Technical Report of *Japanese Morphological Analyzer*

Jun Jiang, Vernkin Chen

June 9, 2009

Abstract

In this TR, we first present a hybrid method for Japanese word segmentation. Word-level information is useful for analysis of known words, while character-level information is useful for analysis of unknown words, and the method utilizes both these two types of information in order to effectively handle known and unknown words. Then we discuss lemma identification in Japanese morphological analysis, which is crucial for a proper formulation of morphological analysis. Since Japanese words often have variation in orthography and the vocabulary of Japanese consists of words of several different origins, it sometimes happens that more than one writing form corresponds to the same lemma and that a single writing form corresponds to two or more lemmas with different readings and/or meanings. The current study focuses on disambiguation of heteronyms, words with the same writing form but with different word forms. To resolve heteronym ambiguity, we make use of goshu information, the classification of words based on their origin. Founded on the fact that words of some goshu classes are more likely to combine into compound words than words of other classes, we employ a statistical model based on CRFs (with hybrid method) using goshu information. The hybrid method and lemma identification are employed together in this TR.

Contents

Chapter 1

Introduction

Automatic morphological analysis is a widely-used technique for the development of NLP systems and linguistically-annotated corpora. Morphological analysis in Chinese and Japanese is an important and difficult task. In these languages, words are not separated by explicit delimiters, and word segmentation must be conducted first in most natural language processing applications. One of the problems which makes word segmentation more difficult is existence of unknown (out-of-vocabulary) words. Unknown words are defined as words that do not exist in a system's dictionary. The word segmentation system has no knowledge about these unknown words, and determining word boundaries for such words is difficult. Accuracy of word segmentation for unknown words is usually much lower than that for known words.

In this TR, we first propose a hybrid method for Chinese and Japanese word segmentation, which utilizes both word-level and character-level information. Word-level information is useful for analysis of known words, and character-level information is useful for analysis of unknown words. We use these two types of information at the same time to obtain high overall performance.

A great amount of studies have attempted to develop software for performing automatic morphological analysis with high accuracy, and several systems for Japanese morphological analysis have been freely available (Asahara and Matsumoto, 2000[1]; Kudo et al., 2004[7]).

In traditional morphological analysis on computer, the task is divided into two sub-tasks: i) segmentation of an input string into a sequence of units, and ii) assignment of a part of speech (POS) tag to each segmented unit. Linguistic analysis of corpora, however, requires more information than one provided by these sub-tasks (see e.g., Mizutani, 1983[3]). For instance, Japanese words often have variation in orthography; verb arawasu (to express) is written either as ”表わす”,”表す,” or ”あらわす,” and noun sakura (a cherry blossom) either as ”桜,” ”サクァ,” or ”さくら” In examining the frequency of words occurring in a text, linguists usually want to collapse these variants. Japanese also has lots of heteronyms[1], two or more words that have the identical writing form but different word forms. For instance, nouns namamono (raw food) and seibutu (a living thing) are both written as ”生物.” These issues are crucial for linguistic analysis of corpora, but are not handled by traditional Japanese morphological analyzers on computer. These tasks, though operating in opposite directions—one mapping two or more different writing forms onto a single word form and the other mapping one writing form onto more than one word form, can be regarded as the same problem; that is identification of lemmas, i.e., entry words in a dictionary. The task of identifying a lemma corresponding to each segmented unit in an input has been totally ignored in the study of automatic morphological analysis of Japanese.

In this paper, we also address a proper approach to Japanese morphological analysis, taking lemma identification into account. We first propose an electronic dictionary for Japanese morphological analysis, UniDic, which employs hierarchical definition of word indexes to represent orthographic variants as well as allomorphs. In this hierarchical structure, heteronyms are represented as different nodes with the same index but with different super-nodes. We then propose a statistical model for resolving heteronym ambiguity, making use of goshu information, the classification of words based on their origin. Founded on the fact that words of some goshu classes are more likely to combine into compound words than words of other classes, we employ a statistical model based on CRFs using goshu information.

[1]An example of heteronym in English is “bow,” which has two different meanings with different sounds, /bou/ and /bau/. In this paper, writing forms refer to representation in orthography, which corresponds to spelling in English. Word forms, on the other hand, are based on kana-reading and roughly correspond to sounds, although in a few cases, e.g., particles wa and e, there is dissociation between kana-reading and sound.

Chapter 2

Related Works

2.1 The Character Tagging Method

This method carries out word segmentation by tagging each character in a given sentence, and in this method, the tags indicate word-internal positions of the characters. We call such tags position-of-character (POC) tags (Xue, 2003[9]) in this paper. Several POC-tag sets have been studied (Sang and Veenstra, 1999[5]; Sekine et al., 1998[6]), and we use the 'B, I, E, S' tag set shown in Table 2.1[1].

