Extracting Low-frequency Translation Pairs from Japanese-English Bilingual Corpora

Keita TSUJI
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku,
Tokyo 101-8430,
JAPAN
+81-3-4212-2571
keita@nii.ac.jp

Kyo KAGEURA
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku,
Tokyo 101-8430,
JAPAN
+81-3-4212-2518
kyo@nii.ac.jp

Abstract

We propose a method for extracting low-frequency translation pairs from Japanese-English bilingual corpora. Many methods have been proposed for extracting translation pairs from bilingual corpora, but most are based on word frequency and are, therefore, not effective in extracting low-frequency pairs. In Japanese-English corpora, many low-frequency translation pairs are loan-word pairs that can be extracted based on transliteration patterns. Our combined method, which relies on both transliteration and word frequency, performed significantly better than methods utilizing word frequency alone. Our method achieved 80% precision at 84% recall against the given corpus, while the word-frequency model achieved just 80% precision at 8% recall.

1 Introduction

In NLP application areas such as machine translation and cross-language information retrieval, a range of translation word pairs are needed. While popular and well-known translation pairs are included in existing bilingual dictionaries, newly coined and minor translation pairs are not yet covered in any resource. Therefore, it would be very useful if we could extract these pairs from proper bilingual corpora. Because newly coined translation pairs have just begun to appear in text and because of the nature of minor translation pairs, the frequencies of both are necessarily low. From this point of view, we aim to develop a method for extracting low-frequency translation pairs from Japanese-English bilingual corpora.

So far, there have been many methods proposed for extracting translation pairs from bilingual corpora. The most typical and intensively studied are word-frequency-based methods that extract word W_J and $W_{E\circ}$ as translation if W_J and $W_{E\circ}$ co-occur in numerous bilingual sentence pairs. Although frequency-based methods have a strong theoretical basis and are often sufficiently effective, they have one weak point, i.e., they are not good at extracting low-frequency pairs. For instance, if W_J and $W_{E\circ}$ occur and co-occur in just one bilingual sentence pair, and another word, $W_{E\times}$, occurs in one and the same sentence, the frequency-based method cannot determine which – $W_{E\circ}$ or $W_{E\times}$ – is the correct translation of W_J.

It is therefore desirable to develop a method for extracting low-frequency translation pairs, filling the gaps in, as well as making good use of, the frequency-based method. Language-pair-independent methods such as the frequency-based method are adequately refined, and we believe that incorporating language-pair-dependent knowledge, which is often ignored, is a step in the right direction. In this, among the first studies based on this framework, we focus on the Japanese and English language pair, although we take special care to keep the method as open to generalization as possible.

In Japanese-English bilingual corpora, many low-frequency translation pairs are loan-word pairs. Although these pairs can be extracted based on some transliteration patterns (without relying on word frequency) and there has been some research related to transliterations (Knight and Graehl, 1997)(Fujii and Ishikawa, 2001), few studies have been conducted for extracting these pairs from bilingual corpora based on transliteration. Two tasks arise: (1) To develop an effective method for extracting transliterated word pairs that include many of the low-frequency Japanese-English translation pairs; and (2) To make good use of the frequency-based method to extract low-frequency and especially non-transliterational translation pairs. Solving these tasks, we propose a combined method that utilizes both transliteration and word frequency.

In the following sections, we will first give the results of an investigation into low-frequency

translation pairs, next describe our method, and then show that significant improvement was observed by our proposed method.

2 Preliminary Investigation

We used one bilingual dictionary and four bilingual corpora for our investigation and experiment. The bilingual dictionary we used was EDICT, which contained 102,380 pairs of Japanese-English translation pairs. We regard EDICT as one example of existing bilingual dictionaries.

The bilingual corpora we used consist of 9,000 pairs of abstracts of academic papers in the fields of (1) information processing field and (2) architecture collected by the National Institute of Informatics of Japan. Each Japanese-English abstract was written by the respective authors of each paper, and there is no strict correspondence between them. We assume that corpora which are strictly correspondent at each sentence level are still hard to obtain, while loosely correspondent ones become more readily available. [1] We chose these abstract corpora to keep our method realistic and applicable to many fields. Also, consequently, we do not try to align sentences in each abstract. Instead, we defined each whole abstract as a segment from which we extracted translation pairs. We also used 9,000 bilingual titles in the above fields for comparison. The basic quantities are shown in Table 1. [2] 'Inf-Abst' and 'Arc-Abst' represent the abstract corpus of information processing and architecture, respectively. 'Inf-Titl' and 'Arc-Titl' represent the title corpus in the same manner.

