CSeg&Tag1.0: A Practical Word Segmenter and POS Tagger for Chinese Texts

Sun Maosong, Shen Dayang, Huang Changning
National Key Lab. of Intelligent Technology & Systems
Department of Computer Science
Tsinghua University
Beijing 100084, P.R. China
lkc-dcs@mail.tsinghua.edu.cn

Abstract

Chinese word segmentation and POS tagging are two key techniques in many applications in Chinese information processing. Great efforts have been paid to the research in the last decade, but unfortunately, no practical system with high performance for unrestricted texts is available up to date. CSeg&Tag1.0, a Chinese word segmenter and POS tagger which unifies these two procedures into one model, is introduced in this paper. The preliminary open tests show that the segmentation precision of CSeg&Tag1.0 is about 98.0% - 99.3%, POS tagging precision about 91.0% - 97.1%, and the recall and precision for unknown words are ranging from 95.0% to 99.0% and from 87.6% to 95.3% respectively. The processing speed is about 100 characters per second on Pentium 133 PC. The work of improving the performance of the system is still ongoing.

1. Background and the Related Issues

In Chinese, there do not exist delimiters, such as spacing in English, to explicitly indicate boundaries between words. Chinese word segmentation is therefore proposed as the first step in any Chinese information processing systems. Then we still face the problem of part-of-speech tagging. These two issues have been intensively studied by the Chinese language computing community in the last decade[1-18]. Unfortunately however, no word segmenter and POS tagger for Chinese with satisfactory performance in treating unrestricted texts are available so far.

Two main obstacles block the progress of Chinese word segmentation: one is *ambiguity*, another is u*nknown word*. The sentences in (1) are examples of ambiguity and the sentence (2) and (3) examples of unknown word.

(1a) 这个**研究所**很有名.

(1b) 这项**研究所**涉及的问题很复杂.

At least two explanations are possible for the fragment "研究所" in (1), resulting in two different segmentations:

correct segmentation for (1a)
这 | 个 | **研究所** | 很 | 有名
this CLASSIFIER institute very famous
(This institute is very famous.)

correct segmentation for (1b)
这 | 项 | **研究** | **所** |
this CLASSIFIER research AUX
涉及 | 的 | 问题 | 很 | 复杂 |.
involve of problem very complex
(The problems involved in this research are very complex.)

Two transliterated foreign personal names(*TFN*), i.e., "穆巴拉克" and "阿斯马特 · 阿卜杜勒 · 马吉德" are involved in the sentence (2):

(2) 随同**穆巴拉克**总统来访的有总理**阿斯马特 · 阿卜杜勒 · 马吉德**, ...

They will be wrongly broken into pieces of isolated characters if not processed:

correct segmentation for (2)
随同 | **穆巴拉克** | 总统 | 来访 | 的 |
accompany TFN1 president visit of
有 | 总理 | **阿斯马特 · 阿卜杜勒 · 马吉德** ...
have premier TFN2
(Visitors accompanying the president TFN1 include the premier TFN2, ...)

wrong segmentation for (2)
随同 | **穆** | **巴** | **拉** | **克** | 总统 | 来访 | 的
| 有 | 总理 | **阿** | **斯** | **马** | **特** | · | **阿** |
卜 | **杜** | **勒** | · | **马** | **吉**| **德** ...

The sentence (3) contains a Chinese personal name(*CN*) “单清”：

(3) 单清楚楚动人.

We have:

<u>correct segmentation for (3)</u>

单清 | 楚楚动人

CN *beautiful*

(CN is beautiful)

<u>wrong segmentation for (3)</u>

单 | 清楚 | 楚 | 动人 |

only clear ChineseSURNAME touching

/ logically ill-formed sentence */*

POS tagging for Chinese is similar to that of English, except that an English tagger only need to tag one word sequence for an input sentence, but in the case of Chinese, to get a correct tag sequence for a sentence, a Chinese tagger may be requested to tag more than one word sequences simultaneously due to the presence of segmentation ambiguities.