Table 2.1: The 'B, I, E, S' Tag Set

B	The character is in the beginning of a word.
I	The character is in the middle of a word.
E	The character is in the end of a word.
S	The character is itself a word.

An example of POC-Tagging:

Sentence: 窓 の 近く に 大きな 木 が あり ます 。

POC Tag: 窓/B の/E 近/B く/E に/S 大/B き/I な/S 木/S が/S あ/B り/E ま/B す/E 。/S

ThePOC-tags can represent word boundaries for any sentences, and the word segmentation task can be reformulated as the POC-tagging task. The

[1] The 'B, I, E, S' tags are also called '$OP-CN$, $CN-CN$, $CNCL$, $OP-CL$' tags (Sekine et al., 1998[6]) or 'LL, MM, RR, LR' tags (Xue, 2003[9]).

tagging task can be solved by using general machine learning techniques such as maximum entropy (ME) models (Xue, 2003[9]) and support vector machines (Yoshida et al., 2003[8]).

This character tagging method can easily handle unknown words, because known words and unknown words are treated equally and no other exceptional processing is necessary. This approach is also used in base-NP chunking (Ramshaw and Marcus, 1995[4]) and named entity recognition (Sekine et al., 1998[6]) as well as word segmentation.

2.2 UniDic

UniDic[2] is with the aim of providing a proper tool for Japanese morphological analysis (Den et al., 2007[10])[3]. The dictionary has the following features:

1. The unit for identifying a word is based on the short unit word (Maekawa, in press), which provides word segmentation in uniform size, without being harmed by too long words.
2. The indexes for words are defined at several levels, including *lemma*, *form*, and *orthography*, which enables us to represent orthographic variants as well as allomorphs.
3. An extensive amount of phonological information, such as lexical accent and sandhi, is also described and can be utilized in speech research.

2.3 Word Segmentation Using Word-Level and Character-Level Information

For the Word-Level segmentation, taking Markov model as example (And we will use CRF). It is observed that the Markov model-based method has high overall accuracy, however, the accuracy drops for unknown words, and the character tagging method has high accuracy for unknown words but lower accuracy for known words (Yoshida et al., 2003[8]; Xue, 2003[9]). This seems

[2] Freely available at http://download.unidic.org/

[3] Throughout the paper, writing forms, i.e., indexes at the orthography level, are written in Japanese characters, and word forms, i.e., indexes at the form level, are written in Romaji. A lemma is expressed by a triple consisting of a standardized word form, a standardized writing form, and a meaning articulated in English.

natural because words are used as a processing unit in the Markov model-based method, and therefore much information about known words (e.g., POS or word bigram probability) can be used. However, unknown words cannot be handled directly by this method itself. On the other hand, characters are used as a unit in the character tagging method. In general, the number of characters is finite and far fewer than that of words which continuously increases. Thus the character tagging method may be robust for unknown words, but cannot use more detailed information than character-level information.

Then, we propose a hybrid method which combines the Word-Level method and the character tagging method to make the most of wordlevel and character-level information, in order to achieve high overall accuracy. The hybrid method is mainly based on word-level Markov models, but both POC-tags and POS-tags are used in the same time and word segmentation for known words and unknown words are conducted simultaneously. (More detail see Hybird Method chapter. *TODO*)

2.4 Goshu Information

Japanese words often have variation in orthography and the vocabulary of Japanese consists of words of several different origins, it sometimes happens that more than one writing form corresponds to the same lemma and that a single writing form corresponds to two or more lemmas with different readings and/or meanings.

Hence, the mapping from a writing form onto a lemma is important in linguistic analysis of corpora. The current study focuses on disambiguation of heteronyms, words with the same writing form but with different word forms. The paper shows that for resolving heteronyms in Japanese is to make use of goshu information. Goshu is the classification of words based on their origin. And it employed Mecab using goshu-related features and Mecab (with goshu) outperform Mecab (without goshu) in the heteronym disambiguation (see section-2.5). And It seems the Mecab is a preferable referenced JMA system while developing our JMA system.

2.5 Comparison of ChaSen and Mecab

ChaSen and Mecab are popular JMA and are with goshu information (see section-2.4).

ChaSen (Asahara and Matsumoto, 2000[1]) run with the published version of UniDic (a dictionary with the aim of providing a proper tool for Japanese morphological analysis) , which employs (an extension of) a hidden Markov model (HMM) to determine the optimal segmentation and POS assignment but can only bring a poor modeling for lemma identification, i.e., the unigram probability of lemmas given a writing form. Incorporating statistical information of goshu classes into an HMM-based analyzer, however, is problematic since in HMMs the only way to utilize goshu information is to introduce new tags that consist of combinations of POS tags and goshu classes and this will easily lead us to the data sparseness.