We define '$min(W_J, W_E)$' of the word pair W_J and W_E as the smaller of the number of segments in which the word W_J occurred and in which W_E occurred. For instance, if W_J and W_E occurred in four and six segments, respectively, '$min(W_J, W_E)$' of that pair is four. In addition, we represent the translation pairs whose $min(W_J, W_E)$ is N as '$min(W_J, W_E) = N$ *translation pairs*'.

We randomly selected 1,000 segments of each corpus and manually identified the translation pairs which should be extracted in each segment. In our estraction experiment, we evaluated the result based on these pairs. The number of these pairs are shown in the column 'Pair' in Table 5. For instance, we can see in Table 5 that there are 548 $min(W_J, W_E) = 1$ *translation pairs* in the information processing abstract corpus. Note that their frequencies were counted based on 9,000 segments, not on 1,000 segments.

	Chasen Words		English Words	
	Token	Type	Token	Type
Inf-Abst	1,389,473	23,072	957,467	34,908
Arc-Abst	1,263,833	23,758	695,542	40,244
Inf-Titl	101,421	7,312	83,357	12,843
Arc-Titl	181,744	9,125	156,479	18,625

Table 1: Basic Quantities of Four Corpora

2.1 Existing Dictionary and Low-frequency Pairs in the Corpora

We investigated what percentages of the single word noun translation pairs in the corpora are listed in EDICT. The results are shown in the column 'R_Dict' in Table 5. Generally speaking, the low-frequency translation pairs (i.e., those whose $min(W_J, W_E)$ are small) are not listed in EDICT. For instance, only 11.78% of $min(W_J, W_E) = 1$ *translation pairs* in the architecture abstract corpus are listed in EDICT. With this fact, we would like to say that the existing bilingual dictionaries do not contain low-frequency translation pairs.

2.2 Problems With the Frequency-based Method for Extracting Low-frequency Pairs

We define 'full co-occurrence' of a word A with a translation pair (X, Y) as a situation in which the word A and X always co-occur in the same segment. There are two types of full co-occurrence: (a) A and X belong to the same language; and (b) A and X belong to different languages.

When type-(a) full co-occurrence as in Table 2 occurs (i.e., A='ノード', (X, Y)=('経路','path')), the frequency-based method estimates that '経路' and 'ノード' are equally likely to be the translation of 'path'. When type-(b) full co-occurrence as in Table 3 occurs (i.e., A='node', (X, Y)=('経路','path')), the frequency-based method estimates that 'node' is more likely to be the translation of '経路' than 'path'. In situations of type-(a), the method

[1] For instance, (Omae et al, 2003) obtained loosely correspondent bilingual texts from the Web by submitting bilingual equivalent keywords to Google and extracting translation pairs from the resultant texts.

[2] The plural forms of English words were converted into singular forms. ChaSen 2.0b and Brill tagger were used for the Japanese and English texts, respectively.

cannot determine which is the correct translation, and in situations of type-(b), the method chooses the wrong word as the translation.

Full co-occurrence is related to 'indirect association' (Melamed, 2000). But this only refers to situations in which word A 'often' – as opposed to 'fully' – co-occurs with X.

Table 5 shows how full co-occurrence can pose a difficult obstacle for the frequency-based method when extracting low-frequency translation pairs. In Table 5, 'R_Full_Cooc' is the ratio (%) of translation pairs that are fully co-occurred by other words against the total translation pairs. For instance, among 314 $min(W_J, W_E) = 1$ *translation pairs* in the architecture abstract corpus, 90.13% are fully co-occurred by other words.

