Concept Unification of Terms in Different Languages for IR

Qing Li, Sung-Hyon Myaeng
Information & Communications
University, Korea
{liqing,myaeng}@icu.ac.kr

Yun Jin
Chungnam National
University, Korea
wkim@cnu.ac.kr

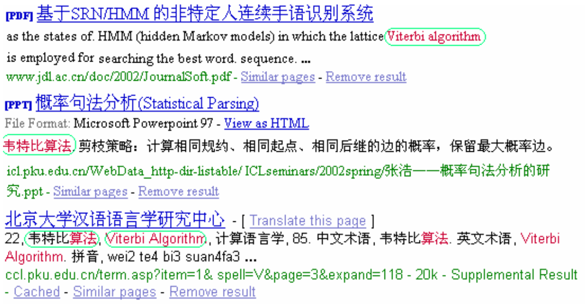
Bo-yeong Kang
Seoul National University,
Korea
comeng99@snu.ac.kr

Abstract

Due to the historical and cultural reasons, English phases, especially the proper nouns and new words, frequently appear in Web pages written primarily in Asian languages such as Chinese and Korean. Although these English terms and their equivalences in the Asian languages refer to the same concept, they are erroneously treated as independent index units in traditional Information Retrieval (IR). This paper describes the degree to which the problem arises in IR and suggests a novel technique to solve it. Our method firstly extracts an English phrase from Asian language Web pages, and then unifies the extracted phrase and its equivalence(s) in the language as one index unit. Experimental results show that the high precision of our conceptual unification approach greatly improves the IR performance.

1 Introduction

The mixed use of English and local languages presents a classical problem of vocabulary mismatch in monolingual information retrieval (MIR). The problem is significant especially in Asian language because words in the local languages are often mixed with English words. Although English terms and their equivalences in a local language refer to the same concept, they are erroneously treated as independent index units in traditional MIR. Such separation of semantically identical words in different languages may limit retrieval performance. For instance, as shown in Figure 1, there are three kinds of Chinese Web pages containing information related with "Viterbi Algorithm (韦特比算法)". The first case contains "Viterbi Algorithm" but not its Chinese equivalence "韦特比算法". The second contains "韦特比算法" but not "Viterbi Algorithm". The third has both of them. A user would expect that a query with either "Viterbi Algorithm" or "韦特比算法" would retrieve all of these three groups of Chinese Web pages. Otherwise some potentially useful information will be ignored.



[PDF] 基于SRN/HMM 的非特定人连续手语识别系统
as the states of. HMM (hidden Markov models) in which the lattice Viterbi algorithm is employed for searching the best word. sequence. ...
www.jdl.ac.cn/doc/2002/JournalSoft.pdf - Similar pages - Remove result

[PPT] 概率句法分析(Statistical Parsing)
File Format: Microsoft Powerpoint 97 - View as HTML
韦特比算法,剪枝策略：计算相同规约、相同起点、相同后继的边的概率，保留最大概率边。
icl.pku.edu.cn/WebData_http-dir-listable/ ICLseminars/2002spring/张浩——概率句法分析的研究.ppt - Similar pages - Remove result

北京大学汉语语言学研究中心 - [Translate this page]
22, 韦特比算法, Viterbi Algorithm, 计算语言学, 85. 中文术语, 韦特比算法. 英文术语, Viterbi Algorithm. 拼音, wei2 te4 bi3 suan4fa3 ...
ccl.pku.edu.cn/term.asp?item=1& spell=V&page=3&expand=118 - 20k - Supplemental Result - Cached - Similar pages - Remove result

Figure 1. Three Kinds of Web Pages

Furthermore, one English term may have several corresponding terms in a different language. For instance, Korean words "디지탈", "디지틀", and "디지털" are found in local Web pages, which all correspond to the English word "digital" but are in different forms because of different phonetic interpretations. Establishing an equivalence class among the three Korean words and the English counterpart is indispensable. By doing so, although the query is "디지탈", the Web pages containing "디지틀", "디지털" or "digital" can be all retrieved. The same goes to Chinese terms. For example, two same semantic Chinese terms "维特比" and "韦特比" correspond to one English term "Viterbi". There should be a semantic equivalence relation between them.

