Automatic Extraction of Word Sequence Correspondences in Parallel Corpora

Mihoko Kitamura
Kansai Laboratory
Oki Electric Industry Co., Ltd.
kita@kansai.oki.co.jp

Yuji Matsumoto
Graduate School of Information Science
Nara Institute of Science and Technology
matsu@is.aist-nara.ac.jp

Abstract

This paper proposes a method of finding correspondences of arbitrary length word sequences in aligned parallel corpora of Japanese and English. Translation candidates of word sequences are evaluated by a similarity measure between the sequences defined by the co-occurrence frequency and independent frequency of the word sequences. The similarity measure is an extension of Dice coefficient. An iterative method with gradual threshold lowering is proposed for getting a high quality translation dictionary. The method is tested with parallel corpora of three distinct domains and achieved over 80% accuracy.

1 Introduction

A high quality translation dictionary is indispensable for machine translation systems with good performance, especially for domains of expertise. Such dictionaries are only effectively usable for their own domains, much human labour will be mitigated if such a dictionary is obtained in an automatic way from a set of translation examples.

This paper proposes a method to construct a translation dictionary that consists of not only word pairs but pairs of arbitrary length word sequences of the two languages. All of the pairs are extracted from a parallel corpus of a specific domain. The method is proposed and is evaluated with Japanese-English parallel corpora of three distinct domains.

Several attempts have been made for similar purposes, but with different settings. (see [Kupiec 93][Kumano & Hirakawa 94][Smadja 96])

Kupiec and Kumano & Hirakawa propose a method of obtaining translation patterns of noun compound from bilingual corpora. Kumano & Hirakawa stand on a different setting from the other works in that they assume ordinary bilingual dictionary and use non-parallel (non-aligned) corpora. Their target is to find correspondences not only of word level but of noun phrases and unknown words. However, the target noun phrases and unknown words are decided in the preprocessing stage.

Brown et al. use a probabilistic measure for estimating word similarity of two languages in their statistical approach of language translation [Brown 88]. In their work of aligning of parallel texts, Kay & Röscheisen used the Dice coefficient as the word similarity for insuring sentence level correspondence [Kay & Röscheisen 93].

Kitamura & Matsumoto use the same measure to calculate word similarity in their work of extraction of translation patterns. The similarity measure is used as the basis of their structural matching of parallel sentences so as to extract structural translation patterns. In texts of expertise a number of word sequence correspondences, not word-word correspondences, are abundant especially in the form of noun compounds or of fixed phrases, which are keys for better performance. Though the method proposed in this paper deals only with consecutive sequences of words and is intended to provide a better base for the structural matching that follows, the results themselves show very useful and informative translation patterns for the domain.

Our method extends the usage of the Dice coefficient in two ways: It deals not only with correspondence between the words but with correspondence between word-sequences, and it modifies the formula measure so that more plausible corresponding pairs are identified earlier.

2 Related Work and Some Results

Brown et. al., used mutual information to construct corresponding pairs of French and English words. A French word f is considered to be translated into English word e_j that gives the maximum mutual information:

$$MI(e_j, f) = \log \frac{P(e_j \mid f)}{P(e_j)}$$

Probabilities $P(e_j \mid f)$ and $P(e_j)$ are calculated from parallel corpus by counting the occurrences and co-occurrences of e_j and f.

Kay & Röscheisen used the following Dice coefficient for calculating the similarity between English word w_e and French word w_f. In the formula, $f(w_e)$, $f(w_f)$ represent the numbers of occurrences of w_e and w_f, and $f(w_e, w_f)$ is the number of simultaneous occurrences of those words in corresponding sentences.

$$sim(w_e, w_f) = \frac{2f(w_e, w_f)}{f(w_e) + f(w_f)}$$

Kitamura & Matsumoto used the same formula for calculating word similarity from Japanese-English parallel corpora. A comparison between the above two method is done on a parallel corpus and the results are reported in [Ohmori 96]. They applied both approaches to a French-English corpus of about one thousand sentence pairs. The results are shown in Table 1 where the correctness is checked by human inspection. Since both methods show very inaccurate results for the words of one occurrence, only the words of two or more occurrences are selected for inspection. Table 1 shows the proportion that a French word is paired with the correct English words checked with the top, three and five highest candidates.

