A Statistical Model for Hangeul-Hanja Conversion in Terminology Domain

Jin-Xia HUANG, Sun-Mee BAE, Key-Sun CHOI
Department of Computer Science
Korea Advanced Institute of Science and Technology/KORTERM/BOLA
373-1 Guseong-dong, Yuseong-gu, Daejeon, 305-701
{hgh, sbae, kschoi}@world.kaist.ac.kr

Abstract

Sino-Korean words, which are historically borrowed from Chinese language, could be represented with both Hanja (Chinese characters) and Hangeul (Korean characters) writings. Previous Korean Input Method Editors (IMEs) provide only a simple dictionary-based approach for Hangeul-Hanja conversion. This paper presents a sentence-based statistical model for Hangeul-Hanja conversion, with word tokenization included as a hidden process. As a result, we reach 91.4% of character accuracy and 81.4% of word accuracy in terminology domain, when only very limited Hanja data is available.

1 Introduction

More than one half of the Korean words are Sino-Korean words (Chang, 1996). These words are historically borrowed from Chinese language, could be represented with both Hanja and Hangeul writings. Hanja writing is rarely used in modern Korean language, but still plays important roles in the word sense disambiguation (WSD) and word origin tracing, especially in the terminology, proper noun and compound noun domain.

Automatic Hangeul-Hanja conversion is very difficult for system because of several reasons. There are 473 Hangeul characters (syllables) have Hanja correspondences, map to 4888 common Hanja characters (Kim, 2003). Each of these Hangeul characters could correspond to from one to sixty-four Hanja characters, so it is difficult to system to select the correct Hanja correspondence. Besides that, the sino-Korean Hangeul characters/words could be also native Korean characters/words according to their meaning. For example, "수술(*susul*): stamen, operation, fringe") could correspond to a native Korean word "수술 (stamen)", a sino-Korean word "手術 (operation)", and a mixed word "繡술 (fringe)" (Bae, 2000). It means in Hangeul-Hanja conversion, the same word may be either converted to Hanja or remain as Hangeul writing. In addition, compound sino-Korean words could be written in both with-space and without-space formats even after part of speech (POS) tagging, because the space using is very flexible in Korean language. For example, "한자변환 (*Hanja bienhuan*) (Hanja conversion)" could be in both "한자변환" and "한자 변환" writing formats. It means a compound word tokenization should be included as a pre-processing in Hangeul-Hanja conversion. Automatic Hangeul-Hanja conversion also suffers from another problem, that there are no enough Hanja corpora for statistical approach. In modern Korean language, only few sino-Korean words are written in Hanja writing generally, and the same sino-Korean word with the same meaning could be in either Hangeul or Hanja writing even in the same text.

This paper presents a sentence-based statistical model for Hangeul-Hanja conversion. The model includes a transfer model (TM) and a language model (LM), in which word tokenization is included as a hidden process for compound word tokenization. To find answer for the issues like adapt the model to character or word level, or limit the conversion target to only noun or expand it to other Part of Speech (POS) tags, a series of experiments has been performed. As a result, our system shows significant better result with only very limited Hanja data, when we compare it to the dictionary-based conversion approach used in commercial products.

In the following of this paper: Section 2 discusses related works. Section 3 describes our model. Section 4 discusses several factors considered in the model implementation and experiment design. Section 5 gives the evaluation approaches and a series of experiment results. Section 6 presents our conclusion.

2 Related Works

There are several related areas according to the tasks and approaches. First is previous Korean Hanja, Japanese Kanji (Chinese characters in Japanese language) and Chinese Pinyin input methods, the second one is English-Korean transliteration.

Korean IME (Haansoft, 2002; Microsoft, 2002) supports word-based Hangeul-to-Hanja conversion. It provides all possible Hanja correspondences to all Hanja-related-Hangeul words in user selected range, without any candidate ranking and sino-Korean word recognition. User has to select sino-Korean words and pick out the correct Hanja correspondence. Word tokenization is performed by left-first longest match method; no context nor statistical information is considered in the correspondence providing, except last-used-first approach in one Korean IME (Microsoft, 2002).