Japanese Part	English Part
	path
経路, ノード	path
経路, ノード	path
経路, ノード	path
	path

Table 2: Full Co-occurrence of Type-(a)

Japanese Part	English Part
	path
経路	path, node
経路	path, node
経路	path, node
	path

Table 3: Full Co-occurrence of Type-(b)

2.3 Loan-word Pairs Among Low-frequency Pairs

The character types of translation pairs are also shown in Table 5. 'Katak', 'Roman', 'Kanji' and 'Other' represents the ratio (%) of the translation pairs whose Japanese words are represented using katakana characters, Roman alphabets, kanji characters and other characters, respectively. In the Japanese language, most of the transliterated loan words are written in katakana or the Roman alphabet, and vice versa. Table 5 shows that the smaller the $min(W_J, W_E)$ becomes, the larger the ratio of the loan-word pairs becomes. The loan-word pairs can be extracted based on some transliteration patterns without relying on word frequency.

3 The Extraction Method

Our preliminary investigations revealed: (1) that most of the low-frequency translation pairs are in the situation of full co-occurrence (as mentioned in Section 2.2) and cannot be extracted based on word frequency alone; and (2) that many of them are transliterated word pairs. Our method is based on these observations; it is as follows:

1) From each segment, extract the translation pairs based on transliteration. For instance, if there are candidate words '経路 (path)' and 'ノード (node)' in the Japanese part of the segment and there are 'path' and 'node' in the English part of the segment, 'ノード' and 'node' are extracted based on transliteration patterns such as 'ノ'='no' and 'ド'='de'.
2) Remove the extracted pairs from the candidate words. As a result, '経路' and 'path' are left in the segment.
3) Based on word frequency, extract translation pairs from the candidate words left in the segment.

Figure 1 shows the flow of our method.

Extraction methods like Steps 1 and 3 used to be studied independently. For instance, (Gale and Church, 1991)(Kupiec, 1993)(Smadja et al., 1996)(Hiemstra, 1997)(Kitamura and Matsumoto, 1996)(Ahrenberg et al., 1998)(Melamed, 2000) focused on word frequency and (Collier et al., 1997)(Jeong et al., 1999) extracted loan-word pairs based on transliteration patterns.

The unique point of our method in a technical sense is that we incorporated Steps 1 and 2 before applying the frequency-based method. These steps are the key to resolving the issue of full co-occurrence. We expect that Step 1 identifies the correct translation of A in Section 2.2 and Step 2 removes A with that correct translation. After these steps, the full co-occurrence by A is resolved and the frequency-based method regains its power to extract translation pairs (X, Y) that remain in the segment.

While many effective frequency-based methods have been proposed, there were only a few transliteration-based methods especially for the Japanese-English language pair. Against this

background, the first author of the present paper developed an effective transliteration-based method (Tsuji, 2002). We adopted it for Step 1. As for Step 3, we adopted the method of (Melamed, 2000) and that of (Hiemstra, 1997). These were chosen because they seem to be the most sophisticated frequency-based methods at present. Note that they are just examples and other similar methods can be adopted as modules in Steps 1 and 3.

One technical problem is that if we extract and remove too many pairs at Steps 1 and 2, we are to wrongly extract Japanese-English word pairs as translations and pass insufficient candidates to Step 3. We assume that there exists optimal threshold for Steps 1 and 2 which can be determined empirically through our experiment.

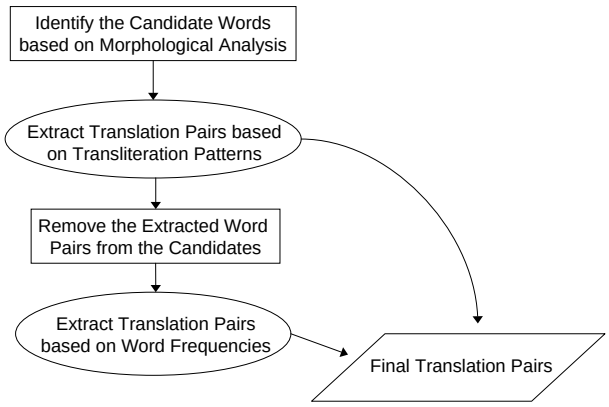

Figure 1: Flow of Extracting Translation Pairs

3.1 The Transliteration-based Method

Our method uses transliteration patterns that are observed in actual loan-word pairs (obtained from bilingual dictionaries, etc.). These patterns are obtained by breaking down the Japanese word into mora units (Japanese syllables) and manually identifying the corresponding character strings in the counterpart English word. Based on these patterns, we extract the loan-word pairs as follows:

