A Uyghur Morpheme Analysis Method based on Conditional Random Fields

[1,2] Batuer Aisha， [3] Maosong Sun

[1,3] Department of Computer Sci. & Technology. State Key Lab on Intelligent Tech. & System. Tsinghua University, Beijing, 100084, China

[2] Research in the Key Laboratory of Multilingual Information Technology, School of Information Science and Engineering, Xinjiang University,Urumqi,Xinjiang,830046,China

{batur601,sunmaosong}@gmail.com

Abstract

Morpheme analysis is very important for Uyghur language processing. Morpheme analysis of Uyghur is quite different from other language, for this task the keys include feature selection and the design of a morpheme annotated corpus . In this paper we propose a new statistical-based Uyghur morpheme analysis method by using Conditional Random Fields (CRFs) model. The preliminary experiment results demonstrate that the proposed method is effective； the F-measure of morpheme analysis reaches 87% in the open test.

Keywords

Xinjiang, Uyghur, Morpheme, *Agglutinative language,* CRFs, Feature.

1. Introduction

Xinjiang Uyghur Autonomous Region covers over 617,763 square miles, one-sixth of China's total territory. With a population of over 21 million, Xinjiang is home to 47 ethnic groups. Uyghur is the major ethnic group with about ten million speakers in Xinjiang. Uyghur language belongs to Altaic language family.

Uyghur is based on Arabic script which consists of 32 letters, 8 vowels and 24 consonants. Each letter may have different shape at the beginning, middle, and end

of a word. Uyghur is written from right to left in text and words splits by a blank space in sentences.

From linguistic perspective Uyghur is an agglutinative language with rich and complex morphology, which is very different from morphology of other languages, such as Chinese (isolating language) and English (inflecting language).

In Uyghur the morpheme is smallest and indivisible units of semantic content or grammatical function which words are made up of. The morpheme is the smallest difference in the shape of a word that correlates with the smallest difference in word or sentence meaning or in grammatical structure, Uyghur morpheme, are not as easily recognizable. Without sufficient morphological knowledge, it is impossible to mark morphemes, for example:

بىنورمالدەك (just like abnormal)

↓ Token

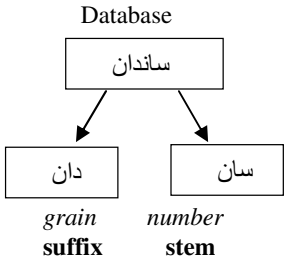
دەك نورمال بى

↓ Morpheme analysis

h بى a نورمال KQ دەك

Morpheme analysis tagging details please look at Table 1.

However, these rule-based approaches may the process due to the sophistication of Uyghur morphology. Look at a simple example:



If we use a rule-based mechanism, e.g., the finite-state automata, to treat the above Uyghur word, “database”, it would be likely to be segmented as a stem (“number”) and a suffix (“grain”) because of the presence of the suffix, this is wrong. In fact, it should be treated as a whole without further decomposition.

so we use machine learning to resolve Uyghur morpheme analysis very difficult problem.

Compared to other part of morpheme analysis tags in Uyghur, the verb is the most grammatically complex, verbs in modern Uyghur divide into many kinds of grammatical categories according to their meaning and grammatical characteristics. According to their morphological composition, they divide into root verbs, derived verbs and constructed verbs.

According to their lexical meaning, they divide into action verbs and state verbs. According to the syntactic function performed they divide into independent and helping verbs.

According to their relationships with their objects they divide into transitive and intransitive verbs. According to whether they possess certain grammatical categories (person, number, tense, voice, positivity and negativity, mood, and aspect categories), they divide into personal and impersonal verbs(Hamit.Tumur,1987;Yi ShenXiu,1998;Kurex.Mahmutjan,2003; Gülnar Eziz, 2007).For example:

ئالدىم (I got);ئالدىڭ (you [singular] got);ئالدى (he/she/it got);قۇئالدۇ (we got); لاراڭئالدى (you [plural] got); نەئالىمەن (I shall get); نەئالىمەن (I shall get); ئالىمىز (we shall get); نەئالىسەن (you [singular] will get); ۇئالىدۇ (he/she/it/they will get); ىزئالىسىز (you [singular] will get); رەئالىسىلەر (you [plural] will get).