	Num. of words	1st candidate	within best 3	within best 5
Mutual Information	697	43.6%	60.0%	65.4%
Dice coefficient	574	46.2%	65.0%	66.5%

Table 1: Comparison of Mutual Information and Dice coefficient

The results show that though Dice coefficient gives a slightly better correctness both methods do not generate satisfactory translation pairs.

[Kupiec 93] and [Kumano & Hirakawa 94] broaden the target to correspondences between word sequences such as compound nouns. Kupiec uses NP recognizer for both English and French and proposed a method to calculate the probabilities of correspondences using an iterative algorithm like the EM algorithm. He reports that in one hundred highest ranking correspondences ninety of them were correct. Although the NP recognizers detect about 5000 distinct noun phrases in both languages, the correctness ratio of the total data is not reported.

Kumono & Hirakawa's objective is to obtain English translation of Japanese compound nouns (noun sequences) and unknown words using a statistical method similar to Brown's together with an ordinary Japanese-English dictionary. Japanese compound nouns and unknown words are detected by the morphological analysis stage and are determined before the later processes. Though they assume unaligned Japanese-English parallel corpora, alignment is performed beforehand. In an experiment with two thousand sentence pairs, 72.9% correctness is achieved by the best correspondences and 83.8% correctness by the top three candidates in the case of compound nouns. The correctness ratios for unknown words are 54.0% and 65.0% respectively.

Smadja proposes a method of finding translation patterns of continuous as well as discontinuous collocations between English and French [Smadja 96]. The method first extracts meaningful collocations in the source language(English) in advance by the XTRACT system. Then, aligned corpora are statistically analized for finding the corresponding collocation patterns in the target language(French). To avoid possible combinational explosion, some heuristics is introduced to filter implausible correspondences.

Getting translation pairs of complex expression is of great importance especially for technical domains where most domain specific terminologies appear as complex nouns. There are still a