Chinese word segmentation and POS tagging techniques can be found many applications in the real world such as information retrieval, text categorization, text proofreading, OCR, speech recognition and text-to-speech conversion systems. For instance, in information retrieval, the incorrect segmentation for the fragment “研究所” in (1a) and (1b) will definitely cause improper access to the texts involving it. Another typical application is in text-to-speech conversion. The over-segmentation of *TFN1* and *TFN2* in the sentence (2) will result in the synthesized speeches choppy. The *CN* in (3) may make the word segmentation and POS tagging of the whole sentence totally wrong, and further, the pronunciation of the character 单 totally wrong (单 should be pronounced as *shan4* if it is referred to a surname, whereas as *dan1* if adjective or adverb).

2. The Complexity of the Task

<u>Combinatorial Explosion 1: Word Segmentation Candidate Space</u>

The number of possible segmentations for some sentences may be rather large. Observe:

(4) 发展中国家用电器材料…

totally 76 possible segmentations will be found if we simply match the sentence with a dictionary:

(seg1) 发展中国家 | 用 | 电器 | 材料 |

(seg2) 发展 | 中国 | 家用 | 电器 | 材料 |

……

(seg75) 发展 | 中 | 国家 | 用电 | 器材 | 料 |

(seg76) 发 | 展 | 中 | 国 | 家 | 用 | 电 | 器 | 材 | 料 | …

Fig.1 shows the word segmentation candidate space for the sentence (4).

The situation will be even complicated as unknown words is under consideration(Fig.2).

Generally, segmentation ambiguities can be classified into three categories:

(a) ambiguities among common words(refer to all arcs in Fig.1)

(b) ambiguities among unknown words(see arcs of representing candidates for Chinese place name and for Chinese personal names in Fig.2)

(c) ambiguities among common words and unknown words(see arcs across Chinese personal name candidates “江爱”, “林江爱”, “王林江爱” and the arc across common word “爱”(love, like) in Fig.2)

In our experience, ambiguities of type (a) will cause about 3% loss on the precision rate of segmentation in the condition of making use of maximal matching strategy, one of the most popular methods employed in word segmentation systems, and type (b) and (c) about 10.0% loss if the processing of unknown words is ignored (unfortunately, type (b) and (c) have received less attention than type (a) in the literature).

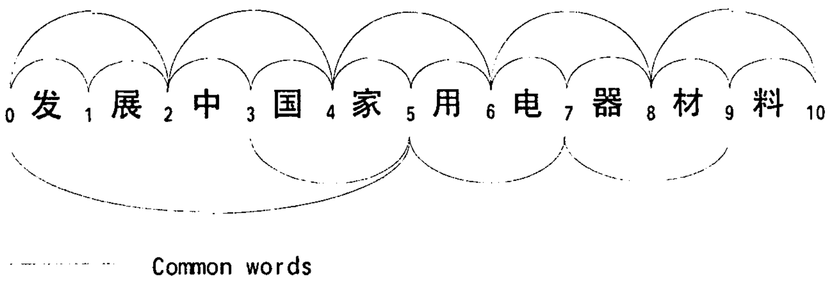

Fig.1 The word segmentation candidate space

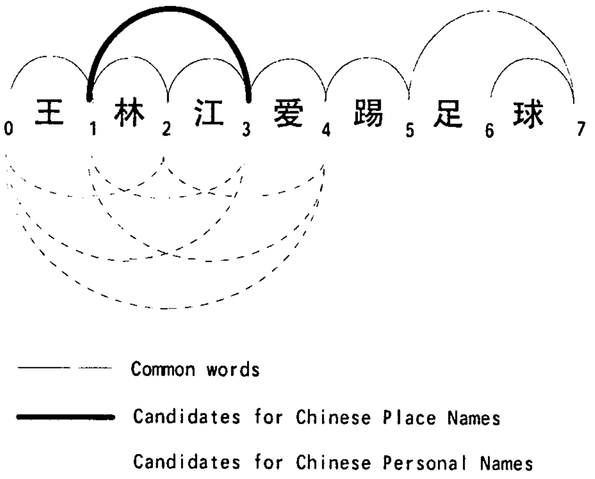
0 王 1 林 2 江 3 爱 4 踢 5 足 6 球 7
Common words
Candidates for Chinese Place Names
Candidates for Chinese Personal Names