The rest of the paper is organized as follows. We introduce structure of Uyghur language and morpheme in section 1. In Section 2 describes related works ,In Section 3 describes feature selection scheme, In Section 4, describes morphology corpus design, In Section 5, describes experiment results and finally, we conclude the paper.

2. Related works

There are traditionally two main approaches to automatic morpheme analysis: one using rules (Brill,1992) and another using statistics (Samuelsson and Voutilainen, 1997), The rules approach involves hand- made rules, which often

based on the researchers' own linguistic intuitions. The statistics approach is by using statistical data to capture the structure of the language under consideration(Gulila Altenbek, 2006).

Uyghur tokenization and morpheme analysis has been described in various researches and implemented in many solutions as it is a required preliminary stage for further processing.

These solutions include morpheme analysis(Bilikiz 2004, Yusup Abaidula 2005, Hong Mei.Niu 2007).

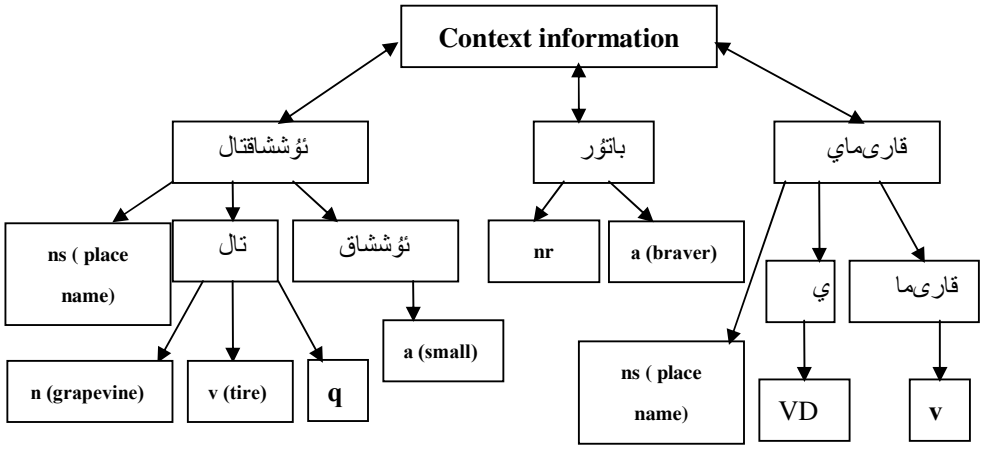
These rule-based approaches can't take full advantage of the context of useful information on the analysis of word ambiguity, in addition some existence morphology problem, e.g. سىز+auxiliary ،سەن+ ،دىمەن+ divided into noun suffix

Uyghur vocabulary and categories existence of the phenomenon, that is, a word belong to two or more different parts of speech. And type of phenomenon is quite common in Uyghur language, so lexical analysis more complicated. For example:

سىز ماڭا قارىماي تۇرۇڭ (Please don't look at me)
مەن قارىماي تۇغرىسىدا كۆپ خەۋەرلەرنى ئاڭلايدىم
(I listen more news about Karimay (Xinjiang city's name))
باشقىلارنىڭ نەرسىسىنى ئالما (Do not take other people's things)
مەن بازاردىن ئالما سېتىۋالدىم (I bought apples from the market)
ئۇ ئۆيدە بار (He was in the house)
سەن مەكتەپكە بار (You go to school)

Sixth, third and first sentences بار ,ئالما , قارىماي are verbs , second sentence قارىماي is place name and fifth sentence بار is a adjective and fourth sentence ئالما is a noun.

In the corpus, we use of context information to resolve this word ambiguity problem. Following figure demonstrates three test words in sentence, these words morpheme analysis(wrong morpheme analysis or right morpheme analysis) dependent their context information, about morpheme analysis tags please look at table1.



In the case, we use CRFs model for Uyghur morpheme analysis tagging corpus which supervised learning when the training data is morpheme analysis tagging and in order to analysis the tagging feature of Uyghur morpheme corpus. The model, by means of an abstract computational structure, implements a set of morpheme analysis tagging rules.

3. Feature selection scheme

Lexical analysis of ruled-based method although they also have a marked effect better, but because of the lack of predictive information on the context morpheme analyse tagging, especially the unknown word accuracy of morpheme analyse, and therefore the use of rule-based method can only use the limited features of the context, otherwise the data will be sparse and so on, resulting in a decline in identification accuracy.

