The Construction of a Dictionary for a Two-layer Chinese Morphological Analyzer

Chooi-Ling Goh[1], Jia Lü[1], Yuchang Cheng[1], Masayuki Asahara[1], and Yuji Matsumoto[1]

[1] Graduate School of Information Science, Nara Institute of Science and Technology,
8916-5 Takayama, Ikoma, Nara 630-0192, Japan
{ling-g,jia-l,yuchan-c,masayu-a,matsu}@is.naist.jp

Abstract. We built a morphological analyzer, which can be freely used by anyone for research purpose. In order to build a pratical system, a dictionary with reasonable size is necessary. The initial dictionary is built from the Penn Chinese Treebank corpus v4.0 and contains only 33,438 entries. Since the initial dictionary is quite small, unknown word detection methods are applied to a huge raw text in order to extract new words to be added into the system dictionary. We have successfully constructed a dictionary with 120,769 entries. Finally, we propose a two-layer morphological analyzer to cater for two sets of outputs. The first layer produces the minimal segmentation units defined by us, and the second layer transforms the output of the first layer to the original segmentation units defined by Penn Chinese Treebank.

Keywords: Morphological analysis, unknown word, dictionary, corpus, machine-learning

1 Introduction

To date, a freely usable Chinese morphological analysis system is still not widely available. Furthermore, since there is no single segmentation standard for all tagged corpora provided by different institutions, some systems are available which are developed by the corpora providers, according to their segmentation standard. As far as we know, there is still no system (freely) available for Penn Chinese Treebank[1] (hereafter CTB) standard. Since this treebank is widely used by a lot of researches that do parsing, probably it is a good idea to build a practical morphological analyzer for CTB standard. The initial system that we built contains only 33,438 entries in the system dictionary. To build a practical system, this number is too small. Therefore, we tried to enlarge the system dictionary using unknown word extraction methods. We intend to extract a large amount of unknown words from a huge raw text corpus. Based on our methods, we have successfully increased the system dictionary to 120,769 entries. Although this number is good enough for a practical system but we still hope to add more in the future.

2 New Segmentation Unit

In Chinese language processing community, no single segmentation standard is agreeable across different instituitions. In SIGHAN bakeoff [1], we could see that different institutions have provided different segmentation standards. Most of the disagreements in the standards come from the segmentation of morphologically derived words [2] and named entities. For example, some would say that “孩子们/NN” (children) as one word and some would prefer to it as two words, “孩子/NN” and “们/M”. For named entities such as Chinese person names, whether a string of a surname and a given name should be one word or two words, is also under argument. It would be nice if we can build a system that suits everyone’s needs but it sounds almost impossible. Wu [2] tried to define tree structures to

[1] http://www.cis.upenn.edu/~chinese/

morphologically derived words but that will need a lot of human efforts as they are all based on rules defined. Gao et al. [3] have tried to modify their current system to adapt for all segmentation standards in SIGHAN bakeoff using the transformation-based learning methods [4]. Since we would like to build a system for CTB, we try to define our segmentation units as close as to the CTB standard, or at least to be able to modify back to the CTB standard easily.

There are a few changes that we have made on the CTB corpus to suit our purpose and to ease our processing. We refer to this new segmentation units as minimal segmentation units. The changes are made on proper names, foreign words and numeral type words only. Figure 1 shows the desired output of minimal unit segmentation and the transformation to the CTB standard. With our new segmentation units, we have increased the number of POS tags from 33 to 42 tags.

