the corpus-derived lexicon + ordinary lexicon and that with the ordinary lexicon for several input terms. These results suggest the effectiveness of the proposed method as well as room for improvement.

6. Discussion

The compositional translation method has been widely used to extract a pair of a word and its translation from corpora, although it is restricted to extracting a pair of compound words. It usually consults an existing bilingual lexicon to generate candidate translations, which then are validated by using a corpus (Cao and Li 2002; Tanaka 2002; Baldwin and Tanaka 2004; Tonoike et al. 2006). In contrast, we proposed consulting a bilingual lexicon derived from a corpus. The experiment demonstrated that our framework improved the possibility of producing a correct translation. Note that unless a correct translation was produced, the validation procedure would be useless. A distinguishing feature of our improved compositional translation method is that it estimates confidence scores for candidate translations. Although there has been work investigating score functions for compositional translation (Tonoike et al. 2006), our score is unique in that it is based on a comparable corpus.

The method described in Sec. 3 is not the only way to derive a bilingual lexicon from a comparable corpus. Alternatively, we can extract parallel sentence pairs from a comparable corpus to acquire a bilingual lexicon with a statistical machine translation tool. This is a common way to exploit comparable corpora for SMT (Fung and Cheung 2004; Munteanu and Marcu 2005; Abdul-Rauf and Schwenk 2009). It is also applicable to augmenting a seed bilingual lexicon for contextual similarity-based bilingual lexicon acquisition from a comparable corpus (Morin and Prochasson 2011). Our method based on co-occurrence statistics in pairs of aligned documents should be evaluated comparatively with this alternative. Our method would be better for very nonparallel corpora, while the alternative would be better for comparable

[3] http://www2.nict.go.jp/r/r312/EDR/index.html
[4] http://www.csse.monash.edu.au/~jwb/edict.html
[5] http://www.alc.co.jp/

Table 1: Summary of experimental results

(a) Artificial Intelligence domain (# of test terms: 1,094)

Bilingual Lexicon	Corpus-derived + ordinary	Corpus-derived	Ordinary
MRR	0.44	0.4	0.22
Top 1 precision	0.402	0.370	0.197
(Bilingual lexicon)	(0.289)	(0.263)	(0.089)
(Compositional translation)	(0.113)	(0.107)	(0.108)
Top 3 precision	0.464	0.428	0.238
(Bilingual lexicon)	(0.326)	(0.297)	(0.112)
(Compositional translation)	(0.138)	(0.131)	(0.125)
Top 10 precision	0.510	0.473	0.351
(Bilingual lexicon)	(0.351)	(0.320)	(0.135)
(Compositional translation)	(0.169)	(0.153)	(0.144)

(b) Natural Language Processing domain (# of test terms: 1,661)

Bilingual Lexicon	Corpus-derived + ordinary	Corpus-derived	Ordinary
MRR	0.35	0.31	0.20
Top 1 precision	0.314	0.282	0.167
(Bilingual lexicon)	(0.231)	(0.202)	(0.102)
(Compositional translation)	(0.083)	(0.081)	(0.066)
Top 3 precision	0.377	0.331	0.217
(Bilingual lexicon)	(0.272)	(0.229)	(0.143)
(Compositional translation)	(0.105)	(0.102)	(0.074)
Top 10 precision	0.415	0.362	0.271
(Bilingual lexicon)	(0.296)	(0.246)	(0,178)
(Compositional translation)	(0.120)	(0.117)	(0.093)

corpora from which many parallel sentence pairs can be extracted.

The followings are the directions for improving our framework. First, we need to improve the bilingual lexicon with correlations. The present corpus-derived bilingual lexicon contains too many spurious pairs. The examples in Table 2 imply that it contains such pairs as (属性<*ZOKUSEI*> (*property*), decision tree), (ネットワーク <*NETTOWAAKU*> (*network*), service), and (反駁 <*HANBAKU*> (*refutation*), PAC learning model). This may be unavoidable as we do not have a seed lexicon. However, once a bilingual lexicon is acquired, we can use it to acquire a less noisy bilingual lexicon. In other words, we can refine our bilingual lexicon incrementally.