These are questions we address with a CRFs, which use the feature of the model forms, effective use of context information, in a certain condition can be consistent with the training data probability distribution, even if is the unknown word, because of its rich context information, it also played the part of speech tagging a good prediction.

Experimental results show that the CRFs obtain a better effects of tagging than the rule-based method.

The Uyghur morpheme analysis tagging system schematic diagram is as shown in Figure 1.

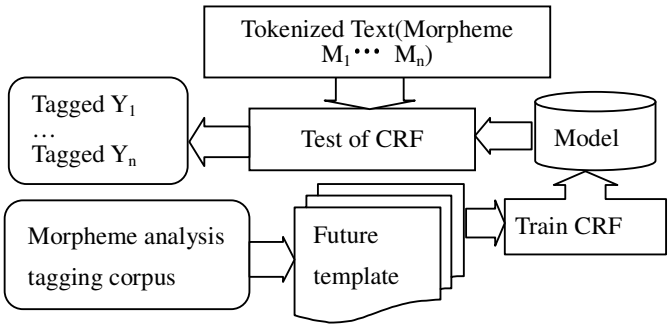


Figure 1. Uyghur morpheme analysis tagging schematic diagram

In this section, firstly we propose our novel morpheme analysis method based on CRFs. Next section, we introduce the background about linear-chain CRFs.

3.1 Features selection for morpheme analysis tagging

Conditional Random Fields are undirected graphical models trained to maximize a conditional probability first introduced by Lafferty (J.Lafferty, 2001), such as natural language processing to the indexing of the string to learning tasks.

Observation series for W = $w_1 w_2$ ... w_n, string sequence tag Y = $y_1 y_2$ ... y_n, CRF for a given string of the label, and its probability is defined as:

$$p_\lambda(Y \mid W) = \frac{1}{Z(W)} \exp\left(\sum_{t=1}^{T} \sum_{k} \lambda_k f_k \left(y_{t-1}, y_t, W, t \right) \right) \qquad (1)$$

In formula 1, Y is a string of tagging sequences, W is a string of tokenized sequences, f_k is the feature function, λ_k is a learned weight associated with feature f_k , Z(W) is the normalized factor, making the probability of all state sequences sum to one; The feature functions can measure any aspect of a state transition, y_{t-1} -> y_t, and the entire observation sequence W, centered at the current time step t .

Decoding process of CRFs, which is marked string to solve the unknown process that needs to calculate the search string on one of the largest joint probability, that is: Y* = arg max P(Y|W) The most probable label sequence for an input W(tokenized text of Uyghur) can be efficiently determined using the Viterbi algorithm. An N-best list of labeling sequences (the results of Uyghur morpheme analysis tags) can also be obtained using modified Viterbi algorithm and A* search (R.Schwartz, 1990; Fuchun Peng, 2004).

Let C0 be the current morpheme of a sentence in a manually morpheme tagged corpus, Ci be the ith morpheme surrounding C0 (i>0 indicates the right-hand side of C0 while i<0 indicates the left-hand side of C0), then the best feature template scheme for CRFs is:

Unigram feature(morpheme) templates: C0,C-1,C1

Bigram feature templates: C0C1, C-1C0, C-1C1

4. Morphology corpus design

We constructed a raw Uyghur corpus, denoted UC, with 594,172 words in UTF-8 code extracted from Uyghur websites (including Uyghur novels, story, law, education, science, conversation, etc.), these raw corpus were preprocessed and checked manually. This corpus is divided into three sub-corpora, i.e., a general-purpose sub-corpus with 162,797 words (denoted GP), a law sub-corpus with 271,372 words (denoted L), and a government sub-corpus with 160,003 words (denoted G). We further constructed two type of corpora from UC, one is manually "stemmed" UC, denoted UCS, another one is manually "lemmatized" UC, denoted UCL. For each sub-corpus, we have GPS, GPL, LS, LL, GS, GL accordingly (Batuer Aisha, 2009).