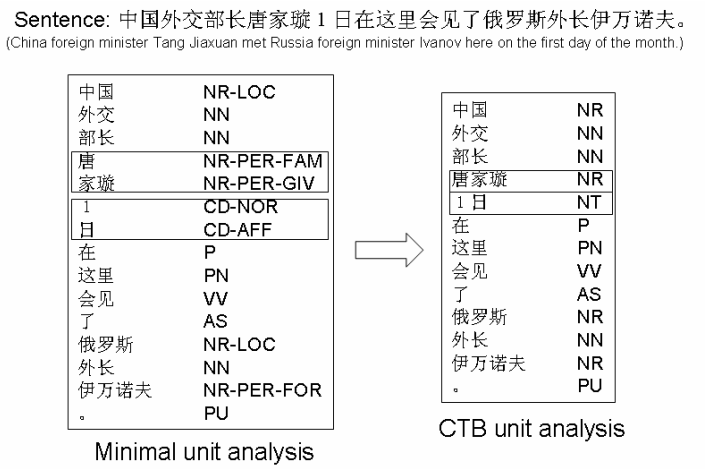

Fig. 1. Transformation from minimal unit segmentation to CTB unit segmentation

There are two advantages of defining the minimal segmentation units. First, we can reduce the size of the dictionary by eliminating those productive numeral type words and foreign words. Second, splitting family names and given names for CJK names in the dictionary can make the combinations of two of them more freely in the text. Moreover, this design can be applied to the analysis of compound words and morphological derived words in the future. These two groups of words are the main types of unknown words in the text. If we define the minimal units for compound words such as "策划/NN室/NN" (planning room), then we can combine them into "策划室/NN" in the second layer. In this case, we can create a compound word dictionary which can show the internal structure of the compound words more precisely. Currently, our implementation does not include the analysis of compound words and morphological derived words. We now describe each modification in the following sections.

2.1 Proper Names

In CTB, all proper names are grouped under one single POS tag as NR. In our new segmentation units, we divide the proper names into 7 new groups. This is because we think that the new groups are more informative and are useful for named entity extraction in the future. First, the person names are identified and 4 groups are introduced, family names (NR-PER-FAM), given names (NR-PER-GIV), foreign names (NR-PER-FOR), and other names (NR-PER-OTH). The family names and given names apply only to Chinese, Japanese and Korean (CJK) names. Originally, CTB does not split family names and given names but we split them into two units. Then, we also define place names (NR-LOC), organization names (NR-ORG) and other proper names (NR-OTH). Based on these definitions, we have manually made the changes to the corpus.

2.2 Foreign Words

Foreign words refer to those that consist of alphabets. As the combination of alphabets is arbitrarily, we do not want to register these words in the dictionary. In our dictionary, only a－z and Ａ－Ｚ are registered as POS tag FW. As a result, we must cut the foreign words into smaller units, meaning one-character units, in our training corpus. For example, the original word "Ｐ／Ｅ值/NN" (P/E value) now becomes "Ｐ/FW ／/PU Ｅ/FW 值/NN". The weakness of his changes is that we will lose the information of the original POS tag, in this case NN. However, the merit is that we do not need to bother about the foreign word segmentation and POS tagging. Although currently we do not register any foreign words in our dictionary, it is possible to add these words in the dictionary if we think that they are frequently used words and it is necessary to register them.

2.3 Numbers

The last changes are on numeral type words. This group of words is very productive and it is impossible to register all possible combinations in the dictionary. Therefore, our dictionary contains only characters ０－９and零–九(zero-nine), 十(ten), 百(hundred), 千(thousand), 万(ten thousand) etc. We also cut the numeral words into one character unit. However, it is simple if we want to combine them in the later stage. There are three types of numeral type words in the corpus, time nouns (NT:三月, March), cardinal numbers (CD:十多, more than ten) and ordinal numbers (OD:第九, number nine). In these examples, there are some characters in the words which are not numbers. Therefore, we introduce 6 new POS tags to tag these numeral words. CD-NOR is used to tag all numbers (０－９and零–九), OD-NOR and NT-NOR are used to tag those words that do not consist of numbers, such as "首" (first) and "半夜" (midnight). Then, we also introduce CD-AFF, OD-AFF and NT-AFF to cater for those characters (affixes) which are not numbers but exist in the numeral type words. Finally the new segmentation for the earlier examples become "三/CD-NOR 月/NT-AFF", "十/CD-NOR 多/CD-AFF" and "第/OD-AFF 九/CD-NOR". We have made the changes in the corpus based on these rules. We also extracted the affixes of these words and added them to the dictionary.