Although tracing the original English term from a term in a native language by back transliteration (Jeong et al., 1999) is a good way to build such mapping, it is only applicable to the words that are amenable for transliteration based on the phoneme. It is difficult to expand the method to abbreviations and compound words.

Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 641–648,
Sydney, July 2006.

Since English abbreviations frequently appear in Korean and Chinese texts, such as "세계무역기구 (WTO)" in Korean, "世界贸易组织 (WTO)" in Chinese, it is essential in IR to have a mapping between these English abbreviations and the corresponding words. The same applies to the compound words like "서울대 (Seoul National University)" in Korean, "疯牛病 (mad cow disease)" in Chinese. Realizing the limitation of the transliteration, we present a way to extract the key English phrases in local Web pages and conceptually unify them with their semantically identical terms in the local language.

2 Concept Unification

The essence of the concept unification of terms in different languages is similar to that of the query translation for cross-language information retrieval (CLIR) which has been widely explored (Cheng et al., 2004; Cao and Li, 2002; Fung et al., 1998; Lee, 2004; Nagata et al., 2001; Rapp, 1999; Zhang et al., 2005; Zhang and Vine, 2004). For concept unification in index, firstly key English phrases should be extracted from local Web pages. After translating them into the local language, the English phrase and their translation(s) are treated as the same index units for IR. Different from previous work on query term translation that aims at finding relevant terms in another language for the target term in source language, conceptual unification requires a high translation precision. Although the fuzzy Chinese translations (e.g. "病毒(virus), 陈盈豪 (designer's name), 电脑病毒 (computer virus)) of English term "CIH" can enhance the CLIR performance by the "query expansion" gain (Cheng et al., 2004), it does not work in the conceptual unification of terms in different languages for IR. While there are lots of additional sources to be utilized for phrase translation (e.g., anchor text, parallel or comparable corpus), we resort to the mixed language Web pages which are the local Web pages with some English words, because they are easily obtainable and frequently self-refresh.

Observing the fact that English words sometimes appear together with their equivalence in a local language in Web texts as shown in Figure 1, it is possible to mine the mixed language search-result pages obtained from Web search engines and extract proper translations for these English words that are treated as queries. Due to the language nature of Chinese and Korean, we integrate the phoneme and semanteme instead of statistical information alone to pick out the right translation from the search-result pages.

3 Key Phrase Extraction

Since our intention is to unify the semantically identical words in different languages and index them together, the primary task is to decide what kinds of key English phrases in local Web pages are necessary to be conceptually unified.

In (Jeong et al., 1999), it extracts the Korean foreign words for concept unification based on statistical information. Some of the English equivalences of these Korean foreign words, however, may not exist in the Korean Web pages. Therefore, it is meaningless to do the cross-language concept unification for these words. The English equivalence would not benefit any retrieval performance since no local Web pages contain it, even if the search system builds a semantic class among both local language and English for these words. In addition, the method for detecting Korean foreign words may bring some noise. The Korean terms detected as foreign words sometimes are not meaningful. Therefore, we do it the other way around by choosing the English phrases from the local Web pages based on a certain selection criteria.

Instead of extracting all the English phrases in the local Web pages, we only select the English phrases that occurred within the special marks including quotation marks and parenthesis. Because English phrases within these markers reveal their significance in information searching to some extent. In addition, if the phrase starts with some stemming words (e.g., for, as) or includes some special sign, it is excluded as the phrases to be translated.

4 Translation of English Phrases

In order to translate the English phrases extracted, we query the search engine with English phrases to retrieve the local Web pages containing them. For each document returned, only the title and the query-biased summary are kept for further analysis. We dig out the translation(s) for the English phrases from these collected documents.