(1) Decompose Japanese word W_J into mora units.
(2) Using all the existing transliteration patterns, generate all the back-transliteration candidates for W_J (henceforth, we represent the i-th candidate as $M_i(W_J)$).
(3) Pick up the English word W_E that co-occurred with W_J and identify the longest common subsequence with $M_i(W_J)$ ($\equiv C(M_i(W_J), W_E)$).
(4) If the following D exceeds a certain threshold, extract the pair W_J and W_E as translation (where $N(W)$ represents the number of characters of string W).

$$D = max_i \frac{N(C(M_i(W_J), W_E)) \times 2}{N(M_i(W_J)) + N(W_E)}$$

If we have transliteration patterns like in Table 4 and W_J is 'グラフ', $3 \times 5 \times 4$ back-transliteration candidates such as <graf>, <graph>, ..., <gulerfe> are obtained. And if W_E is 'graph':

$$\begin{aligned} D &= max\Big(\frac{N(C(\text{graf}, \text{graph})) \times 2}{N(\text{graf}) + N(\text{graph})}, \\ &\quad \frac{N(C(\text{graph}, \text{graph})) \times 2}{N(\text{graph}) + N(\text{graph})}, \\ &\quad ... \\ &\quad \frac{N(C(\text{gulerfe}, \text{graph})) \times 2}{N(\text{gulerfe}) + N(\text{graph})} \Big) \\ &= max(0.67, 1.00, ..., 0.33) \\ &= 1.00 \end{aligned}$$

It indicates that 'グラフ' and 'graph' are very likely to be translation. This method is more effective than that of (Collier et al., 1997). The main reason is that the transliteration patterns are more realistic and we take into account the length of $M_i(W_J)$.

グ	g	gue	gu		
ラ	ra	ru	la	lu	ler
フ	f	ph	ff	fe	

Table 4: Transliteration Patterns

3.2 The Frequency-based Method

The method of (Melamed, 2000) and that of (Hiemstra, 1997) link the word W_J and W_E as translation based on iterative algorithms. We extracted a word pair as translation if its score P was high:

$$P = \log(L(W_J, W_E) + 1) \frac{L(W_J, W_E) \times 2}{L(W_J, .) + L(., W_E)}$$

where $L(W_J, W_E)$ is the number of links between W_J and W_E at the convergence stage.

$L(W_J, .)$ is the sum of the number of links between W_J and all the other words. $L(., W_E)$ is defined as the same.[3]

4 Extraction Experiment

Two kinds of extraction experiments were performed. The targets of extraction were $min(W_J, W_E) = 1$ *translation pairs*, which we think are difficult to extract but believe to be valuable pairs in the sense that they include the most newly coined translation pairs.

As we previously-mentioned, we evaluated the result based on the randomly-selected 1,000 segments in the corpus. From the extracted pairs, we eliminated the pairs that could not be extracted from these 1,000 segments and calculated the precision and recall. We calculated them every time we extracted one pair whose P was highest at Step 3.

(A) Basic Extraction Experiment: We extracted $min(W_J, W_E) = 1$ *translation pairs* and evaluated the result. The threshold D at Steps 1 and 2 were 0.9 or 0.8. This experiment was performed to show the overall effectiveness of our method to extract low-frequency translation pairs.

(B) Non-transliterational Pairs Extraction Experiment: We evaluated the result of extracting $min(W_J, W_E) = 1$ *translation pairs* that were not loan-word pairs. This experiment was performed to show the effectiveness of incorporating Steps 1 and 2 before Step 3. If our method extracted these non-transliterational pairs with higher precision and recall than the frequency-based method alone, it is very likely that part of the issue of full co-occurrence was resolved through Steps 1 and 2. The threshold D at Steps 1 and 2 were 0.9, 0.8 or 0.7.

4.1 Result of the Basic Experiment

The results of the basic experiment against the information processing abstract corpus and the architecture abstract corpus are shown in Figures 2 and 3, respectively. In Figures 2 and 3, 'TLS(D=X)+Melamed' represents the precision and recall curve of the method whose threshold at Step 2 was X and whose frequency-based method used at Step 3 was that of (Melamed, 2000). Just 'Melamed' represents the result of using the method of (Melamed, 2000) alone. The notation of Hiemstra is the same. From Figures 2 and 3, we can say the following:

(1) The results gained through the frequency-based methods alone are not good. One of the reasons for this should be the issue of full co-occurrence.