3 Preparation of System Dictionary

In order to build a practical system, we need a dictionary with reasonable size. We can retrieve the words from the training corpus but yet the size is too small when talking about real world text. In this paper, we describe some methods that we have used to enlarge the size of our dictionary. For evaluation purpose, we have used CTB version 4.0 (about 437,000 words) as our training corpus. The exclusive parts from version 5.0 (about 110,000 words) are used as testing data. Our initial dictionary is build from training data only. There exists about 7.8% unknown words (8,603 tokens and 5,343 types) in the test data (9.2% if we include unknown POS, meaning those words exists in the dictionary but are with different POS tags.). Out of the 5,343 types, only 746 are proper nouns. Most of the are common nouns (eg. "村妇" (village woman), "带领人" (leader) and "感想" (impression)) and verbs (eg. "都市化" (urbanization), "返家" (go home), "唤" (to call)). These unknown words are not only compound words but are also word morphemes. In other words, the initial dictionary created from the corpus is different from a normal dictionary prepared for language learners that even some common words are not in there. Therefore, we need to extract more words to be added to the dictionary.

3.1 Extraction from CTB

Based on the segmentation units that have been described above, we made the changes to the corpus accordingly. After that we extracted all words from CTB 4.0 to build our initial dictionary. We leave the exclusive part in CTB 5.0 as testing data (for evaluation). We have also removed some noise which we

found not suitable to be used as the entries in the dictionary[2]. Finally, we built an initial dictionary that contains 33,438 entries (word/POS pairs, a word can have more than one POS tag). There are 28,390 words if we consider the word tokens only. To build a practical system, this number is too small. Therefore, we must find some ways to increase the number of entries in the dictionary.

3.2 Collection of Proper Nouns from Web

We collected various proper nouns from the web. These include place names (5,365 place names in China), country and capital names (391), and Chinese family names (436). These names are quite common on the web and they can be used directly in our system.

3.3 Unknown Word Extraction from Chinese Gigaword

Chinese Gigaword (CGW) is a raw text corpus provided by LDC. The size is about 1,118,380K Chinese characters. We use this corpus to extract new words to add into our system dictionary.

General Unknown Word Extraction and POS Tag Guessing. The unknown word extraction method used is similar to [5]. In this approach, we assign each character with a character type such as NUMber, ALPphabet, SYMbol or HANzi, and label each character with BIES tagset[3]. We use Maximum Entropy models (hereafter ME) for the character-based tagging. We found that this method gives us the best unknown word recall although the precision is a bit lower. In [5], some pruning steps have been applied to delete some false unknown words. However, since this step deteriorates the recall, we do not do the pruning here as our purpose is to collect as many unknown words as possible.

Our evaluation is carried out using the test data in CTB 5.0. With the initial dictionary, there are about 9.2% of unknown word/POS pairs in the test data. Out of this, 7.8% are unknown words, 1.3% are unknown POS (the words exist in the dictionary but are with different POS tags). Currently, our method solves only the problem of unknown word but not the unknown POS. We leave the unknown POS problem as the future work.

The features that we use for tagging are: 2 characters each from left and right contexts, character types, and 2 previously tagged labels. We get 72.2% recall for tokens, 72.1% for types, and 50.6% precision for types. In other words, we get quite high recalls but the precision is not so good. Only about half of the words extracted are correct. However, since we want to increase the size of the dictionary, higher recall means that we get more words.

After unknown word extraction, we need to assign POS tags to them. We use the same method as described in [6]. We also apply ME model as the classification model. The training data are those words that appear only once in the corpus. This covers all major unknown POS tags. The features used are the context words, POS tags (unigram and bigram) and the internal component features (first character, last character and word length). The context features (known words and POS tags) are taken from the morphological analysis during the first layer analysis. Our experimental results show that the accuracy of tagging unknown words is 68.2% if only context features are used and 75.2% if all features are used.