Fig.2 The word segmentation candidate space regarding unknown words

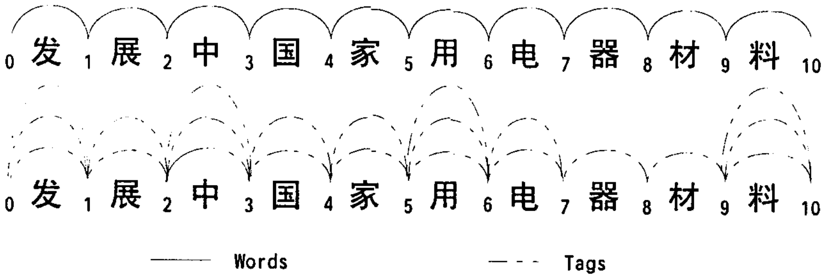
0 发 1 展 2 中 3 国 4 家 5 用 6 电 7 器 8 材 9 料 10
0 发 1 展 2 中 3 国 4 家 5 用 6 电 7 器 8 材 9 料 10
Words
Tags

Fig.3 The POS tagging candidate space

Combinatorial Explosion 2: POS Tagging Candidate Space

Given that:

TAG(发) = {vgm, qnq, ngm}
TAG(展) = {vgm, ns}
TAG(中) = {j, dm, vgm)
TAG(国) = {ngm, ns}
TAG(家) = {ngm, k}
TAG(用) = {vgm, um, pgm}
TAG(电) = {vgm, ngm}
TAG(器) = {ngm}
TAG(材) = {ngn}
TAG(料) = {ngm, vgm, qnq}

we will get 1296 possible tag sequences solely for seg(76) in the sentence 4 (Fig.3).

Combinatorial Explosion 1 × Combinatorial Explosion 2: An Integrated Model

We find out through experiments that the word segmentation and POS tagging are mutually interacted, the performance of the both will increase if they are integrated together[18]. Scholars ever tried to do so. The method reported in [11] is: (a) finding out the N-best segmentation candidates explicitly in terms of word frequency and length; (b) POS tagging each of the N-best segmentation candidates, resulting in the N-best tag sequences accordingly; and (c) using a score with weighted contributions from (a) and (b) to select the best solution. Note that the model used in (a) is just word unigram, and (a) and (b) are being done successively (denoted as "(a)+(b)"). It is a kind of pseudo-integration. More truly one, in our point of view, should be: (a) taking all segmentation possibilities into account; (b) expanding every segmentation candidate of the input sentence into a number of tag sequences one by one, deriving a considerable huge segmentation and tagging candidate space; and (c) seeking the optimal path over such space with a bigram model, obtaining then both word segmentation and POS tagging result from the path found. In the case, (a) and (b) are being done simultaneously (denoted by "(a)||(b)"). We regard this as a basic strategy and testbed for conducting our system. Obviously, a much more serious combinatorial problem is encountered here.

3. CSeg&Tag1.0: System Architecture and Algorithm Design

Although great efforts have been paid to the related researches by Chinese information processing community in the last decade, we still have not a practical word segmenter and POS tagger at hand yet. What is the problem? The crucial reason, we believe, lies in the "knowledge". As indicated in section 2, we meet a very serious difficulty, without relevant knowledge, even humanbeings will definitely fail to solve it. The focus of the research should be no longer solely on the 'pure' or 'new' formal algorithms — no matter what it will be, instead, what is urgently required is on two issues, i.e., (1) what sorts of and how many knowledges are needed; and (2) how these various konwledges can be represented, extracted, and cooperatively mastered, in a system.

This is also the philosophy in designing Cseg&Tag1.0, an integrated system for Chinese word segmentation and POS tagging, which is being developed at the National Key Lab. of Intelligent Technology and Systems, Tsinghua University. The aim of CSeg&Tag is to be able to process unrestricted running texts. Fig.4 gives its architecture.

Roughly speaking, Cseg&Tag1.0 can be viewed as a three-level multi-agent(the concept of "agent" means an entity that can make decision independently and communicate with others) system plus some other necessary mechanisms. They are: (1) agents at the low level for treating unknown words; (2) a competition agent at the intermediate level for resolving conflicts among low level agents; (3) a bigram-based agent at the high level for coping with all the remaining ambiguities; (4) mechanisms employing the so-called "global statistics" and "local statistics" (cache); and (5) a rule base. We will introduce them briefly in turn(the detailed discussion of each part is beyond the scope of this paper).