Second, there is room for refining the confidence score. Currently, we do not consider the relation or compatibility between constituent translations. A possible refinement of the confidence score is to multiply the harmonic means of the confidence scores for constituent translations by the correlation between the constituent translations, which can be estimated from a target-language monolingual corpus. We have an alternative to this refinement. That is, producing unlikely translations as candidates and validating the candidates by using a target-language monolingual corpus or the Web may not be problematic (Dagan and Itai 1994; Grefenstette 1999; Way and Gough 2003).

Third, we need to extend our compositional translation model to allow word order to be changed. For example, while a Japanese term is a noun sequence, its English translation can include a prepositional phrase. A factor of structural transfer should be incorporated into our confidence score. Some previous work has addressed compositional translation involving changes in word order (Baldwin and Tanaka 2004).

Table 2: Example of compositional translation results

#	Input term	Rank	Corpus-derived + ordinary		Ordinary	Reference translation
			Translation	Score	Translation	
1	属性継承 <*ZOKUSEI KEISHOU*>	1	attribute inheritance	0.060	attribute inheritance	property inheritance
		2	attribute succession	0.023	***property inheritance***	
		3	decision tree inheritance	0.021	characteristic inheritance	
2	単純再帰ネットワーク <*TANJUN SAIKI NETTOWAKU*>	1	simple recursive network	0.021	-	simple recurrent network
		2	simple recursion network	0.018	-	
		3	simple recursive service	0.017	-	
3	統合データベース <*TOUGOU DETABESU*>	1	***integrated database***	0.188	integration data base	integrated database
		2	intermolecular	0.069	synthesis data base	
		3	information database	0.058	fusion data base	
4	統計的機械翻訳 <*TOUKEI TEKI KIKAI HONYAKU*>	1	***statistical machine translation***	0.062	statistic object machine translation	statistical machine translation
		2	statistical method machine translation	0.047	statistic target machine translation	
		3	statistical machine translation system	0.046	statistic aim machine translation	
5	統計的統語解析 <*TOUKEI TEKI TOUGO KAISEKI*>	1	statistical syntactic analysis	0.040	-	statistical parsing
		2	statistical method syntactic analysis	0.033	-	
		3	statistical syntactic structure	0.032	-	
6	反駁 <*HANBAKU*>	1	PAC learning model	0.089	counterblast	refutation
		2	・ F・	0.067	negation	
		3	***refutation***	0.062	rebuttal	
7	ベイズ決定理論 <*BEIZU KETTEI RIRON*>	1	***Bayes decision theory***	0.056	-	Bayes decision theory
		2	unknown datum theory	0.034	-	
		3	Bayesian decision theory	0.034	-	
8	命題様相論理 <*MEIDAI YOUSOU ROMMRI*>	1	proposition modal logic	0.062	proposition aspect logic	propositional modal logic
		2	***propositional modal logic***	0.036	problem aspect logic	
		3	proposition modal	0.032	proposition state logic	

[Note] Bold and Italicized translations were judged as correct.

7. Conclusion

We improved the compositional term translation method with comparable corpora. A bilingual lexicon consisting of word sequence pairs within terms and their correlations is acquired from a document-aligned corpus. The correlations between word sequences in two languages are calculated based on their co-occurrence in aligned document pairs. Then, for an input term, candidate translations are compositionally produced together with their confidence scores, which are defined based on the correlations between the constituents. Thus, the correct translation for the input term can be selected from among as many candidate ones as possible.

An experiment with a comparable corpus consisting of Japanese and English scientific-paper abstracts demonstrated that compositional translation with the corpus-derived bilingual lexicon outperformed that with an ordinary bilingual lexicon. Future work includes the incremental improvement of the bilingual lexicon with correlations, the refinement of the confidence score, and the extension of the compositional translation model to allow word order to be changed.