4.1 Extraction of Candidates for Selection

After querying the search engine with the English phrase, we can get the snippets (title and summary) of Web texts in the returned search-result pages as shown in Figure 1. The next step then is to extract translation candidates within a window of a limited size, which includes the

English phrase, in the snippets of Web texts in the returned search-result pages. Because of the agglutinative nature of the Chinese and Korean languages, we should group the words in the local language into proper units as translation candidates, instead of treating each individual word as candidates. There are two typical ways: one is to group the words based on their co-occurrence information in the corpus (Cheng et al., 2004), and the other is to employ all sequential combinations of the words as the candidates (Zhang and Vine, 2004). Although the first reduces the number of candidates, it risks losing the right combination of words as candidates. We adopt the second in our approach, so that, return to the aforementioned example in Figure 1, if there are three Chinese characters (韦特比) within the predefined window, the translation candidates for English phrases “Viterbi” are “韦”,“特”, “比”, “韦特 ”, “特比”, and “韦特比”. The number of candidates in the second method, however, is greatly increased by enlarging the window size k . Realizing that the number of words, n , available in the window size, k , is generally larger than the predefined maximum length of candidate, m , it is unreasonable to use all adjacent sequential combinations of available words within the window size k . Therefore, we tune the method as follows:

1. If $n \le m$, all adjacent sequential combinations of words within the window are treated as candidates

2. If $n > m$, only adjacent sequential combinations of which the word number is less than m are regarded as candidates. For example, if we set n to 4 and m to 2, the window “$w_1 w_2 w_3 w_4$” consists of four words. Therefore, only “$w_1 w_2$”, “$w_2 w_3$”, “$w_3 w_4$”, “w_1”, “w_2”, “w_3”, “w_4” are employed as the candidates for final translation selection.

Based on our experiments, this tuning method achieves the same performance while reducing the candidate size greatly.

4.2 Selection of candidates

The final step is to select the proper candidate(s) as the translation(s) of the key English phrase. We present a method that considers the statistical, phonetic and semantic features of the English candidates for selection.

Statistical information such as co-occurrence, Chi-square, mutual information between the English term and candidates helps distinguish the right translation(s). Using Cheng’s Chi-square method (Cheng et al., 2004), the probability to find the right translation for English specific term is around 30% in the top-1 case and 70% in the top-5 case. Since our goal is to find the corresponding counterpart(s) of the English phrase to treat them as one index unit in IR, the accuracy level is not satisfactory. Since it seems difficult to improve the precision solely through variant statistical methods, we also consider semantic and phonetic information of candidates besides the statistical information. For example, given the English Key phrase “Attack of the clones”, the right Korean translation “클론의습격” is far away from the top-10 selected by Chi-square method (Cheng et al., 2004). However, based on the semantic match of “습격” and “Attack”, and the phonetic match of “클론” and “clones”, we can safely infer they are the right translation. The same rule applies to the Chinese translation “克隆人的进攻”, where “克隆人” is phonetically match for “clones” and “进攻” semantically corresponds to “attack”.

In selection step, we first remove most of the noise candidates based on the statistical method and re-rank the candidates based on the semantic and phonetic similarity.

4.3 Statistical model

There are several statistical models to rank the candidates. Nagata (2001) and Huang (2005) use the frequency of co-occurrence and the textual distance, the number of words between the Key phrase and candidates in texts to rank the candidates, respectively. Although the details of the methods are quite different, both of them share the same assumption that the higher co-occurrence between candidates and the Key phrase, the more possible they are the right translations for each other. In addition, they observed that most of the right translations for the Key phrase are close to it in the text, especially, right after or before the key phrase (e.g. “ …연방수사국(FBI)이…”). Zhang (2004) suggested a statistical model based on the frequency of co-occurrence and the length of the candidates. In the model, since the distance between the key phrase and a candidate is not considered, the right translation located far away from the key phrase also has a chance to be selected. We observe, however, that such case is very rare in our study, and most of right translations are located within 5~8 words. The distance information is a valuable factor to be considered.

In our statistical model, we consider the frequency, length and location of candidates together. The intuition is that if the candidate is the right translation, it tends to co-occur with the key phrase frequently; its location tends to be close to the key phrase; and the longer the candidates' length, the higher the chance to be the right translation. The formula to calculate the ranking score for a candidate is as follows:

$$w_{FL}(q,c_i) = \alpha \times \frac{len(c_i)}{\max_{len}} + (1-\alpha) \times \frac{\sum_k \frac{1}{d_k(q,c_i)}}{\max_{Freq-len}}$$

where $d_k(q,c_i)$ is the word distance between the English phrase q and the candidate c_i in the k-th occurrence of candidate in the search-result pages. If q is adjacent to c_i , the word distance is one. If there is one word between them, it is counted as two and so forth. α is the coefficient constant, and $\max_{Freq-len}$ is the max reciprocal of $d_k(q,c_i)$ among all the candidates. $len(c_i)$ is the number of characters in the candidate c_i .