A multiple-knowledge-source based Hangeul-Hanja conversion method was also proposed (Lee, 1996). It was a knowledge based approach which used case-frame, noun-noun collocation, co-occurrence pattern between two nouns, last-used-first and frequency information to distinguish the sense of the sino-Korean words and select the correct Hanja correspondence for the given Hangeul writing. Lee (1996) reported that for practical using, there should be enough knowledge base, including case-frame dictionary, collocation base and co-occurrence patterns to be developed.

There are several methods were proposed for Japanese Kana-Kanji conversion, including last-used-first, most-used-first, nearby character, collocation and case frame based approaches. The word co-occurrence pattern (Yamashita, 1988) and case-frame based approach (Abe, 1986) were reported with a quite high precision. The disadvantages include, there should be enough big knowledge-base developed before, and syntactic analyzer was required for the case frame based approach.

Chinese Pinyin conversion is a similar task with Hangeul-Hanja conversion, except that all Pinyin syllables are converted to Chinese characters. To convert Pinyin P to Chinese characters H, Chen and Lee (2000) used Bayes law to maximize $\Pr(H|P)$, in which a LM $\Pr(H)$ and a typing model $\Pr(P|H)$ are included. The typing model reflects online typing error, and also measures if the input is an English or Chinese word. As the report, the statistical based Pinyin conversion method showed better result than the rule and heuristic based Pinyin conversion method.

Hangeul-Hanja conversion normally does not need to convert online input. So we assume the user input is perfect, and employ a transfer model instead of the typing model in Chen and Lee (2000).

The third related work is transliteration. In statistical based English-Korean transliteration, to convert English word E to Korean word K, a model could use Korean LM $\Pr(K)$ and TM $\Pr(E|K)$ (Lee, 1999; Kim et.al, 1999) to maximize $\Pr(K|E)$, or use English LM $\Pr(E)$ and TM $\Pr(K|E)$ to maximize $\Pr(E,K)$ (Jung et, al., 2000).

3 The Model

Different from previous Hangeul-Hanja conversion method in Korean IMEs, our system uses statistical information in both sino-Korean word recognition and the best Hanja correspondence selection. There are two sub-models included in the model, one is Hangeul-Hanja TM, and the other one is Hanja LM. They provide a unified approach to the whole conversion processing, including compound word tokenization, sino-Korean word recognition, and the correct Hanja correspondence selection.

Let S be a Hangeul string (block) not longer than a sentence. For any hypothesized Hanja conversion T, the task is finding the most likely T^*, which is a most likely sequence of Hanja and/or Hangeul characters/words, so as to maximize the highest probability $\Pr(S, T)$: $T^* = \mathrm{argmax}_T \Pr(S, T)$.

$\Pr(S, T)$ could be transfer probability $\Pr(T|S)$ itself. And like the model in Pinyin IME (Chen and Lee, 2000), we also try to use a Hanja LM $\Pr(T)$, to measure the probabilities of hypothesized Hanja and/or Hangeul sequences. The model is also a sentence-based model, which chooses the probable Hanja/Hangeul word according to the context. Now the model has two parts, TM $\Pr(T|S)$, and LM $\Pr(T)$. We have:

$$T^* = \Pr(S,T) = \operatorname*{argmax}_T \Pr(T \mid S)\Pr(T) \qquad (1)$$

T is a word sequence which composed by t_1, t_2, …, t_n, where t_i could be either Hanja or Hangeul word/character. We can see the model in equation (1) does not follow the bayes law. It is only a combination model of TM and LM, in which TM reflects transfer probability, and LM reflects context information. Using linear interpolated bigram as LM, the model in equation (1) can be rewritten as equation 2.

$$\Pr(S,T) \approx \prod_{i=1}^{n} \Pr(t_i \mid s_i)\{\beta \Pr(t_i \mid t_{i-1}) + (1-\beta)\Pr(t_i)) \qquad (2)$$

Word tokenization is also a hidden process in model (2), so both $T=t_1, t_2, \ldots, t_n$ and $T'=t'_1, t'_2, \ldots t'_m$ can be the correspondences of given source sentence S. In practice, a Viterbi algorithm is used to search the best T^* sequence.