Using this method, we tried on a small part of the CGW corpus for testing purpose. We estimated that 74.9% of the words extracted from CGW corpus are usable in our dictionary. In a real run of the method applied to CGW, we extracted 51,412 word/POS pairs from file xin200209. Then we hired 4 native Chinese to check on the words manually in one month time. 14,537 words are correct, 10,643 words have been corrected with their POS tags. Since it is done manually, we also ask the checkers to correct some of the word boundaries to obtain correct words (7,785 words). Finally, manual checking on the

[2] For example, the phrase （/PU 四/NR-LOC ）/PU 川/NR-LOC （/PU 西/NR-LOC ）/PU 藏/NR-LOC 公路/NN (the road between Sichuan and Tibet) has four location names which we think are not real abbreviations to be used normally.

[3] B - begin, I - inside, E - end, S - single

words give us a total of 26,281 (51.12%) correct words[4]. Although the result is a bit lower than our estimation, we still manage to get quite a number of new words.

Person Name Extraction. In [7, 8], we have seen that one can get better results if the extraction is focused on a certain type of unknown words, such as personal names. This is because we can train the system to be more precise to the type by providing specific features to it. For example, a Chinese person name normally comprises of a family name and a given name. A family name is normally one character long (very few with two characters) and it is almost a closed set. If we can provide the information about the family names, then it will be easier to guess the given names. At our disposal, we also have a set of characters that are possible to be used in transliteration foreign names. These provide some extra features for extraction.

The method that we use here is similar to the one described in [7]. First, an HMM-based analyzer is used to segment and POS tag the text, then an SVM-based chunker is used to extract the person names. Since our target is the Chinese given names and foreign names, we create a dictionary which consists of none of the both. It will make the HMM analyzer to wrongly segment all the names. In the second step, names are extracted by chunking process using SVM. We also provide family names and transliteration characters as the features. We assign each character with one of these 4 tags, FAM (family name), FOR (transliteration character), BTH (can be used for both), OTH (not in use for both). Currently we have collected 482 family names and 581 transliteration characters to be used for the training features. The context window is three characters at both left and right sides.

We have conducted an experiment using the CTB 5.0 test data. In CTB 4.0 there exist 4,190 given name and 926 foreign name instances. We use these data for training. In the test data, there are 1,157 given names and 194 foreign names. Table 1 shows the results of our method. Although we could get quite good accuracy with CJK given names, we could not get a good result with foreign names. This may be because the training data for the foreign names is not enough.

Using this method, we extracted 4,622 person names from CGW, file xin199101[5]. After manual checking, we obtained 3,976 (86%) words which are usable to our system. Since it is done manually, we also asked the checkers to correct some of the wrong POS and reassign boundaries if necessary. The accuracy for given names and foreign names only is about 66%, which follows our estimation during the testing experiments.

Table 1. Results for person name extraction

	Recall	Precision	F-measure
CJK given name	89.0	70.1	78.5
Foreign name	39.7	56.6	46.7
Average	81.9	69.0	74.9

Checking with other Resources. From our past exprience, we realize that manual checking on unknown words in a time consuming task. Therefore, we also look for other solutions to speed up the process. One way is to use other resources for double checking as described below.

Sinica corpus[6] is the first tagged balanced corpus which contains about 5 millions words. Texts are collected from different areas and classified according to five criteria: genre, style, mode, topic, and source. Therefore, this corpus is a representative sample of modern Chinese language. Moreover, the size is 10 times larger than CTB.