4.4 Phonetic and semantic model

Phonetic and semantic match: There has been some related work on extracting term translation based on the transliteration model (Kang and Choi, 2002; Kang and Kim, 2000). Different from transliteration that attempts to generate English transliteration given a foreign word in local language, our approach is a kind a match problem since we already have the candidates and aim at selecting the right candidates as the final translation(s) for the English key phrase.

While the transliteration method is partially successful, it suffers form the problem that transliteration rules are not applied consistently. The English key phrase for which we are looking for the translation sometimes contains several words that may appear in a dictionary as an independent unit. Therefore, it can only be partially matched based on the phonetic similarity, and the rest part may be matched by the semantic similarity in such situation. Returning to the above example, "clone" is matched with "클론" by phonetic similarity. "of" and "attack" are matched with "의" and "습격" respectively by semantic similarity. The objective is to find a set of mappings between the English word(s) in the key phrase and the local language word(s) in candidates, which maximize the sum of the semantic and phonetic mapping weights. We call the sum as SSP (Score of semanteme and phoneme). The higher SSP value is, the higher the probability of the candidate to be the right translation.

The solution for a maximization problem can be found using an exhaustive search method. However, the complexity is very high in practice for a large number of pairs to be processed. As shown in Figure 2, the problem can be represented as a bipartite weighted graph matching problem. Let the English key phrase, *E*, be represented as a sequence of tokens $< ew_1,...,ew_m >$, and the candidate in local language, *C*, be represented as a sequence of tokens $< cw_1,...,cw_n >$. Each English and candidate token is represented as a graph vertex. An edge (ew_i,cw_j) is formed with the weight $\omega(ew_i,cw_j)$ calculated as the average of normalized semantic and phonetic values, whose calculation details are explained below. In order to balance the number of vertices on both sides, we add the virtual vertex (vertices) with zero weight on the side with less number of vertices. The SSP is calculated:

$$\text{SSP=argmax} \sum_{\text{i=1}}^{\text{n}} \omega(kw_i, ew_{\pi(i)})$$

where π is a permutation of {1, 2, 3, …, n}. It can be solved by the Kuhn-Munkres algorithm (also known as Hungarian algorithm) with polynomial time complexity (Munkres, 1957).

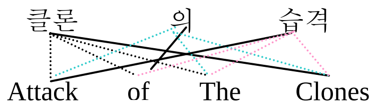


Figure 2. Matching based on the semanteme and phoneme

Phonetic & Semantic Weights: If two languages have a close linguistic relationship such as English and French, cognate matching (Davis, 1997) is typically employed to translate the untranslatable terms. Interestingly, Buckley et al., (2000) points out that "English query words are treated as potentially misspelled French words" and attempts to treat English words as variations of French words according to lexicographical rules. However, when two languages are very distinct, e.g., English–Korean, English–Chinese, transliteration from English words is utilized for cognate matching.

Phonetic weight is the transliteration probability between English and candidates in local language. We adopt the method in (Jeong et al., 1999) with some adjustments. In essence, we compute the probabilities of particular English key phrase *EW* given a candidate in the local language *CW*.

$$\omega_{phoneme}(EW,CW) = \omega_{phoneme}(e_1,...,e_m,c_1,...,c_k)$$
$$= \omega_{phoneme}(g_1,...,g_n,c_1,...,c_k) = \frac{1}{n}\sum_j \log P(g_j \mid g_{j-1}) P(c_j \mid g_j)$$

where the English phrase consists of a string of English alphabets $e_1,...,e_m$, and the candidate in the local language is comprised of a string of phonetic elements. $c_1,...,c_k$. For Korean language, the phonetic element is the Korean alphabets such as “ㄱ”, “ㅣ”, “ㄹ” , “ㅎ” and etc. For Chinese language, the phonetic elements mean the elements of “pinying”. g_i is a pronunciation unit comprised of one or more English alphabets (e.g., ‘ss’ for ‘ㅅ’, a Korean alphabet).