We do not use the noisy channel model $\Pr(T|S)=\mathrm{argmax}_T \Pr(S|T)\Pr(T)$ to get T^*, because most of the Hanja characters has only one Hangeul writing, so that most of the $\Pr(S|T)$ tend to be 1. So if we use the noisy channel model in Hangeul-Hanja conversion, the model would be weakened to Hanja LM $\Pr(T)$ in most of the cases.

4 Implementation

There are several factors should be considered in the model implementation. For example, we could adapt the model to character-level or word-level; we could adopt a TM weight as an interpolation coefficient, and find out the suitable weight for best result; we can also consider about utilizing Chinese corpus to try to overcome the sparsness problem of Hanja data. We can also limit the sino-Korean candidates to only noun words, or expand the candidates to noun, verb, modifier and affix and so on, to see what kind of POS-tag-restriction is better for the Hangeul-Hanja conversion.

We adopt previous dictionary-based approach as our base-line system. To get the higher precision in the base-line experiments, we also want to check if the big dictionary or small dictionary would be better for the Hangeul-Hanja conversion.

4.1 Word Level or Character Level

There are two kinds of levels in the model implementation. In word level implementation, the s_i in equation (2) is a Hangeul word. In character level implementation, s_i is a sequence of Hangeul characters.

In word level implementation, there is no word tokenization after POS tagging, so unknown word or compound word is considered as one word without further tokenization. The advantage of word level implementation is, there is no noisy caused by tokenization error. Its disadvantage is that, the system is weak for the unknown and compound word conversion.

To the contrary, in character level implementation, word tokenization are performed as a hidden process of the model. There are several reasons for why word tokenization is required even after POS tagging. First, it is because the morph analysis dictionary is different from the Hangeul-Hanja word dictionary, so the compound word in the morph dictionary still could be unknown word in Hangeul-Hanja dictionary. Second, there are some unknown words even after POS tagging, and this situation is quite serious in terminology or technical domain. Character level implementation will tokenize a given word to all possible character strings, and try to find out the best tokenization way by finding the most likely T^* via equation (2).

Obviously, character level implementation is better than word level implementation for unknown and compound word conversion, but it also raises the risk of bringing too much noise because of the tokenization error. We have to distinguish which one is better through the experiment.

4.2 Transfer Model Weight

Our model in equation 2 is not derived from Bayes law. We just use the conditional probability Pr(*T*|*S*) to reflect the Hangeul-Hanja conversion possibility, and assume Hanja LM Pr(*S*) would be helpful for the output smoothing. The model is only a combination model, so we need a interpolation coefficient α - a TM weight, to get the best combination way of the model. Get the log of the equation, the equation (2) can be rewritten as equation (3).

$$T^* \approx \arg\max_T \sum_{i=1}^{n} \{\alpha \log(\Pr(t_i \mid s_i)) + (1-\alpha)\log\{\beta \Pr(t_i \mid t_{i-1}) + (1-\beta)\Pr(t_i)\}\} \quad (3)$$

where, α = [0,1] is the TM weight.

When α takes a value between 0 to 1, it's a combination model. When α=1, the model is a TM; and when α=0, the model is a LM.

To the LM, we test both unigram and bigram in word level experiment. The interpolated bigram in equation (3) is used for character level implementation.

4.3 Language Resource Utilization

There is no much Hanja data could be used for Hangeul-Hanja conversion. So we treat Hangeul-Hanja word dictionary as a Dictionary corpus, which is 5.3Mbytes in our experiment, to get unigram, bigram and transfer probability. The extracted data from dictionary is called dictionary data D.