Sinica corpus uses a different POS tagset as CTB corpus. It has 46 simplified POS tags, as compared to 33 tags in CTB. Basically the segmentation standard between CTB and Sinica is very similar but there are also some differences. From Sinica corpus, we could get around 150,000 distincts words. Leaving out the copyright problem to use the resources from Sinica, we cannot use the list of words directly from Sinica in our system since the segmentation standard is different. Therefore, we choose to use it in another way. First, we extract the new words from CGW using our unknown word extraction

[4] There are some overlappings among these groups, so the final total is not the same as the total of all groups.
[5] We use a different part of CGW so that the results will not overlap with the previous unknown word extraction.
[6] http://www.sinica.edu.tw/SinicaCorpus

model. Instead of manual checking, we double check the words with Sinica corpus entries. If the words are found, then we assume that these words are correct ones. Since using a corpus requires copyright clearance, we have obtained the permission verbally from the Academia Sinica to use their corpus as a reference.

In order to do this, first we need to compare the POS tagsets to find out equivalent POS tags. Table 2 shows the equivalent POS tags that we use for comparison. We omitted some POS tags that cannot be matched directly, such as proper names, numbers, time nouns etc. As a results, we obtained a list of 105,030 word/POS pairs for comparison. We applied the unknown word extraction model in Section 3.3 to the whole CGW corpus. We managed to extract 33,286 new entries which we are sure to be correct ones since they also exist in Sinica corpus.

Table 2. Matching between Sinica and CTB POS tags

POS Tag	Sinica Tag	CTB Tag
Adjective	A	JJ
Adverb	D, Da, Dfa, Dfb, Dk	AD
Common Noun	Na	NN
Localizer	Ncd	LC
Measure Noun	Nf	M
Verb	V?[?], (+nom)	VV, NN
Stative Verb	VH[?], (+nom)	VA, NN

We also managed to download a list of Chinese names from the web[7]. They provide a list of family names and a list of given names together with their frequencies. From a total of 217,913 unique names, they give 619 distinct family names and 75,581 distinct given names. We found out that there are quite a lot of noise in the files because the way they cut the unique names into family names and given names are not so reliable. Therefore, we decided not to use the family name list since we already have quite a number of them. However, we also do not want to use the given name list directly because it might contain error names as well. Our approach is the same as using Sinica corpus as a reference. First we extract the given names from the CGW using the method as described in Section 3.3, then we double check with the provided given name list to see if the names are inside the list. If they are in the list, then we assume that they are correct given names. By this way, we managed to extract 18,818 given names from CGW automatically.

Composition of Current Dictionary. Overall, 27% (33,438, includes 4,824 proper names) of the current dictionary is extracted from CTB4.0, 5% (6,192) is collected from the web, 25% (30,257) is extracted from manual checking and 43% (52,104) from auto checking. There are some overlapped entries between these groups. In total, we have collected 120,769 (includes 40,336 proper names) entries in our dictionary.

4 Two-layer Morphological Analysis

We propose a two-layer morphological analysis in our system. The first layer produces the segmentation and POS tags based on our definition, meaning the minimal segmentation units. The second layer transforms the output of the first layer to CTB original segmentation units. Figure 2 shows the overview of the system. The right hand side shows the process used for preparation of the training data and the system dictionary. The left hand side shows the process of two-layer analysis. The preparation of the training data and the system dictionary has already been described in the previous sections. We will describe the methods used in each layer of analysis in the following sections. Using this approach, we produce two sets of outputs which give different sizes of segmentation units for certain types of words.