8. Acknowledgements

We thank the Japan Science and Technology Agency for permitting us to use the JST Japanese and English scientific-paper abstracts. This work was partly supported

by Grant-in-Aid for Scientific Research, MEXT (22300032).

9. References

Abdul-Rauf, Sadaf and Holger Schwenk. 2009. On the use of comparable corpora to improve SMT performance. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, pp. 16-23.

Andrade, Daniel, Tetsuya Nasukawa, and Jun'ichi Tsujii. 2010. Robust measurement and comparison of context similarity for finding translation pairs. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 19-27.

ANLP (Association for Natural Language Processing). 2010. Gengo Shori Gaku Jiten (Encyclopedia of Natural Language Processing). Kyoritsu Publishing Co. (Tokyo).

Baldwin, Timothy and Takaaki Tanaka. 2004. Translation by machine of complex nominals: Getting it right. In *Proceedings of the 2nd ACL Workshop on Multiword Expressions: Integrated Processing*, pages 24-31.

Cao, Yunbo and Hang Li. 2002. Base noun translation using Web data and the EM algorithm. In *Proceedings of the 19h International Conference on Computational Linguistics*, pp. 127-133.

Dagan, Ido and Alon Itai. 1994. Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, vol. 20, No. 4, pp. 563-596.

Fung, Pascale and Percy Cheung. 2004. Mining very-non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and EM. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 57-63.

Fung, Pascale and Lo Yuen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, pp. 414-420.

Grefenstette, Gregory. 1999. The World Wide Web as a resource for example-based machine translation tasks. In *Proceedings of the ASLIB Conference on Translating and the Computer*, Vol. 21.

Ismail, Azniah and Suresh Manandhar. 2010. Bilingual lexicon extraction from comparable corpora using in-domain terms. In *Proceedings of the 23rd International Conference on Computational Linguistics*, Poster Volume, pages 481-489.

JSAI (Japanese Society for Artificial Intelligence). 2008. Jinkou Chiou Gaku Jiten (Encyclopedia of Artificial Intelligence). Kyoritsu Publishing Co. (Tokyo).

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the ACL*, pp. 48-54.

Matsumoto, Yuji and Takehito Utsuro. 2000. Lexical knowledge acquisition. In *R. Dale, H. Moisl, and H. L. Somers (ed.). Handbook of Natural Language Processing*, Ch. 24, pp. 563-610 (Marcel Dekker Inc.).

Morin, Emmanuel and Emmanuel Prochasson. 2011. Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora, ACL 2011*, pages 27-34.

Munteanu, Dragos Stefan and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, vol. 31, No. 4, pp. 477-504.

Och, Franz Josef and Hermann Ney. 2003. A Systematic comparison of various statistical alignment models. *Computational Linguistics*, vol. 29, No. 1, pp. 19-51.

Rapp, Reinhard. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 320-322.

Tanaka, Takaaki. 2002. Measuring the similarity between compound nouns in different languages using non-parallel corpora. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 981-987.

Tonoike, Masatsugu, Mitsuhiro Kida, Toshihiro Takagi, Yasuhiro Sasaki, Takehito Utsuro, and Satoshi Sato. 2006. Comparative Study on Compositional Translation Estimation using a Domain/Topic-Specific Corpus collected from the Web. In *Proceedings of the 2nd International Workshop on Web as Corpus*, pp. 11-18.

Utsuro, Takehito, Takashi Horiuchi, Kohei Hino, Takeshi Hamamoto, and Takeaki Nakayama. 2003. Effect of cross-language IR in bilingual lexicon acquisition from comparable corpora. In *Proceedings of the 10th Conference of the European Chapter of the ACL*, pp. 355-362.

Way, Andy and Nano Gough. 2003. *wEBMT*: Developing and validating an example-based machine translation system using the World Wide Web. *Computational Linguistics*, vol. 29, No. 3, pp. 421-457.