Mecab (Kudo et al., 2004[7]) is more recent and is based on a novel statistical method, conditional random fields (CRFs) (Lafferty et al., 2001[2]), which overcome several problems of HMMs including label bias, length bias, and difficulty in using features that can co-occur at the same position such as the POS tag and the goshu class. With the lexicon contained in ChaSen's standard dictionary and the RWCP Text Corpus as the training data, Mecab is shown to outperform ChaSen in the segmentation and POS assignment tasks.

2.6 Conclusion

From the survey, CRF are better choice than HMM and ME. Also we perform a small experiment on the CRF (from the CRF module of Jun's CMA) and the precision is 0.964 (Corpus are generated from the Mecab). And we think it is good enough for our JMA to develop based on the CRF. Also, JMA will use hybird method and the goshu information which are discussed above to gain a higher precision than 0.964.

Chapter 3

Dictionary

Dictionary plays an important role in our JMA. That's because the character tagging method (in 2.1) cannot directly reflect lexicons which contain prior knowledge about word segmentation. We cannot ignore a lexicon since over 90% accuracy can be achieved even using the longest prefix matching with the lexicon.

Basing on the approach used in MeCab, we introduce the dictionary usage in our JMA.

3.1 Source Files

The source files are listed below.

- dicrc

  The configuration file. The minimum setting in this file is item "bos-feature", which defines the feature for sentence begining symbol.

- *.csv

  The dictionary file in text format (CSV). Multiple CSV files could be defined and each line is an entry of one word. The CSV file format is described in 3.3.

- char.def

  The definition file of character category for unkown word processing. The characters are categorized into symbol, numeric, Kanji, Hiragana, Katakana, etc.

- unk.def

The definition file of POS and word cost for each character category used in unkown word processing.

- matrix.def

  The definition file for the connection matrix. Each value in the matrix represents a cost value for connecting adjacent tokens, which cost value would be used in CRF analysis.

3.2 Binary Files

3.3 CSV Format

3.4 Dictionary Build Process

Appendix A

Appendix - Project schedules and milestones

	Milestone	Start	Finish	In Charge	Description	Status
1	Survey on the project	2008-05-04	2008-05-15	Vernkin	1. Have a general overview of current Japanese Segmentation techniques and their differences. 2. Select a suitable one and do more research on it.	Finished
2	Experiment using CMA approach	2009-05-11	2009-05-15	Jun	Experiment CMA approach using only character-level information, which result reached 0.96 on ASAHI text. To improve the result further for JMA, we would adopt the hybrid approach, which utilizes both word-level and character-level information.	Finish
2	Analyze requirement specification, and study the potential solutions	2009-05-18	2009-05-22	Vernkin	Analyze requirements, and study the hybrid approach.	Finish
3	Implement the Japanese morphological analyzer	2009-05-25	2009-06-26	Jun	Implement an initial version of this project.	On going
4	Experiment and debugging	2009-06-29	2009-07-24	Vernkin	Experiment the system on training data and test data, and debugging the system.	
5	Refine the code, finish TR and doxygen document	2009-07-27	2009-07-31	Jun	Optimize and enhance the code. Integrate the documents available into TR and generate doxygen documents.	

Bibliography

[1] Masayuki Asahara and Yuji Matsumoto. Extended models and tools for high-performance part-of-speech tagger. *Proceedings of the 18th International Conference on Computational Linguistics*, pages 21–27, 2000.

[2] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[3] Shizuo Mizutani. Vocabulary (in japanese). *Asakura New Series on Japanese*, 2, 1983.

[4] Lance Ramshaw and Mitch Marcus. Text chunking using transformation-based learning. *Proceedings of the 3rdWorkwhop on Very Large Corpora*, pages 88–94, 1995.

[5] Erik F. Tjong Kim Sang and Jorn Veenstra. Representing text chunks. *Proceedings of 9th Conference of the European Chapter of the Association forComputational Linguistics*, pages 173–179, 1999.

[6] Ralph Grishman Satoshi Sekine and Hiroyuki Shinnou. A decision tree method for finding and classifying names in japanese texts. *Proceedings of the 6th Workshop on Very Large Corpora*, pages 171–177, 1998.

[7] Kaoru Yamamoto Taku Kudo and Yuji Matsumoto. Applying conditional random fields to japanese morphological analysis. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237, 2004.

[8] Kiyonori Ohtake Tatsumi Yoshida and Kazuhide Yamamoto. Performance evaluation of chinese analyzers with support vector machines. *Natural Language Processing*, 10(1):109–131, 2003.

[9] Nianwen Xue. Chinese word segmentation as character tagging. *International Journal of Computational Linguistics and Chinese*, 8(1):29–48, 2003.

[10] Hideki Ogura Atsushi Yamada Nobuaki Minematsu Kiyotaka Uchimoto Yasuharu Den, Toshinobu Ogiso and Hanae Koiso. The development of an electronic dictionary for morphological analysis and its application to japanese corpus linguistics (in japanese). *Japanese Linguistics*, 22:101–123, 2007.