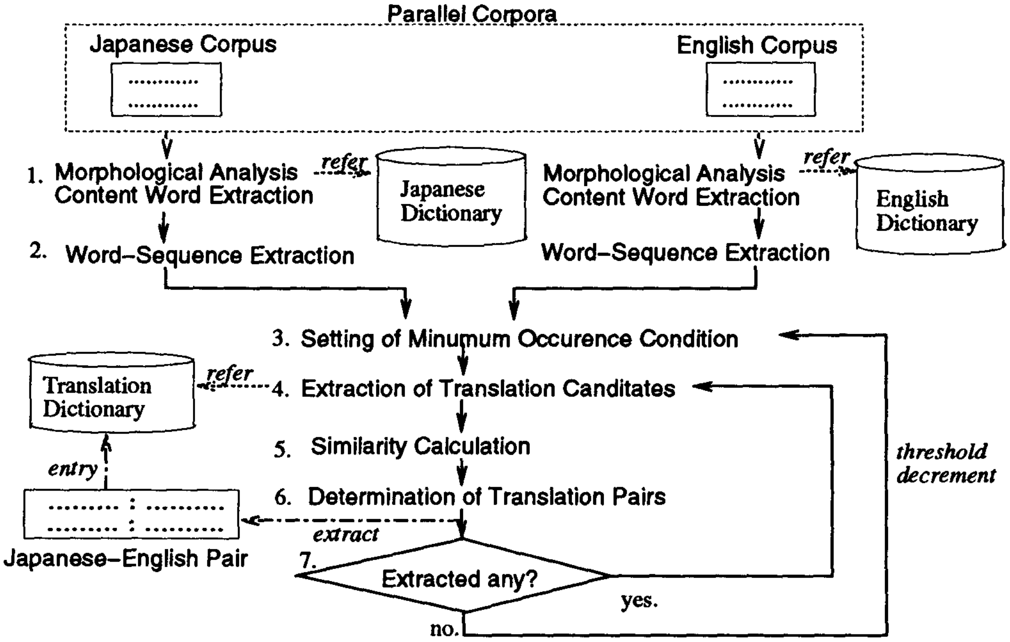

Figure 1: The flow of finding the correspondences of word sequences

number of other interesting and meaningful expression that should be translated in a specific way. We propose a method of finding corresponding translation pairs of arbitrary length word sequences appearing in parallel corpora and an algorithm that gradually produces "good" correspondences earlier so as to reduce noises when extracting less plausible correspondences.

3 Overview of the Method

Figure 1 shows the flow of the process to find the correspondences of Japanese and English word sequences. Both Japanese and English texts are analyzed morphologically.

We make use of two types of co-occurrences: Word co-occurrences within each language corpus and corresponding co-occurrences of those in the parallel corpus. In the current setting, all words and word sequences of two or more occurrences are taken into account. Since frequent co-occurrence suggests higher plausibility of correspondence, we set a similarity measure that takes co-occurrence frequencies into consideration. Deciding the similarity measure in this way reduces the computational overhead in the later processes. If every possible correspondence of word sequences is to be calculated, the combination is large. Since high similarity value is supported by high co-occurrence frequency, a gradual strategy can be taken by setting a threshold value for the similarity and by iteratively lowering it. Though our method does not assume any bilingual dictionary in advance, once words or word sequences are identified in an earlier stage, they are regarded as decisive entries of the translation dictionary. Such translation pairs are taken away from the co-occurrence data, then only the remaining word sequences need be taken into consideration in the subsequent iterative steps. Next section describes the details of the algorithm.

4 The Algorithm

The step numbering of the following procedure corresponds to the numbers appearing in Figure 1. In the current implementation, the Translation Dictionary is empty at the beginning. Steps 1 and 2 are performed on each language corpus separately.

1. Japanese and English texts are analyzed morphologically and all content words (nouns, verbs, adjectives and adverbs) are identified.

2. All content words of two or more occurrences are extracted. Then, word sequences of length two that are headed by a previously extracted word are extracted, provided they appear at least twice in the corpus. In the same way, a word sequence w of length $i + 1$ is taken into consideration only when its prefix of length i has been extracted and w appears at least twice in the corpus. This process is repeated until no new word sequences are obtained. The subsequent steps handle only those extracted word sequences. It would be natural to set a maximum length for the candidate word sequences, which we really have it be between 5 and 10 in the experiments.
3. A threshold for minimum frequency of occurrence (f_{min}) is decided, and the following process is repeated, every time decrementing the threshold by some extent.
4. For the word sequence occurring more than f_{min} times, the numbers of total occurrence and total bilingual co-occurrence are counted. This is done for all the pairs of such Japanese and English word sequences. It is not the case for a pair that already appeared in the Translation Dictionary.
5. For each pair of bilingual word sequences, the following similarity value ($sim(w_J, w_E)$) is calculated, where w_J and w_E are Japanese and English word sequences, and f_j, f_e and f_{je} are the total frequency of w_J in the Japanese corpus, that of w_E in the English corpus and the total co-occurrence frequency of w_J and w_E appearing in corresponding sentences.