The first term in the product corresponds to the transition probability between two states in HMM and the second term to the output probability for each possible output that could correspond to the state, where the states are all possible distinct English pronunciation units for the given Korean or Chinese word. Because the difference between Korean/Chinese and English phonetic systems makes the above uni-gram model almost impractical in terms of output quality, bi-grams are applied to substitute the single alphabet in the above equation. Therefore, the phonetic weight should be calculated as:

$$\omega_{phoneme}(E,C) = \frac{1}{n}\sum_j \log P(g_j g_{j+1} \mid g_{j-1} g_j) P(c_j c_{j+1} \mid g_j g_{j+1})$$

where $P(c_j c_{j+1} \mid g_j g_{j+1})$ is computed from the training corpus as the ratio between the frequency of $c_j c_{j+1}$ in the candidates, which were originated from $g_j g_{j+1}$ in English words, to the frequency of $g_j g_{j+1}$. If $j=1$ or $j=n$, g_{j-1} or g_{j+1} , c_{j+1} is substituted with a space marker.

The semantic weight is calculated from the bilingual dictionary. The current bilingual dictionary we employed for the local languages are Korean-English WorldNet and LDC Chinese-English dictionary with additional entries inserted manually. The weight relies on the degree of overlaps between an English translation and the candidate

$$w_{semanteme}(E,C) = argmax \frac{No. of \ overlapping \ units}{total \ No. of \ units}$$

For example, given the English phrase “Inha University” and its candidate “인하대 (Inha University), “University” is translated into “대학교”, therefore, the semantic weight between “University” and “대” is about 0.33 because only one third of the full translation is available in the candidate.

Due to the range difference between phonetic and semantic weights, we normalized them by dividing the maximum phonetic and semantic weights in each pair of the English phrase and a candidate if the maximum is larger than zero.

The strategy for us to pick up the final translation(s) is distinct on two different aspects from the others. If the SSP values of all candidates are less than the threshold, the top one obtained by statistical model is selected as the final translation. Otherwise, we re-rank the candidates according to the SSP value. Then we look down through the new rank list and draw a “virtual” line if there is a big jump of SSP value. If there is no big jump of SSP values, the “virtual” line is drawn at the bottom of the new rank list. Instead of the top-1 candidate, the candidates above the “virtual” line are all selected as the final translations. It is because that an English phrase may have more than one correct translation in the local language. Return to the previous example, the English term “Viterbi” corresponds to two Chinese translations “维特比” and “韦特比”. The candidate list based on the statistical information is “编码, 算法, 译码, 维特比,…,韦特比”. We then calculate the SSP value of these candidates and re-rank the candidates whose SSP values are larger than the threshold which we set to 0.3. Since the SSP value of “维特比(0.91)” and “韦特比(0.91)” are both larger than the threshold and there is no big jump, both of them are selected as the final translation.

5 Experimental Evaluation

Although the technique we developed has values in their own right and can be applied for other language engineering fields such as query translation for CLIR, we intend to understand to what extent monolingual information retrieval effectiveness can be increased when relevant terms in different language are treated as one unit while indexing. We first examine the translation precision and then study the impact of our approach for monolingual IR.

We crawls the web pages of a specific domain (university & research) by WIRE crawler provided by center of Web Research, university of Chile (http://www.cwr.cl/projects/WIRE/). Currently, we have downloaded 32 sites with 5,847

Korean Web pages and 74 sites with 13,765 Chinese Web pages. 232 and 746 English terms were extracted from Korean Web pages and Chinese Web pages, respectively. The accuracy of unifying semantically identical words in different languages is dependant on the translation performance. The translation results are shown in table 1. As it can be observed, 77% of English terms from Korean web pages and 83% of English terms from Chinese Web pages can be strictly translated into accurate Korean and Chinese, respectively. However, additional 15% and 14% translations contained at least one Korean and Chinese translations, respectively. The errors were brought in by containing additional related information or incomplete translation. For instance, the English term “blue chip” is translated into “蓝芯(blue chip)”, “蓝筹股 (a kind of stock)”. However, another acceptable translation “绩优股（a kind of stock)” is ignored. An example for incomplete translation is English phrase “ SIGIR 2005” which only can be translate into “国际计算机检索年会 (international conference of computer information retrieval” ignoring the year.