Second, we extract user data U from a very small user corpus (0.28Mbytes in our open test), which is in the same domain with the testing data.

Finally, we assume that Chinese corpus is helpful for the Hangeul-Hanja conversion because of the historical relation between them, although they may not exactly the same words in the two language. We convert the code of the Hanja words to Chinese ones (GB in our experiment) to get Chinese data D (unigram and bigram) for the Hanja words from Chinese corpus, which is 270Mbytes corpus in news domain (TREC9, 2000).

We want to know how much these different data D, U, C can help for Hangeul-Hanja conversion, and testify that through experiment.

4.4 POS Tag Constraint

We compare two cases to see the influence of the POS tag constraint in sino-Korean recognition. The first case is only treat Noun as potential sino-Korean, and in the other case we extend noun to other possible POS tags, including noun, verb, modification, suffix, and affix. The sign, foreign, junction words are excluded from the potential sino-Korean candidates. It is because these words would never be sino-Korean in practice. A POS tagger is employed for the pre-processing of our system.

Actually, most of the sino-Korean words that need Hanja writing are noun words, but in practice, the POS tagger normally shows tagging errors. Such kind of tagging error is much more serious in terminology and technical domain. It is one of the reasons why we want to expand the noun words to other possible POS tags. Another reason is, the more restricted the POS tag constraint is, the lower the coverage is, although the higher precision could be expected. So we should have a test to see if the constraint should be more restrict or less.

4.5 Dictionary Size

We develop a dictionary-based conversion system as our base line system. This dictionary-based system follows the approach used in the previous Korean IMEs. The difference is our system uses POS tagger, and gives the best candidate for all sino-Korean words, when previous IMEs only provide all possible candidates without ranking and let user to select the correct one.

Intuitively, the bigger the dictionary is, the better the conversion result would be. But generally, the word in bigger dictionary has more candidates, so it is still possible that bigger dictionary will low down the conversion performance. So we want to distinguish which one is better for Hangeul-Hanja in practical using.

We used two dictionaries in the experiments, one contains 400k Hangeul-Hanja word entries, and one contains 60k Hangeul-Hanja word entries.

5 Experiment

This chapter shows the experiments on the model in equation 3 and some different implementations we have discussed above.

There are two parts in the experiments, first one is mostly related to word level model implementation, in which the basic issues like language resource utilization and POS tag restriction, and some word level related issues like bigram or unigram for LM in word level are tested. The second part is mostly character level related.

Several evaluation standards are employed in the experiments. The adopted standards and evaluation approaches are reported in the first section of the experiments.

5.1 Evaluation Standard and Approach

We use several evaluation standards in the experiments. To reflect the readability from the user viewpoint, we adopt word and phrase (sentence) level accuracy, precision and recall; to compare the automatic conversion result with the standard result – from the developer viewpoint, Dice-coefficient based similarity calculation is employed also; to compare with previous Chinese Pinyin input method, a character based accuracy evaluation is also adopted.

An automatic evaluation and analysis system is developed to support large scale experiments. The system compares the automatic result to the standard one, and performs detailed error analysis using a decision tree.

5.2 Word Level Experiment

In this part, the basic issues like language resource utilization and POS tag restriction, and the word level related issues, like bigram or unigram for LM are performed.

The objects of **the first experiment** are, firstly, compare a simple LM based statistical approach with the base line - dictionary based approach; secondly, see if large dictionary is better than small dictionary in dictionary based conversion; thirdly, see if Chinese corpus does help to the Hangeul-Hanja conversion.

A small dictionary based conversion (Dic), large dictionary based conversion (BigDic), a unigram (Unigram) and a bigram based (Bigram) word level conversion, are performed to compared to the each other.