   $$sim(w_J, w_E) = (\log_2 f_{je})\frac{2f_{je}}{f_j + f_e}$$

   This formula is a modification of the Dice coefficient, weighting their similarity measure by logarithm of the pair's co-occurrence frequency. Only the pairs with their $sim(w_J, w_E)$ value greater than $\log_2 f_{min}$ are considered in this step. The fact that no word sequence occurring less than f_{min} times cannot yield greater similarity value than $\log_2 f_{min}$ assures that all pairs of word sequences with the occurrence more than f_{min} times are surely taken into consideration.
6. The most plausible correspondences are then identified using the similarity values so calculated:
   (a) For an English word sequence w_E, let $WJ = \{w_{J1}, w_{J2}, \cdots, w_{Jn}\}$ be the set of all Japanese word sequences such that $sim(w_{Ji}, w_E) \geq \log_2 f_{min}$. The set is called the candidate set for w_E. For each Japanese word sequence w_J its candidate set is constructed in the same way.
   (b) Of the candidate set WJ for w_E, if the candidate with the highest similarity value ($w_{Ji} = arg \max_{w_{Jk} \in WJ} sim(w_{Jk}, w_E)$) again selects w_E as the candidate with the highest similarity ($w_E = arg \max_{w_{Em} \in WE} sim(w_{Ji}, w_{Em})$), where WE is the candidate set for w_{Ji}, the pair $\langle w_{Ji}, w_E \rangle$ is regarded as a translation pair.
7. The approved translation pairs are registered in the Translation Dictionary until no new pair is obtained, then the threshold value f_{min} is lowered, and the steps 4 through 6 are repeated until f_{min} reaches a predetermined value.

5 Experiments of Translation Pair Extraction

5.1 The settings

We used parallel corpora of three distinct domains: (1) a computer manual (9,792 sentence pairs), (2) a scientific journal (12,200 sentence pairs), and (3) business contract letters (10,016 sentence pairs). All the Japanese and English sentences are aligned and morphologically analyzed[1].

The settings of the experiments are as follows: The maximum length of the extracted word sequences is set at 10. The initial value of f_{min} is set at the half of the highest number of occurrences of extracted word sequences and is lowered by dividing by two until it reaches to or under 10, then it is lowered by one in each iteration until 2.

[1] Japanese and English morphological analyzers of Machine Translation System PENSÉE were used. PENSÉE is a trademark of Osaka Gas corporation, OGIS-RI, and Oki Electric Industry Co.,Ltd.

	single word				word seq.	
	total		occurrence ≥ 2		occurrence ≥ 2	
	Eng.	Jap.	Eng.	Jap.	Eng.	Jap.
Business	2,300	3,739	2,218	3,568	73,026	72,574
Science	7,254	9,415	6,764	8,856	16,555	24,998
Manual	3,701	4,926	3,478	4,799	32,049	38,796

Table 2: Numbers of extracted words and word sequences

threshold	Num. of pairs	correct	near miss	incorrect	correctness (+near) %	accumulative correctness %
1151	2	2	0	0	100(100)	100(100)
575	3	3	0	0	100(100)	100(100)
287	4	4	0	0	100(100)	100(100)
143	12	12	0	0	100(100)	100(100)
71	19	18	1	0	94.7(100)	97.5(100)
35	48	48	0	0	100(100)	98.9(100)
17	103	101	2	0	98.1(100)	99.0(100)
10	164	155	8	1	94.5(99.4)	96.6(99.7)
9	53	51	2	0	96.2(100)	96.6(99.8)
8	67	63	4	0	94.0(100)	96.2(99.8)
7	82	75	6	1	91.5(98.8)	95.5(99.6)
6	134	114	20	0	85.1(100)	93.5(99.7)
5	163	145	15	3	89.0(98.2)	92.6(99.4)
4	318	257	50	11	80.8(96.5)	89.4(98.6)
3	755	502	195	59	66.5(92.2)	80.4(96.1)
total	1,927	1,550	302	75	80.5(96.1)	—

Table 3: Results of Business Letter

Table 2 summarizes the numbers of word sequences extracted by Step 2. For each corpus the table shows the numbers of distinct content words, those of two or more occurrences, and the numbers of word sequences of two or more occurrences.