[7] http://technology.chtsai.org/namefreq

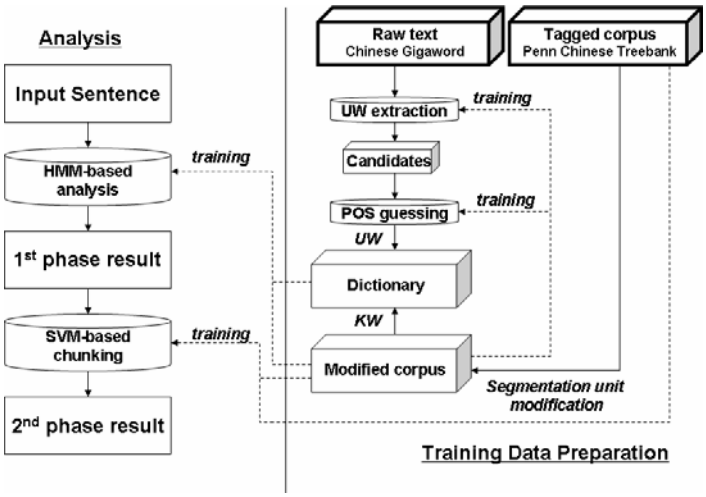

Fig. 2. Overview of two-layer morphological analysis

4.1 Minimal Unit Analysis -ChaSen

We use *ChaSen* [9] in our first layer analysis. Although *ChaSen* is originally built for Japanese language, it can be adopted easily to Chinese with slight modification. In fact, it is easier to setup the system in Chinese as we do not need to define grammar in Chinese since it does not have morphological changes such as inflection. We just need a training corpus and a dictionary for the training. The system is based on Hidden Markov Models (HMM).

Top part of Table 3 shows the results of the first layer analysis. The results are calculated based on the minimal segmentation units. [CTB4 Dic] contains only the entries extracted from CTB 4.0, which is 33,438 entries. During the first phase of manual extraction from CGW and collection from the web, we managed to increase the dictionary to 68,626 entries [+ manual extraction]. At the second phase of extraction, with auto-checking with other resources, we further increased the dictionary to 120,769 entries [+ auto extraction]. We can see the improvement on the analysis results with the increment of size of dictionary. We realize a decreasing in the unknown word rates. The row [no unknown] shows the results by retrieving all the entries from both training and testing data for building the dictionary. There are 39,896 entries in total, 6,458 entries more than [CTB4 Dic]. The training of HMM takes only the training data and the dictionary into account. The row [closed] shows the results where the training of HMM also includes the testing data. We can say that the [closed] is the perfect case of the system. From the results, we can see that our system is still far from perfect. Besides increasing the entries in the dictionary, we must also find a better way to improve the accuracy of POS tagging.

Table 3. Results of first and second layer analysis

			Segmentation			POS Tagging		
Layer	Dictionary	Unknown rate	Rec.	Prec.	F	Rec.	Prec.	F
First Layer	CTB4 Dic	9.2%	90.0	83.1	86.4	82.1	75.8	78.8
	+ manual extraction	7.4%	91.3	86.3	88.8	83.3	78.7	80.9
	+ auto extraction	5.4%	92.8	90.0	91.4	84.7	82.2	83.5
	no unknown	0%	97.1	97.8	97.4	90.1	90.7	90.4
	closed	0%	97.3	98.1	97.7	91.1	91.8	91.5
Second Layer	CTB4 Dic	-	88.5	81.1	84.6	80.2	73.6	76.7
	+ manual extraction	-	89.8	84.8	87.2	81.4	76.8	79.1
	+ auto extraction	-	91.4	88.8	90.1	83.0	80.6	81.8

4.2 CTB Unit Analysis -YamCha

The second layer takes the output from the first layer and joins the words by chunking. In order to obtain the original segmentation and POS tags, our task is to join up family names and given names, numbers, numeral type time nouns, and foreign words. The only difference with the original POS tags is that we cannot get back the original POS tags for foreign words. We used *YamCha* [10] for chunking as it is proved to be efficient for this task. This system is based on Support Vector Machines. The feature sets used are two words and POS tags at both left and right sides of the current word, plus the previous two output labels. The output labels are NR-PER-B, NR-PER-I, CD-B, CD-I, OD-B, OD-I, NT-B, NT-I, FW-B, FW-I and O.

Bottom part of Table 3 shows the results of the second layer analysis. While the results of first layer are based on minimal units, the results of second layer are based on CTB units. Since the results are based on different segmentation units, we cannot do a direct comparison. However, for a rough comparison, the difference between the first and second layer analysis is quite small. This also means that the accuracy for chunking is high since the upper bound of the second layer depends on the accuracy of the first layer. By this way, we can easily convert the minimal unit segmentation back to CTB standard.