Compared to other part of morpheme analysis tags in Uyghur, the verb is the most grammatically complex because a verb communicates by way of the predicate, various actions, states, thoughts, and understandings, it influences and governs all parts of a sentence, setting their grammatical forms and purposes

(Hamit.Tumur,1987; Yi ShenXiu,1998; Kurex.Mahmutjan,2003; Gülnar Eziz, 2007). For example:

ئۇما (to walk, move in space);رۇت (to stand; to stay; to live);باشـــلا (to begin);سـانا(to count);ئاقـار (to turn white);ھئى-زد (to search);شـارىلدا (to splash, to burble);قۇيـول (to encounter);بىرلىـــك (to come together);ارقى-راۋ(to shout);يىغ-ۋك (to increase); قارشـــىئال (to welcome).

We insert these all verbs changing form to our morpheme analysis corpus, According to our statistical analysis from a large-scale Uyghur corpus can be found most cases unknown word is a noun .

The following sentences selected from originally corpus:

سىز ئاۋۇال كېلىڭ

مەن پەقەت چىدىيالمىدىم

كۆڭلىڭىزگە ئالماڭ

ئشنى باشلايلى

مەن راستتىنلا ھېرىپ ئۆلەي دېدىم

بار كۆچۈمنى چىقاردىم

راستلا شۇنداقمۇ؟

Following sentences selected our morpheme segmentation corpus:

سىز ئاۋۇال كېل ىڭ

مەن پەقەت چىدى يالمى دىم

كۆڭلى ڭىز گە ئالما ڭ

ئش نى باشلا يلى

مەن راست تىن لا ھېرىپ ئۆل ەي دي دىم

بار كۆچۈم نى چىقار دىم

راست لا شۇنداق مۇ؟

Our morpheme analysis corpus consisted of 15 basic morpheme analysing tagging increase person names nr, place name ns, organization name nt, proper nouns nz; from the linguistic point of view has also increased some of morpheme

analysing tags, beside, it also involves some tags and additional composition tags a total of 32 tags used.

The list of morpheme analysing tags used in our experiment is shown in Table 1.

Tag	Name	Examples	Tag	Name	Examples
a	adjectives	قىزىل،سۈننىي، ئېگىز	nt	Organization	ئۇيغۇرسوفت
c	conjunction	ۋە، بىلەن،چۈنكى	nz	proper nouns	قارقاش دەرياسى ، بۇغدا كۆلى
nY	Inflective Suffixes of Noun	شۇناس ،پەرۋەر	o	onomatopoeia	غاژ-غۇژ،گاژ-گۇژ، پال-پۇل
aY	Inflective Suffixes of adjective	راق،سىز	q	quantifier	يۈتۈم،سەر،چارەك
VY	Inflective Suffixes of verb	سۇن،غاي	r	pronoun	مەن،بىز،ئۇ
mY	Inflective Suffixes of numeral	لىغان،نچى	u	auxiliary	ئىدىم،-دۇ،-سىز
m	numeral	3.14،بەش،بىر	v	verb	بار-،ياز-،كۆل-
d	adverb	بۈگۈن،دەرھال،قەستەن	VD	Adverbial Forms	سېرى،كۆچە
e	interjection	پاھ،ئاپلا،ئەستاغپۇرۇللا	VN	Gerundal Forms	قۇ،مەك
h	Before the next component	بى،بەت،نا	VA	Verbal Participal Forms	دىكەن،گەن
k	followed by component	ئۆچۈن،ئارقىلىق،ئائىت	w	punctuation	. ؛ ، ؟
n	noun	تاغ،ئىرادە	y	Tone of the word	مۇ،لا،ئېھتىمال
nr	Person names	باتۇر،ئەيسا	YQ	Inflective Suffixes	ەن ، انە
ns	place name	خوتەن،بېيجىڭ،مەككە	TQ	Derivative Suffixes	لار، لەر
XQ	Possessor affixation	سى،ىم	KQ	Case affixation	گە،دە
vY	Verbal Predicative Forms	سا ، دىم ، دى	MP	Target verb	ماقچى ، مەكچى

Table 1: Morpheme analysis tags used in our experiments

The frequency of the different morpheme (part of pos) tags in the corpus as shown in Figure 2.