5.2 The results

Tables 3, 4 and 5 shows the statistics obtained from the experiments. The columns specify the numbers of approved translation pairs. The correctness of the translation pairs are checked by a human inspector. A "near miss" means that the pair is not perfectly correct but some parts of the pair constitute the correct translation.

It is noticeable that the pairs with high frequencies give very accurate translation in the cases of the computer manual and the business letters, whereas the scientific journal does not necessarily gives high accuracy to highly frequent pairs. The reason is that the former two corpora are really in a homogeneous domain, while the corpus of scientific journal is a complex of distinct scientific fields. The former two corpora reveal a worse performance with the pairs with low frequency threshold. This is because those corpora frequently contain a number of lengthy fixed expression or particular collocations. One such example is that "p type (silicon)" frequently collocates with "n type (silicon)," making the correspondence uncertain.

The science journal shows a stable accuracy of translation pair extraction. The accuracy exceeds 90% in most of the stages. The reason would be that scientific papers do not repeat fixed expression and the terminologies are used not in a fixed way.

Table 6 summarizes the combination of the length of English and Japanese word sequences. The fraction in each entry shows the number of correct pairs over the number of extracted pairs. This table indicates that translation pairs of lengthy or unbalanced sequences are safely regarded

threshold	Num. of pairs	correct	near miss	incorrect	correctness (+near) %	accumulative correctness %
68	1	1	0	0	100(100)	100(100)
34	21	19	1	1	90.5(95.2)	90.9(95.5)
17	69	64	5	0	92.8(100)	92.3(98.9)
10	142	133	8	1	93.7(99.3)	93.1(99.1)
9	52	49	3	1	94.2(98.1)	93.3(97.9)
8	69	69	0	0	100(100)	94.6(99.2)
7	66	63	2	1	95.5(98.5)	94.7(99.0)
6	105	99	6	0	94.3(100)	94.7(99.2)
5	168	155	12	1	92.3(98.8)	94.1(99.1)
4	292	263	25	4	90.1(98.6)	92.9(99.0)
3	536	494	34	8	92.2(97.2)	92.6(98.4)
2	1,307(500)	(445)	(46)	(9)	89.0(97.4)	91.7(98.1)
total	2,828(2,021)	(1,854)	(129)	(38)	91.7(98.1)	—

Table 4: Results of Science Journal

threshold	Num. of pairs	correct	near miss	incorrect	correctness (+near) %	accumulative correctness %
209	1	1	0	0	100(100)	100(100)
104	4	4	0	0	100(100)	100(100)
52	19	19	0	0	100(100)	100(100)
26	55	54	0	1	98.1(98.1)	98.7(98.7)
13	145	140	5	0	96.6(100)	97.3(99.6)
10	81	76	5	0	93.8(100)	96.4(99.7)
9	58	55	2	1	94.8(98.3)	96.1(99.4)
8	75	68	5	2	90.7(93.6)	95.2(99.1)
7	106	99	7	0	93.4(100)	94.9(99.3)
6	126	118	7	1	93.7(99.2)	94.6(99.3)
5	214	198	13	3	92.5(98.6)	94.1(99.1)
4	367	330	26	11	89.9(97.0)	92.9(98.5)
3	629	519	97	13	82.5(97.9)	89.4(98.3)
2	1,401(500)	(395)	(87)	(18)	79.0(96.4)	87.2(97.9)
total	3,281(2,380)	(2,076)	(254)	(50)	87.2(97.9)	—