(2) The results of the combined method are by far better than those of the frequency-based method only. For instance, the combined method (TLS(D=0.8)+Melamed) achieved 80% precision at 84% recall against the information processing abstract corpus while the method of (Melamed, 2000) achieved 80% precision at just 8% recall. Incorporating the transliteration-based method to extract Japanese-English seems to have been effective, as we hypothesized in our preliminary investigation.

(3) Generally speaking, the recall against the information processing abstract corpus was better than that against the architecture corpus. Table 5 shows that the amount of loan-word pairs in the information processing corpus is larger than that in the architecture corpus. The extraction result indicates that our method will function more effectively in fields or language pairs where word loans are common.

(4) The maximum of the recall when we set $D = 0.8$ was larger than that when we set $D = 0.9$. It means that, to achieve overall good recall, we should rely on transliteration-based method and greedily extract loan-word pairs.

These tendencies were also observed in the results against two title corpora, though the precision and recall were higher.

4.2 Result of Extracting Non-transliterational Pairs

The results of the experiment (B) against the information processing abstract corpus and the architecture abstract corpus are shown in Figures 4 and 5, respectively. We can see in Figure 4 that the performance of the combined method for extracting non-transliterational pairs is better than that of (Melamed, 2000) alone. For instance, the combined method (TLS(D=0.8)+Melamed) achieved 64% precision at 2% recall, while the method of (Melamed, 2000) achieved 19% precision at 2% recall. Similar results were obtained in other corpora and when we used the method of (Hiemstra, 1997). This indicates that Steps 1 and 2 resolved a number of cases of full co-occurrence among non-transliterational $min(W_J, W_E) = 1$ *translation pairs* and enabled the frequency-

[3] We adopted the method 'B' of (Melamed, 2000). We tried his '$score_B$' as the final criteria, too, but the above score P produced better results.

based method to extract them.

We can see in Figures 5 that the precision when we set $D = 0.9$ was lower than those when we set $D = 0.8$ and $D = 0.7$. It means that if we extract and remove only the highly-matched loan-word pairs, the full co-occurrence is not resolved. We observed that the best value for D was around 0.8.

5 Conclusions

From the standpoint that the low-frequency translation pairs in bilingual corpora include many useful pairs, we developed a method for extracting them from corpora. The investigation and experiment clarified the following four points: (1) The frequency-based method is not effective in extracting low-frequency translation pairs because of full co-occurrence; (2) In Japanese-English bilingual corpora, low-frequency translation pairs are often loan-word pairs and they can be extracted using the transliteration-based method; (3) The performance of the combined method for extracting low-frequency translation pairs is higher than that of the frequency-based method alone; (4) Extracting and removing the loan-word pairs using the transliteration-based method leads to the resolution or amelioration of full co-occurrence and enables the frequency-based method to extract non-transliterational low-frequency pairs.

What is special about our research is that it focused on word pairs that have often been ignored and that it proposed a method addressing the circumstances of the Japanese-English language pair quite well. We feel more attention should be paid to language-pair-dependent knowledge and we developed an optimized method for Japanese and English based on this framework. We would like to add that, by modifying the transliteration patterns, similar methods should work fairly well against other language pairs where word loans are common.

References

L. Ahrenberg et al. 1998. "A Simple Hybrid Aligner for Generating Lexical Correspondences in Parallel Texts". In *Proceedings of the COLING-ACL'98*, pages 29–35.

J. Breen. EDICT. http://www.csse.monash.edu.au/~jwb/.

N. Collier et al. 1997. "Acquisition of English-Japanese Proper Nouns from Noisy-Parallel Newswire Articles Using KATAKANA Matching". In *Proceedings of the Natural Language Processing Pacific Rim Symposium 1997*, pages 309–314.

A. Fujii and T. Ishikawa. 2001. "Japanese/English Cross-Language Information Retrieval: Exploration of Query Translation and Transliteration". *Computers and the Humanities*, 35(4):389–420.

W. A. Gale and K. W. Church. 1991. "Identifying Word Correspondences in Parallel Texts". In *Proceedings of the DARPA Speech and Natural Language Workshop*, pages 152–157.

D. Hiemstra. 1997. "Deriving a Bilingual Lexicon for Cross-Language Information Retrieval". In *Proceedings of the Fourth Groningen International Information Technology Conference for Students*, pages 21–26.