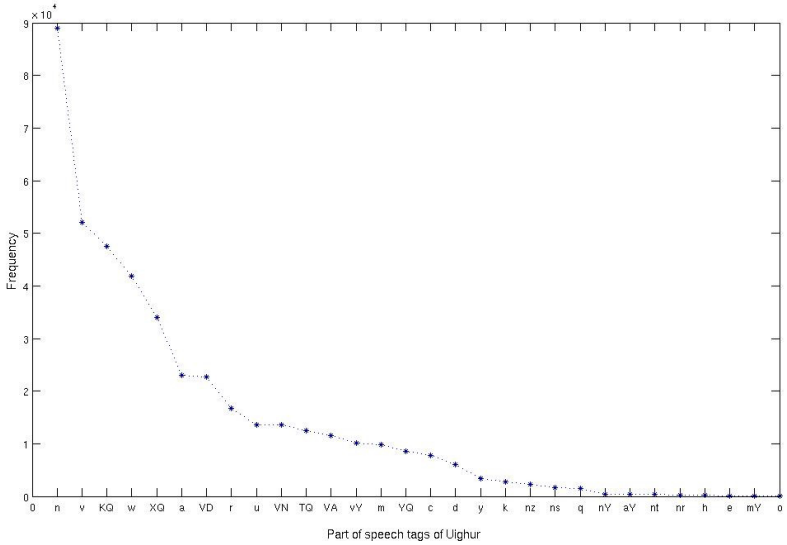


Figure 2: Tags and their frequency in tagging corpus.

We use our previous work (Batuer Aisha,2009) changing these original corpus to tokenized corpus(morpheme segmentation+harmonization), following sentences selected our tokenized corpus:

سىز ئاۋۇال كەل ىكّ

مەن پەقەت چىدا يالما دىم

كۆكّلى كّىز گە ئالما كّ

ئىش نى باشلا يلى

مەن راست تىن لا ھبرىپ ئول ەي ده دىم

بار كۇچۇم نى چىقار دىم

راست لا شۇنداق مۇ ؟

Following sentences will be appaer after morpheme analysing tagging experiment:

XQ كّ ى v كەل d ئاۋۇال r سىز

vY دىم vY يالما v چىدا y پەقەت r مەن

vY كّ v ئالما TKQ گە XQ كّىز n كۆكّلى

vY يلى v باشلا KQ نى n ئىش

vY دىم v ده vY ەي v ئۆل n ھېرىپ y لا KQ تىن a راست r مەن

vY دىم v چىقار KQ نى XQ ۇم n كۆچ a بار

w ؟ y مۇ r شۇنداق y لا a راست

5. Experiment and Results

5.1 Experimental design

The number of tokens of each UCS or UCL becomes 1,054,668, after all the punctuation marks and numerals are excluded. Similarly, that of GPS or GPL, LS or LL, GS or GL becomes 277,799, 455,621 and 321,248 respectively (Batuer Aisha,2009). To make a comprehensive evaluation, we use two of datasets in experiment. A summary of the datasets is shown in Table 2 .

Training set (number of morpheme)	**Test set (number of morpheme)**	**OOV rate (%) in test set**
733,420(GP+L)	277,799 (part of GP)	12.0
	455,621 (part of L)	10.9
	321,248 (G)	10.5

Table 2: Dataset in morpheme tagging experiment

$$\textbf{Precision} = \frac{\textbf{Correct}_{\text{number of morpheme tags}}}{\textbf{All}_{\text{number of morpheme tags}}} * 100\%$$

$$\textbf{Recall} = \frac{\textbf{Correct}_{\text{number of morpheme tags}}}{\textbf{Standard Answer}_{\text{number of morpheme tags}}} * 100\%$$

$$\textbf{F1} = \frac{2*\textbf{Precision}*\textbf{Recall}}{\textbf{Precision} + \textbf{Recall}}$$

The standard scoring program is used to calculate precision, recall, F score (**F1**), On the inside the out-of-vocabulary (OOV) rate for the test corpus, the recall on

OOV words (Roov), and the recall on in-vocabulary (Riv) words borrowed Chinese words segmentation (Richard Sproat, 2003).

5.2 Overall results

In this subsection, We borrow the tokenization approach to reformulate morpheme analysis task as a sequence labeling problem. This approach is used in both train and test step.

Table 3 shows results of morpheme analysis tagging with word precision, word recall, F measure(F score), OOV recall rate and IV recall rate.

Test corpus	Precision	Recall	F1	R_{oov}	R_{iv}
GP(closed test)	0.837	0.877	0.857	0.885	0.868
L(closed test)	0.876	0.909	0.892	0.915	0.901
G(open test)	0.850	0.890	0.870	0.899	0.878

Table 3: Results of Morpheme analysis tagging

Experiment results demonstrate our morpheme analysis tagging based on statistical method better than rule-based method and our CRFs-based approach can yield promising performances.