4.3 Related Work

There are some systems which are downloadable from the web. ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System)[8] [11] is an integrated system that uses an approach based on multi-layer Hidden Markov Models. ICTCLAS provides word segmentation, POS tagging and unknown word recognition. Their experiment results show that ICTCLAS achieved 98.25% accuracy for word segmentation, 95.63% for POS tagging with 24 tags and 93.38% with 48 tags. Their system is trained on Peking University corpus.

Microsoft Research Asia (MSR) also provides a free segmenter for download (S-MSRSeg)[9] [3]. It is a simplified model and does not provide the functionalities of new word identification, morphology analysis and adaptation to various standards. They applied a source-channel approach to word segmentation, and a class-based model and context model for new word identification. They obtained 95.5–96.2% recall and 95.0–95.6% precision for word segmentation and 60.4–78.4% recall and 46.2–68.1% precision for new word identification. MSR also defines their own segmentation standard.

5 Conclusion

As a conclusion, a dictionary is very important in Chinese morphological analysis. The accuracy is worse if we have a small size dictionary. Our purpose is to build a practical system, therefore we look for some ways to enlarge the dictionary. We have increased the entries of our dictionary from 33,438 entries to 120,769 entries. However, we still wish to add more in the future as the accuracy of the system is still not near to the perfect. We have designed a two-layer morphological analyzer for Chinese text. The first layer produces minimal unit segmentation with detailed POS tags and the second layer transforms the minimal units into CTB standard. The design enables us to reduce the size of the dictionary by splitting some high productive words into smaller units. In the future, we would like to apply the same design for the analysis of compound words and morphological derived words.

References

1. Sproat, R., Emerson, T.: The First International Chinese Word Segmentation Bakeoff. In: Proceedings of Second

[8] http://www.ict.ac.cn/freeware/003_ictclas.asp
[9] http://131.107.65.76/research/downloads/default.aspx

SIGHAN Workshop. (2003) 133–143
2. Wu, A.: Customizable Segmentation of Morphologically Derived Words in Chinese. International Journal of Computational Linguistics and Chinese Language Processing 8(1) (2003) 1–28
3. Gao, J., Wu, A., Li, M., Huang, C.N., Li, H., Xia, X., Qin, H.: Adaptive Chinese Word Segmentation. In: Proceedings of ACL. (2004) 463–470
4. Brill, E.: Some Advances in Transformation-based Part of Speech Tagging. In: Proceedings of AAAI. (1994)
5. Goh, C.L., Asahara, M., Matsumoto, Y.: Pruning False Unknown Words to Improve Chinese Word Segmentation. In: Proceedings of PACLIC 18. (2004) 139–149
6. Goh, C.L., Asahara, M., Matsumoto, Y.: Machine Learning-based Methods to Chinese Unknown Word Detection and POS Tag Guessing. Journal of Chinese Language and Computing (to appear)
7. Goh, C.L., Asahara, M., Matsumoto, Y.: Chinese Unknown Word Identification Using Character-based Tagging and Chunking. In: Companion Volumn to the Proceedings of ACL 2003, Interactive Poster/Demo Sessions. (2003) 197–200
8. Zhang, H.P., Liu, Q., Zhang, H., Cheng, X.Q.: Automatic Recognition of Chinese Unknown Words Based on Roles Tagging. In: Proceedings of First SIGHAN Workshop. (2002)
9. Matsumoto, Y., Kitauchi, A., Yamashita, T., Hirano, Y., Matsuda, H., Takaoka, K., Asahara, M.: Morphological Analysis System ChaSen 2.2.9 Manual. Nara Institute of Science and Technology. (2002) http://chasen.naist.jp/.
10. Kudo, T., Matsumoto, Y.: Chunking with Support Vector Machines. In: Proceedings of NAACL. (2001) 192–199
11. Zhang, H.P., Yu, H.K., Xiong, D.Y., Liu, Q.: HHMM-based Chinese Lexical Analyzer ICTCLAS. In: Proceedings of Second SIGHAN Workshop on Chinese Language Processing. (2003) 184–187