The small dictionary Dic has 56,000 Hangeul-Hanja entries; while the large dictionary BigDic contains 280,000 Hangeul-Hanja entries. The unigram and bigram are extracted from Chinese data C. The test set is a small test set with 90 terms (180 content words) from terminology domain. Word level precision and recall with F1-measure are employed as evaluation standard.

	Dic	BigDic	unigram	bigram
P	57.1%	50.0%	78.6%	78.6%
R	25.7%	44.0%	70.6%	70.6%
F1	35.4%	46.8%	74.4%	74.4%

Table 1. Base line (small dic vs. large dic) vs. Statistical approach (unigram vs. bigram)

From the result shows in table 1, we can get the conclusions that, 1) compare to the small dictionary, large dictionary reaches better F1-measure because of the enhancement in recall, although the precision is slightly low downed because of more Hanja candidates for given Hangeul entry; 2) Statistical approach shows obvious better result than the dictionary based approach, although it is only a very simple LM; 3) Chinese data does help to the Hangeul-Hanja conversion. We have to evaluation its impact by comparing it with other Hanja data in further experiments. 4) Bigram shows similar result with unigram in word level conversion, it shows that data sparseness problem is still very serious.

The objects of **the second experiment** include the evaluation on different POS tag constraints and the comparison between different language resources.

First is evaluation on different POS tag constraints. Let the system employs unigram based Hangeul-Hanja conversion approach, which uses dictionary data D (word unigram from large dictionary at here). Our experiment wants to compare the case of only considering noun as potential sino-Korean words ("Dn" in table 2), with the case of extending the POS tags to verb, modification and affix ("De" in table 2). Second evaluation is comparison on different language resources. As we have mentioned above, D is data from large dictionary (word unigram is used at here), U is data from very small user corpus, and C is data from Chinese corpus. We want to compare the different combination of these language resources. In the second evaluation, extended POS tag constraint is employed.

The experiment uses a test set with 5,127 terms (12,786 content words, 4.67 Hanja candidates per sino-Korean word in average) in computer science and electronic engineering domain. User data U is from user corpus, which is the same with the test set at here (so it is a closed test). In evaluation, a dice-coefficient based similarity evaluation standard is employed.

	Dn	De	U	C	DC	DU	UC	DUC
Sim	0.71	0.75	0.81	0.72	0.75	0.82	0.82	0.81

Table 2. POS tag constraint and language resource evaluation

From the table 2, we can see that, 1) the extended POS tag constraint ("De" in table 2) shows better result than the noun POS tag constraint ("Dn"); 2) User data U shows better result than dictionary data D ("U" ⇔ "De", "UC" ⇔ "DC" in table 2), and dictionary data D shows better result than Chinese data C ("De" ⇔ "C"), although Chinese corpus (where C is from) is 270MB, and much larger than the Hangeul-Hanja dictionary (5.3MB here, where D is from). It shows that the effect of Chinese data is quite limited in despite of its usefulness.

The object of the third experiment is to find out which TM weight α is better for the word model.

	α=0	α=0.5	α=1
P	78.6%	76.52%	84.80%
R	70.6%	70.70%	77.31%
F1	74.4%	73.4%	80.8%

Table 3. TM weight in word model

Let α to be 0, 0.5, 1, and so the model in equation (3) is LM, combined model, and TM, with the same environment of the second experiment, we get the result in table 3. Word level precision and recall with F1-measure is evaluated. We can see the TM with α=1 shows the best result.

5.3 Character Level Experiment

In the character level experiments, first, we compare the character level model with base line dictionary based approach; Second, compare the character level model with the word level model; Third, to find out the best TM weight for the character level model.

This part of experiments uses a new test set, which has 1,000 terms in it (2,727 content words; 3.9 Hanja candidates per sino-Korean word in average). The user data U has 12,000 Hangeul-Hanja term pairs in it. U is from the same domain of the test set (computer science and electronic engineering domain at here), but there is no overlap with the test set (so it is a opened test).