	Korean		**Chinese**	
	No.	%	No.	%
Exactly correct	179	77%	618	83%
At least one is correct but not all	35	15%	103	14%
Wrong translation	18	8%	25	3%
Total	232	100%	746	100%

Table 1. Translation performance

We also compare our approach with two well-known translation systems. We selected 200 English words and translate them into Chinese and Korean by these systems. Table2 and Table 3 show the results in terms of the top 1, 3, 5 inclusion rates for Korean and Chinese translation, respectively. “Exactly and incomplete” translations are all regarded as the right translations. “LiveTrans” and “Google” represent the systems against which we compared the translation ability. Google provides a machine translation function to translate text such as Web pages. Although it works pretty well to translate sentences, it is ineligible for short terms where only a little contextual information is available for translation. LiveTrans (Cheng et al., 2004) provided by the WKD lab in Academia Sinica is the first unknown word translation system based on web-mining. There are two ways in this system to translate words: the fast one with lower precision is based on the “chi-square” method (χ^2) and the smart one with higher precision is based on “context-vector” method (CV) and “chi-square” method (χ^2) together. “ST” and “ST+PS” represent our approaches based on statistic model and statistic model plus phonetic and semantic model, respectively.

		Top -1	**Top-3**	**Top -5**
Google		56%	NA	NA
Live Trans	“Fast” χ^2	37%	43%	53.5%
	“Smart” χ^2 +CV	42%	49%	60%
Our Methods	ST(d_k=1)	28.5 %	41%	47%
	ST	39 %	46.5%	55.5%
	ST+PS	93%	93%	93%

Table 2. Comparison (Chinese case)

		Top -1	**Top-3**	**Top -5**
Google		44%	NA	NA
Live Trans	“Fast” χ^2	28%	37.5%	45%
	“Smart” χ^2 +CV	24.5%	44%	50%
Our Methods	ST(d_k=1)	26.5 %	35.5%	41.5%
	ST	32 %	40%	46.5%
	ST+PS	89%	89.5%	89.5%

Table 3. Comparison (Korean case)

Even though the overall performance of LiveTrans’ combined method (χ^2 +CV) is better than the simple method (χ^2) in both Table 2 and 3, the same doesn’t hold for each individual. For instance, “Jordan” is the English translation of Korean term “요르단”, which ranks 2nd and 5th in (χ^2) and (χ^2 +CV), respectively. The context-vector sometimes misguides the selection.

In our two-step selection approach, the final selection would not be diverted by the false statistic information. In addition, in order to examine the contribution of distance information in the statistical method, we ran our experiments based on statistical method (ST) with two different conditions. In the first case, we set $d_k(q,c_i)$ to 1, that is, the location information of all candidates is ignored. In the second case, $d_k(q,c_i)$ is calculated based on the real textual distance of the candidates. As in both Table 2 and Table 3, the later case shows better performance.

As shown in both Table 2 and Table 3, it can be observed that “ST+PS” shows the best performance, then followed by “LiveTrans (smart)”, “ST”, “LiveTrans(fast)”, and “Google”. The statistical methods seem to be able to give a rough estimate for potential translations without giving high precision. Considering the contextual words surrounding the candidates and the English phrase can further improve the precision but still less than the improvement made by the phonetic and semantic information in our approach. High precision is very important to the practical application of the translation results. The wrong translation sometimes leads to more damage to its later application than without any translation available. For instance, the Chinese translation of “viterbi” is “算法(algorithm)” by LiveTrans (fast). Obviously, treating “Viterbi” and “算法(algorithm)”as one index unit is not acceptable.

We ran monolingual retrieval experiment to examine the impact of our concept unification on IR. The retrieval system is based on the vector space model with our own indexing scheme to which the concept unification part was added. We employed the standard $tf \times idf$ scheme for index term weighting and idf for query term weighting. Our experiment is based on KT-SET test collection (Kim et al., 1994). It contains 934 documents and 30 queries together with relevance judgments for them.

In our index scheme, we extracted the key English phrases in the Korean texts, and translated them. Each English phrases and its equivalence(s) in Korean is treated as one index unit. The baseline against which we compared our approach applied a relatively simple indexing technique. It uses a dictionary that is Korean-English WordNet, to identify index terms. The effectiveness of the baseline scheme is comparable with other indexing methods (Lee and Ahn, 1999). While there is a possibility that an indexing method with a full morphological analysis may perform better than our rather simple method, it would also suffer from the same problem, which can be alleviated by concept unification approach. As shown in Figure 3, we obtained 14.9 % improvement based on mean average 11-pt precision. It should be also noted that this result was obtained even with the errors made by the unification of semantically identical terms in different languages.