Table 5: Results of Computer Manual

Business		Length of Eng. Seq.									
Letters		1	2	3	4	5	6	7	8	9	10
	1	823/843	43/58	0/6	0/1	0	0	0	0	0	0
	2	32/45	401/450	17/55	1/23	0/5	0/4	0/1	0	0/1	0
	3	0	79/122	72/90	7/23	0/8	0/4	0/4	0	0	0
Length	4	0	6/21	29/45	15/23	2/5	1/2	0/1	0	0	0/1
of	5	0	3/10	2/13	7/14	3/10	2/3	0	0/1	0/1	0/1
Jap.	6	0	0	2/4	2/3	0/1	0/2	0	0/1	0/1	0/2
Seq.	7	0	0/1	0	0/2	0	0/1	0	0/1	0	0
	8	0	0/1	0	0/1	0/1	0/1	0	0	0	0/1
	9	0	0	0	0	0	0	0	0	0	0
	10	0	0/1	0/1	0	0	1/1	0	0	0	0/6

Table 6: Length Combination of Word Sequences and their Accuracy (Business Letter)

	Japanese	English	Similarity
— 1. Business Letter —			
	独占 実施 権	exclusive license	4.95
△	紛争 (、) 論争 (又は) 意見	dispute(,) controversy (or) difference (which may) arise	4.34
	営業 秘密	trade secret	3.72
	契約 (の) 発効 日	effective date (of this) agreement	3.12
	営業 時間	business hour	2.92
	特許 実用新案 商標 意匠 著作権	utility model(,) trademark(,) design (or) copyright	2.81
	確認 (付) 取消 不能 信用状	irrevocable confirm(ed) letter (of) credit	2.81
	技術 製造 ノウハウ	technique manufacture know-how	2.62
	特許 (、) ノウハウ (又は) 技術 情報	patent(s)(,) know-how (or) technical information	1.06
— 2. Science Journal —			
	出血 熱 ウイルス	hemorrhage fever virus	3.19
	アクリル 酸 メチル	methyl acrylate	3.17
	ロスアラモス 国立 研究所	Los Alamos national laboratory	2
	n 型 シリコーン	n type	1.78
*	n 型 シリコーン	p type	1.78
	カリフォルニア 大学 デービス 校	university (of) California (at) Davis	1.58
	ワイヤレス ネットワーク	wireless network	1.19
	光 ファイバー ネットワーク	fiber(-)optic network	1.19
	出血 熱	hemorrhage fever	1.14
— 3. Computer Manual —			
	インタネット	internet	5.25
	インタネット アドレス	internet address	2.83
	倍精度 浮動 小数点	double precision float point	1.79
	インタネット プロトコル ＩＰ	internet protocol IP	1.78
	インタネット プロトコル	internet protocol	1.66
*	ホスト テーブル	DoD internet	1.6
	ネーム トゥ アドレス マッピング	name (to) address map(ping)	1.58
	インタネット サービス	internet service	1.45

△ indicates "near miss" and * indicates "incorrect".

Table 7: Samples of Corresponding Word Sequences

as incorrect correspondences.

Tables 7 and 8 list samples of translation pairs extracted from the experiments. Table 7 lists some of typical word sequence pairs. Many of Japanese translation of English technical terms are automatically detected. Table 8 lists the top 30 pairs from the experiment on the business contract letters.

The method is capable of getting interesting translation patterns. For example, "営業秘密" and "営業時間" are found to correspond to "trade secret" and "business hour" respectively. Note that Japanese word "営業" is translated into different English words according to their occurrences with distinct word.

Table 9 shows the recall ratio based on the results of the experiments. The figures show the numbers of words that are included at least one extracted translation pairs. The recall rates are shown in parentheses, which indicates how much proportion of the words with two or more occurrences in the corpora are finally participated in at least one translation pair. The major reason that the recall is not sufficiently high is that we decided to use a rather severe condition on selecting a translation pairs in Step 6 in the algorithm. The condition may be loosen to get better recall ratio though we may lose high precision. We have not yet tested our method with other conditions.