5.3 Error Analysis

Firstly, L documents morpheme analysis tagging accuracy higher two documents, because L documents structures and style very formal no unknown words ,many sentences and words repeated ,it is very useful to train models .

Secondly, G documents morpheme analysis tagging accuracy lower L documents because include so many person name (mainly Chinese person name and Uyghur person name), organization name and place name. If Chinese person name ئەنچىڭپىڭ (Chingping An) tokenized right, thus morpheme analysis right and vice versa.

For example:

ئەنچىكىپكە tokenized wrong(ئەن چىكە پ ى كە), so following morpheme analysis results is wrong(ئەن a چىكە VD پ a ى كە XQ).

Thirdly, GP documents morpheme analysis tagging accuracy lower than other two documents accuracy, because include so many unknown word especially new words (about science and medicine etc.), person name (Uyghur person name, Chinese person name , foreign person name), organization name and place name (include China and other country's city name).

6. Conclusions

Exist Uyghur word morpheme analysis methods (Bilkiz, 2004; Yusup Abaidula, 2005; Gulila Altenbek, 2006; Hong Mei.Niu, 2007) are mainly rule-based. To the best of our knowledge, our method is the first work of statistical-based morpheme analysis method of Uyghur language. Experiments prove that the problem with word ambiguity to be resolved with our statistical-based method, so it is more effective than rule-based one, the accuracy of word morpheme analysis reach 90%. Contrary to rule-based methods, the advantage of our method is effective use of context feature and information, which can be consistent with the training of the probability distribution of data, morpheme analysis of the sentence has played a very good prediction. This will be very useful to development of Uyghur language syntax analysis and related research.

7. Acknowledgement

We would like to thank the anonymous reviewers for their valuable comments and suggestions during the research.

8. References

Hamit.Tumur, “Modern Uyghur Grammar (morphology)”, Beijing National Publishing House, 1987(in Uyghur)

Bilkiz, “Uyghur Corpus experimental study of part of speech tagging”, Xinjiang fifth Annual Youth Symposium, Xinjiang People's Publishing House in October 2004 (in Chinese)

Hong Mei.Niu, Jamila.Wushur, Turgun.Ibrayim , “Modern Uyghur proofing of-speech tagging technology research”, Yili Normal University (Natural Science Edition) 2007 ,pp. 43-46 (in Cinese)

Kurex.Mahmutjan,Amine. Litip,Yari.Abaydulla, “Modern Uyghur Language” First Edition. Xinjiang People's Press, 2003(in Uyghur)

Yi ShenXiu, Gao ShiJie, “Uyghur Grammar”, Beijing National Publishing House, 1998 (in Chinese)

J.Lafferty, A.McCallum, and F.Pereira. “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data”, In Proceedings of the 18th International Conf. on Machine Learning, 2001, pp. 282–289.

Gülnar Eziz, “Resistance to Borrowing of Uyghur Verbs”, Annual Conference University of Washington, October18-21, 2007, pp. 1-15

Fuchun Peng , Fangfang Feng , PAndrew McCallum , “Chinese segmentation and new word detection using conditional random fields”, Proceedings of the 20th international conference on on Computational Linguistics, 2004,pp. 562-568

Gulila Altenbek, “Automatic Morphological Tagging of Contemporary Uyghur Corpus”, The 2006 IEEE International Conference on Information Reuse and Integration, 2006, pp. 557-560

R.Schwartz and Y.Chow, The N-best Algorithm:An Efficient and Exact Procedure for Finding the N most Likely Sentence Hypotheses. In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 1990.

Yusup Abaidula, Rezwangul, Abdiryim Sali “The Research and Development of Computer Aided Contemporary Uyghur Language Tagging System” Volume Journal of Chinese Language and Computing pp.203-210, 2005

Batuer Aisha , Maosong Sun,A Statistical Method for Uyghur Tokenization,In Proceedings of IEEE International Conference on Natural Language Processings and Knowledge Engineering,2009

Richard Sproat and Thomas Emerson, The First International Chinese Word Segmentation Bakeoff, The Second SIGHAN Workshop on Chinese Language Processing, Sapporo, Japan, July 2003