Several different evaluation standards are employed. As the first column of table 4, "CA", "WA" and "SA" mean character, word, sentence (terms) accuracy, respectively. "Sim" is the similarity based evaluation, and F1 is the value of word level F1-measure which is from word precision/recall evaluation.

%	Dic	wD1	wDUC1	D.5	DU0	DU.2	DU.5	DU.8	DU1
CA	62.9	69.1	75.0	73.1	81.0	89.3	90.2	91.0	91.4
WA	49.9	73.8	75.3	64.6	72.4	77.1	82.3	82.1	81.4
SA	18.8	43.4	51.2	34.7	48.2	67.0	67.5	67.1	68.1
Sim	68.4	75.5	79.7	77.9	82.5	90.4	91.2	91.7	92.1
F1	39.0	65.6	69.7	51.2	60.8	75.7	75.9	75.9	76.2

Table 4: Character level model vs. word level model vs. base line (dictionary based approach)

The first row of table 4 shows the Hangeul-Hanja conversion approach with the employed data and TM weight α. "Dic" is the base line dictionary based approach; "w" means word level model; "D" means dictionary data (extracted from the large dictionary with 400,000 Hangeul-Hangeul and Hangeul-Hanja entries), U means user data described above, C means Chinese data. The digital value like ".5" is TM weight. So, as an example, "wDUC1" means word model with α=1 and using all data resources D, U and C; "DU.2" means character model with α=0.2 and using data D and U.

From the table 4, we can get the conclusions that, 1) all statistical model based approaches shows obviously better performance than the base line dictionary based approach "Dic" (Dic ⇔ others). 2) In most cases, character models show better results than word model (DUx ⇔ wDUCw1). But when there is no user data, word mode is better than character model (wD1⇔D.5). 3) Among character models, the TM with α=1 shows the best result ("DU1" ⇔ "DU.x"). 4) User data has positive impact on the performance ("Dw1 ⇔ DUCw1", "D.5 ⇔ DU.5"), and it is especially important to the character model ("D.5 ⇔ DU.5"). It is because character model may cause more noise because of word tokenization error when there is no user data.

From the table 4, we can see the best result is gotten from character based TM with using dictionary and user data D, U ("DU1"). The best character accuracy is 91.4%, when the word accuracy is 81.4%. The character accuracy is lower than the typing and language model based Chinese Pinyin IME, which was 95% in Chen & Lee (2000). But consider that in our experiment, there is almost no Hanja data except dictionary, and also consider the extra difficulty from terminology domain, this comparision result is quite understandable. Our experiment also shows that, compare to using only LM like it in Chen & Lee (2000), TM shows significantly better result in character accuracy (from 81.0% to 91.4% in our experiment: "DU0" ⇔ "DU1", table 4). Our user evaluation also shows that, to the terminology domain, the automatic conversion result from the system shows even better quality than the draft result from untrained human translator.

5.4 Different Evaluation Standards

Figure 1 shows the trends of different evaluation standards in the same experiment shown in table 4. We can see character accuracy "CA" shows similar trend with similarity based standard "Sim", while word accuracy "WA" and sentence (terms) accuracy "SA" show similar trends with F1-measure "F1", in which "F1"is based on word precision and recall.

From the user viewpoint, word/sentence accuracy and F1-measure reflects readability better than character accuracy. It is because, if there is a character wrongly converted in a word, it affects the readability of whole word but not only that character's. However, character accuracy is more important to the system evaluation, especially for the character level model implementation. It is because the character accuracy can reflect the system performance in full detail than the word or sentence (term) based one.