Japanese	English	Similarity	Pair.Freq.	Jap.Freq.	Eng.Freq.
— Freq.Stage 1151 —					
会社	company	10.73	3952	4081	4720
ライセンシー	licensee	10.47	2436	2521	2715
— Freq.Stage 575 —					
販売店	distributor	9.55	1471	1562	1679
契約 品	product	9.26	2511	2996	3127
売り手	seller	9.24	999	1039	1116
— Freq.Stage 287 —					
買い手	buyer	8.92	940	970	1112
当事者	party	8.84	1276	1394	1584
書面	writing	8.39	754	860	858
条	article	8.34	778	837	955
— Freq.Stage 143 —					
b	b	8.07	332	345	344
a	a	8.01	324	335	340
ABC	ABC	7.99	354	362	388
情報	information	7.87	489	549	561
XYZ	XYZ	7.77	327	333	370
特許	patent	7.65	455	545	505
技術	technical	7.64	520	558	670
権利	right	7.60	664	869	769
商標	trademark	7.50	369	401	438
c	c	7.41	218	231	226
地域	territory	7.26	668	693	1033
必要	necessary	7.26	332	356	410
— Freq.Stage 71 —					
技術 情報	technical information	7.08	214	227	241
△ 受託	consignee	6.99	225	244	259
ロイヤルティ	royalty	6.84	295	377	331
△ 契約 以下	hereinafter	6.82	198	223	220
d	d	6.76	126	130	130
販売	sale	6.75	626	804	920
独占	exclusive	6.74	235	278	271
製造	manufacture	6.72	578	930	648
義務	obligation	6.59	228	255	287
:	:	:	:	:	:

△ indicates "near miss".

Table 8: Sample of Top Correspondences (Business Letter(Best 30))

Corpus	English	(recall)	Japanese	(recall)
Business	867	(39.1%)	1,005	(28.2%)
Science	2,240	(33.1%)	2,359	(26.6%)
Manual	1,922	(55.3%)	2,224	(46.3%)

Table 9: Numbers of words identified

6 Conclusion

A method for obtaining translation dictionary from parallel corpora was proposed, in which not only word-word correspondences but arbitrary length word sequence correspondences are extracted. This work is originally motivated for the purpose of improving the performance of our translation pattern extraction from parallel corpora [Kitamura & Matsumoto 95], in which translation patterns are extracted by syntactically analyzing both Japanese and English sentences and by structurally matching them. Some discrepancy is caused by poor quality of translation dictionary. This is why we tried to pursue a way to obtain better translation dictionary from parallel corpora. We believe that the proposed method gives results of good performance compared with previous related work. The translation pairs obtained through our method are directly usable as the base resource for MT systems based on *translation memory* [Lehmann 95].

We hope to acquire better translation patterns by combining the current results with our work of structural matching for finding out fine grained correspondence.

References

P.F. Brown. A Statistical Approach to Language Translation. In *COLING-88*, volume 1, pages 71–76, 1988.

M. Kay and M. Röscheisen. Text-Translation Alignment. *Computational Linguistics*, 19(1):121–142, 1993.

M. Kitamura and Y. Matsumoto. A Machine Translation System based ón Translation Rules Acquired from Parallel Corpora. In *Recent Advances in Natural Language Processing*, pages 27–44, 1995.

A. Kumano and H. Hirakawa. Building an MT Dictionary from Parallel Texts Based on Linguistic and Statistical Information. In *COLING-94*, volume 1, pages 76–81, 1994.

J. Kupiec. An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora. In *31st Annual Meeting of the Association for Computational Linguistics * Proceedings of the Conference (ACL93)*, pages 23-30, 1993.

H. Lehmann. Machine Translation for Home and Business Users. In *Proceedings of MT Summit V*,1995

K. Ohmori, J. Tsutsumi, and M. Nakanishi. Building Bilingual Word Dictionary Based on Statistical Information . In *Proceedings of The Second Annual Meeting of The Association for Natural Language Processing*, pages 49–52, 1996. *(in Japanese)*

F. Smadja, K.R. McKeown and V. Hatzivassiloglou. Translating Collocations for Bilingual Lexicons: A Statistical Approach. *Computational Linguistics*, 22(1):1–38, 1996.