6 Conclusion

In this paper, we showed the importance of the unification of semantically identical terms in different languages for Asian monolingual information retrieval, especially Chinese and Korean. Taking the utilization of the high translation accuracy of our previous work, we successfully unified the most semantically identical terms in the corpus. This is along the line of work where researchers attempt to index documents with concepts rather than words. We would extend our work along this road in the future.

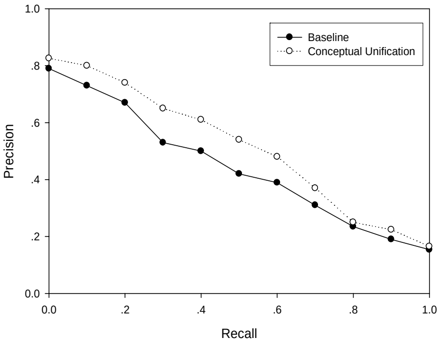


Figure 3. Korean Monolingual IR

Reference

Buckley, C., Mitra, M., Janet, A. and Walz, C.C.. 2000. Using Clustering and Super Concepts within SMART: TREC 6. Information Processing & Management. 36(1): 109-131.

Cao, Y. and Li., H.. 2002. Base Noun Phrase Translation Using Web Data and the EM Algorithm. In Proc. of. the 19th COLING.

Cheng, P., Teng, J., Chen, R., Wang, J., Liu,W., Chen, L.. 2004. Translating Specific Queries with Web Corpora for Cross-language Information Retrieval. In Proc. of ACM SIGIR.

Davis, M.. 1997. New Experiments in Cross-language Text Retrieval at NMSU's Computing Research Lab. In Proc. Of TREC-5.

Fung, P. and Yee., L.Y.. 1998. An IR Approach for Translating New Words from Nonparallel, Comparable Texts. In Proc. of COLING/ACL-98.

Huang, F., Zhang, Y. and Vogel, S.. 2005. Mining Key Phrase Translations from Web Corpora, In Proc. of the Human Language Technologies Conference (HLT-EMNLP).

Jeong, K. S., Myaeng, S. H., Lee, J. S., Choi, K. S.. 1999. Automatic identification and back-transliteration of foreign words for information retrieval. Information Processing & Management. 35(4): 523-540.

Kang, B. J., and Choi, K. S. 2002. Effective Foreign Word Extraction for Korean Information Retrieval. Information Processing & Management, 38(1): 91-109.

Kang, I. H. and Kim, G. C.. 2000. English-to-Korean Transliteration using Multiple Unbounded Overlapping Phoneme Chunks. In Proc. of COLING .

Kim, S.-H. et al.. 1994. Development of the Test Set for Testing Automatic Indexing. In Proc. of the 22nd KISS Spring Conference. (in Korean).

Lee, J, H. and Ahn, J. S.. 1996. Using N-grams for Korean Test Retrieval. In Proc. of SIGIR.

Lee, J. S.. 2004. Automatic Extraction of Translation Phrase Enclosed within Parentheses using Bilingual Alignment Method. In Proc. of the 5th China-Korea Joint Symposium on Oriental Language Processing and Pattern Recognition.

Munkres, J.. 1957. Algorithms for the Assignment and Transportation Problems. J. Soc. Indust. Appl. Math., 5 (1957).

Nagata, M., Saito, T., and Suzuki, K.. 2001. Using the Web as a Bilingual Dictionary. In Proc. of ACL '2001 DD-MT Workshop.

Rapp, R.. 1999. Automatic Identification of Word Translations from Unrelated English and German corpora. In Proc. of ACL.

Zhang, Y., Huang, F. and Vogel, S.. 2005. Mining Translations of OOV Terms from the Web through Cross-lingual Query Expansion, In Proc. of ACM SIGIR-05.

Zhang, Y. and Vines, P.. 2004. Using the Web for Automated Translation Extraction in Cross-Language Information Retrieval. In Proc. of ACM SIGIR-04.