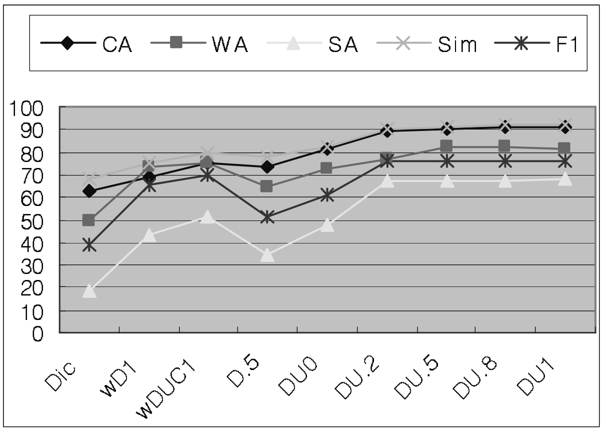

Figure1.The trends of different evaluation standards

6 Conclusion

This paper proposes a sentence based statistical model for Hangeul-Hanja conversion in Korean language. This model provides a unified approach to the whole conversion processing, which includes word tokenization, sino-Korean word recognition and the correct Hanja correspondence selection. A series of experiments have been done for the issues in model and system implementation. Including, adapting the model to character-level or word-level, the influence of the TM weight, the different POS tag constraints on the sino-Korean word recognition, etc.

The experiments show that best result is achieved from character based TM with using both dictionary and user data. The best character accuracy in computer science and electronic engineering terminology domain is 91.4%, which is even better than the draft result from untrained human translator.

This paper also uses several different evaluation standards to see which method is the most suitable one. As a result, we found that the word/term accuracy and word based precision/recall can reflect the user readability well, when the character accuracy is more suitable to the system performance evaluation in full detail.

We are doing further research on general domain, especially about utilizing the concept hierarchy of thesaurus to solve data sparseness problem. We are also considering about use Japanese corpus for Hangeul-Hanja, because the Kanji in Japanese language also has some overlap with the Hanja in Korean language.

7 Acknowledgements

This work was supported in part by Ministry of Science & Technology of Korean government and Korea Science & Engineering Foundation.

References

Abe, M. & Y. Oshima, 1986. *A Kana-Kanji Translation System for Non-segmented Input Sentences Based on Syntactic and Semantic Analysis*, in the Proceedings of COLING-86, 280-285, 1986.

Chen, Zheng and Kai-Fu Lee. 2002. *A New Statistical Approach To Chinese Pinyin Input*. The 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000.)

Dice, L.R. 1945. *Measures of the amount of ecologic association between species*. Journal of Ecology, 26:297-302

Gao, Jianfeng, Hai-Feng Wang, Mingjing Li and Kai-Fu Lee. 2000. *A Unified Approach to Statical Language Modeling for Chinese.* ICASSP-2000, Istanbul, Turkey, June 5 - 9, 2000

Haansoft. 2002. Hangeul 2002. Haansoft Inc.

Kim, Kyongsok. 2003. *Hangeul correspondence for Hanja character in KSX1001.* http://asadal.cs.pusan.ac.kr/hangeul/code/ksx1001-name-hj-v02.txt

Kim J.J., J.S. Lee, and K-S. Choi., 1999. *Pronunciation unit based automatic English-Korean transliteration model using neural network*, In Proceedings of Korea Cognitive Science Association (in Korean).

Lee, Jaeseong. 1999. *An English-Korean Transliterationand retransliteration model for cross-lingual information retrieval.* Ph.D Dissertation. KAIST

Lee, Jong-Hyeok, 1996. *A Sense-analysis-based Hangeul-Hanja conversion System. In Computational Semantics and Application, Meanum Company*, 1996. pp247-278. (in Korean)

Microsoft, 2002. *Korean Input Method 2002*. Microsoft Corporation.

Chang, Suk-Jin. 1996. *Korean*. London Oriental and African Language Library 4. Philadelphia, PA.: John Benjamins. pp.2

Jung , SungYoung, SungLim Hong & EunOk Paek. 2000. *An English to Korean Transliteration Model of Extended Markov Window.* 18th International Conference on Computational Linguistics

TREC9. 2001. http://trec.nist.gov/

Yamashita, M. & F. Obashi, 1988. *Collocational Analysis in Japanese Text Input*, in the Proceedings of COLING-88, 770-772, 1988.