K. S. Jeong et al. 1999. "Automatic Identification and Back-transliteration of Foreign Words for Information Retrieval". *Information Processing and Management*, 35(4):523–540.

M. Kitamura and Y. Matsumoto. 1996. "Automatic Extraction of Word Sequence Correspondences in Parallel Corpora". In *Proceedings of the 4th Workshop on Very Large Corpora*, pages 79–87.

K. Knight and J. Graehl. 1997. "Machine Transliteration". In *Proceedings of the 35th Annual Conference of the ACL*, pages 128–135.

J. Kupiec. 1993. "An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora". In *Proceedings of the Sixth Conference of the EACL*, pages 17–22.

I. D. Melamed. 2000. "Models of Translational Equivalence among Words". *Computational Linguistics*, 26(2):221–249.

H. Omae et al. 2003. "Extraction of Compound Word Translations from Non-parallel Japanese-French Text in World Wide Web". In *Proceedings of the Workshop PAPILLON-2003 on Multilingual Lexical Databases.* (No Pagenation).

F. Smadja et al. 1996. "Translating Collocations for Bilingual Lexicons: A Statistical Approach". *Computational Linguistics*, 22(1):1–38.

K. Tsuji. 2002. "Automatic Extraction of Translational Japanese-KATAKANA and English Word Pairs from Bilingual Corpora". *International Journal of Computer Processing of Oriental Languages*, 15(3):261–279.

Corpus	$min(W_J, W_E)$	Pair	R_Dict	R_Full_Cooc	Katak	Roman	Kanji	Other
Inf-Abst	1	548	8.76	84.12	17.15	55.47	14.96	12.41
	2	236	22.03	17.80	23.31	50.00	19.92	6.78
	3	150	24.00	7.33	16.67	47.33	26.00	10.00
	4	151	29.14	2.65	17.22	47.68	28.48	6.62
	5–9	425	40.47	0.94	24.00	30.12	39.06	6.83
	10+	1,907	62.87	0.05	21.31	24.47	48.11	6.12
Arc-Abst	1	314	11.78	90.13	25.48	18.79	46.50	9.24
	2	147	25.85	19.05	28.57	16.33	49.66	5.44
	3	102	37.25	3.92	30.39	12.75	50.00	6.86
	4	51	49.02	0.00	31.37	7.84	56.86	3.92
	5–9	230	48.26	0.43	18.26	6.96	72.17	2.61
	10+	1,357	67.65	0.00	15.81	6.36	73.65	4.18
Inf-Titl	1	248	17.34	55.65	18.95	35.08	35.89	10.08
	2	102	27.45	18.63	19.61	39.22	34.31	6.86
	3	79	43.04	3.80	17.72	22.78	50.63	8.86
	4	63	49.21	0.00	34.92	12.70	47.62	4.76
	5–9	181	45.86	0.00	24.72	5.39	68.03	1.86
	10+	538	61.52	0.00	22.05	17.84	55.08	5.04
Arc-Titl	1	147	21.77	55.78	32.65	12.24	50.34	4.76
	2	81	30.86	17.28	22.22	13.58	60.49	3.70
	3	48	52.08	8.33	33.33	10.42	54.17	2.08
	4	44	65.91	2.27	20.45	4.55	72.73	2.27
	5–9	153	60.78	0.65	16.99	2.61	77.78	2.61
	10+	588	66.50	0.17	15.93	4.52	77.29	2.26

Table 5: Translation Pairs in Dictionary, Fully Co-occurred and Their Character Types

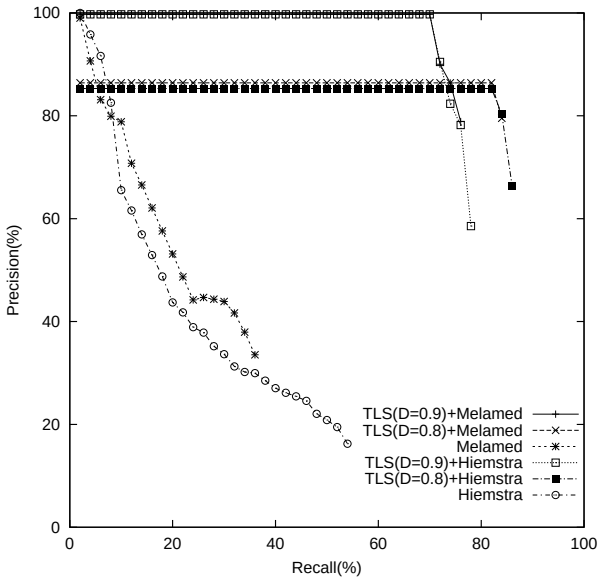

Figure 2: Extracting $min(W_J, W_E) = 1$ *Translation Pairs* (Inf-Abst)

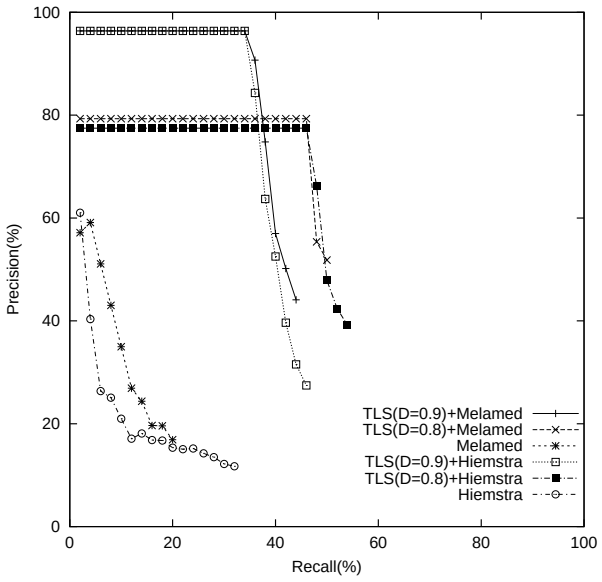

Figure 3: Extracting $min(W_J, W_E) = 1$ *Translation Pairs* (Arc-Abst)

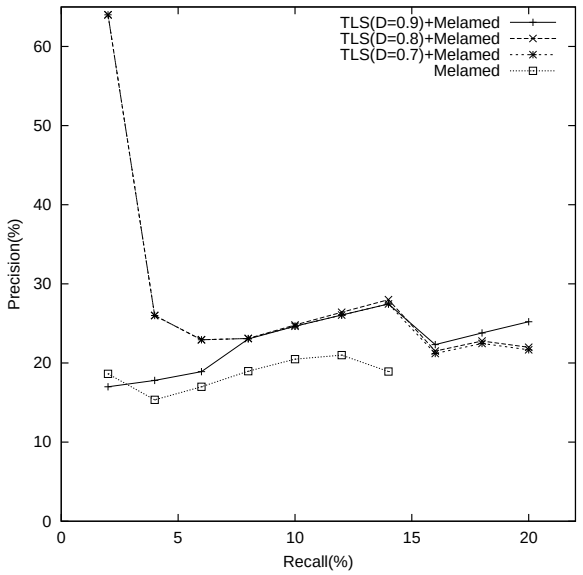

Figure 4: Extracting Non-transliterational $min(W_J, W_E) = 1$ *Translation Pairs* (Inf-Abst)

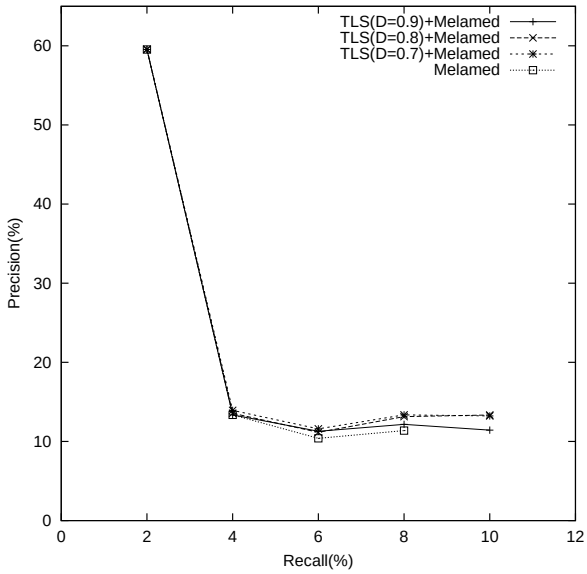

Figure 5: Extracting Non-transliterational $min(W_J, W_E) = 1$ *Translation Pairs* (Arc-Abst)