3.1. Agents at the Low Level for Treating Unknown Words

The types of unknown words CSeg&Tag1.0 currently concerns include *Chinese personal names(CN)*, *transliterated foreign personal names(TFN)* and *Chinese place names(CPN)*. They can not be enumerated in any dictionary even with numerous size.

To study unknown words systematically, we build up there relevant banks:

- *CN Bank(CNB)*: 200,000 samples
- *TFN Bank(TFNB)*: 38,769 samples
- *CPN Bank(CPNB)*: 17,637 samples

The difficulty of identifying unknown words in Chinese arises from characteristics of them:

(a) no any explicit hint such as capitalization in English exists to signal the presence of unknown words, and the character sets used for unknown words are strict subsets of Chinese characters(the size of the complete Chinese character set is 6763), with some degree of decentralized distributions;

	# of chars in char set
CN (surname)	*729*
CN (given name)	*3345*
TFN	*501*
CPN	*2595*

(b) the length of unknown words may vary arbitrarily;

(c) some characters used in unknown words may also be used as mono-syllabic common words in texts;

(d) the mono-syllabic words identified above fall into the syntactic categories not only notional words but also function words;

(e) the character sets are mutually intersected to some extent;

(f) some multi-syllabic words may occur in unknown words.

In our system, three agents, *CNAgent*, *TFNAgent* and *CPNAgent* are set up to be responsible for finding candidates in input texts accordingly. A candidate can be regarded as a "guess" with a value of belief. Three steps are involved in all the three agents in general:

Step 1: Applying MM(maximal matching) first as a pre-processing, then finding candidates over the resulting fragments of characters

There are two strategies for seeking candidates in the input sentence. One is simply viewing it as character string, finding candidates over whole of it in terms of the relevant character set:

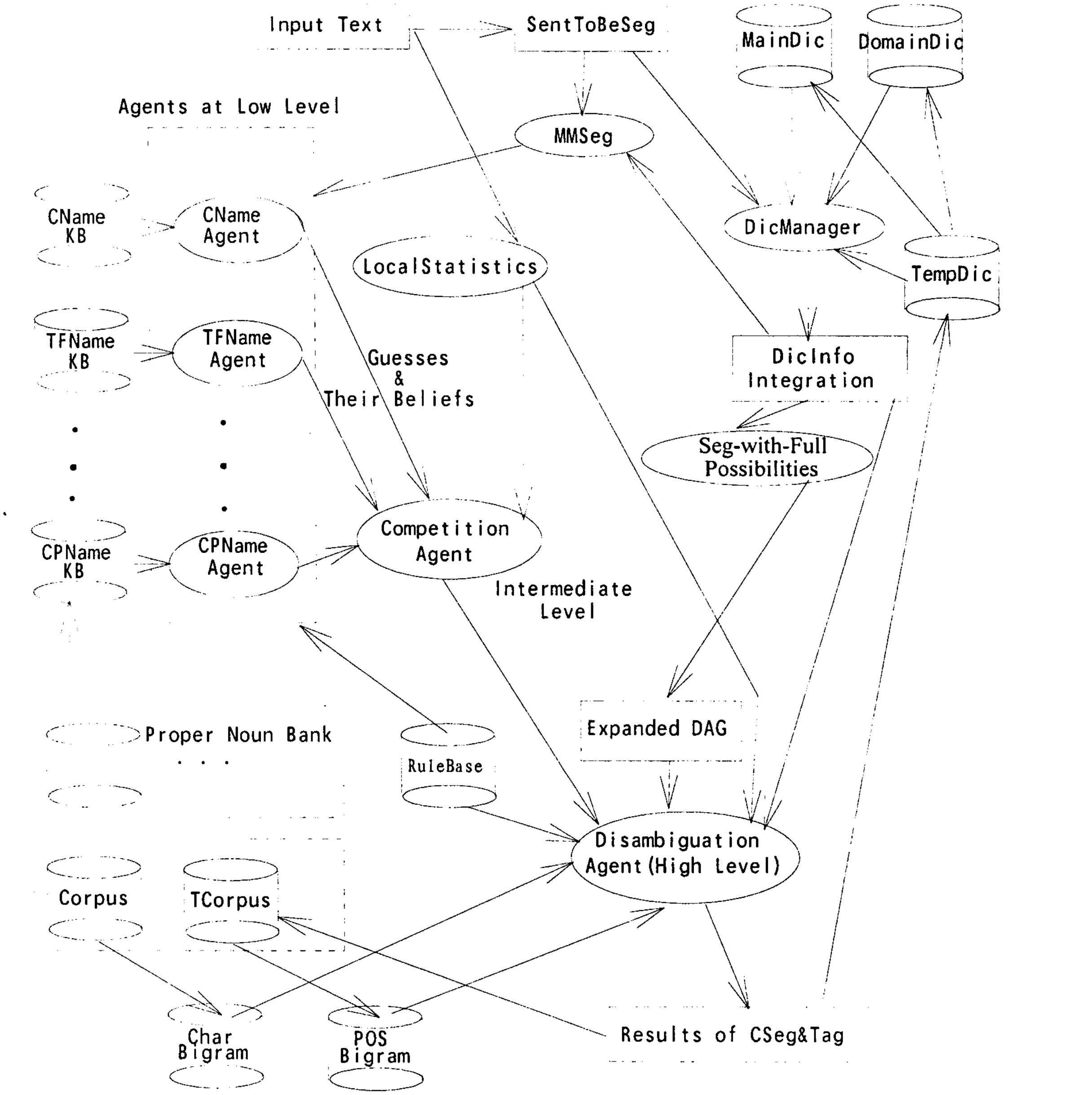

Fig.4 The system architecture of CSeg&Tag1.0

(5a) 王雪芝要来参加中国科学杂志社的庆典
CN1 CN2 CN3

Many noises will be unnecessarily introduced, as *CN2* and *CN3* in (5a). Another way is viewing input as word string, applying *MM* segmentation as a pre-processing first, then trying to find candidates only over the fragments composed of successive single characters:

(5b) 王 | 雪 | 芝 | 要 | 来 | 参加 | 中国 |
CN1 will come attend China
科学 | 杂志社 | 的 | 庆典
science journal of celebration
(CN1 will come here and attend the celebration of the journal of "Science in China")

Step 2: Drawing back some multi-syllabic words into the candidates

Look at:

(6) 他叫白金汉希尔
(His name is Buckinghamshire)

after *MM*, we get

他 | 叫 | 白金 | 汉 | 希 | 尔 |

obviously, 白金(*platinum*) should be drawn back and added into the *TFN* candidate.

Such multi-syllabic words can be collected from the banks.

Step 3: Further determining boundaries of the candidates

All of the useful information, usually language-specific and unknown-word-type-specific, are activated to perform this work.

internal information

(i) statistical information

Each candidate will be assigned a belief according to the statistics derived from the banks.

(ii) structural information

# nature of characters

· absolute closure characters for *CNs*

They will definitely belong to a Chinese surname once falling into the control domain of it:

李逵　　郑筱云　　刘景藜

· relative closure characters for *CNs*

In certain conditions, they function as absolute closure characters:

(7a) 胡戎睿 | 十分 | 聪明
CN1　　*very*　*clever*
(CN1 is very clever)

(7b) 胡戎 | 睿智 | 过人
CN2　*clever*　*very*
(CN2 is very clever)

· open characters for *CNs*

For this sort of characters, possibilities of being included in a name and excluded out of the name must be reserved:

(8a) 张玉爱 | 读 | 小说
CN1　　*read*　*novel*
(CN1 is reading a novel)

(8b) 张玉 | 爱 | 读 | 小说
CN2　*like*　*read*　*novel*
(CN2 likes to read novels)

# position in unknown words

For instance, "在" always occurs in the first position of given name of *CNs*, illustrated as "于在河""王在明""陈在铁""金在荣""谭在树""白在桥""邓在军". The *CN* candidate "邓军在" in (9)

(9)　邓军在唱歌

will be therefore properly filtered out, leaving the correct one: "邓军".

# affix

Affix(e.g. suffix of *CPNs*) will be beneficial to locating the boundaries of some unknown words.

# constructions

CPNs　==>　Chinese surname + "家" +
　　　　　　mono-syllabic *CPN* suffix

external information

(i) statistical information

Refer to "global & local statistics".

(ii) structural information

# titles

# special verbs

# special syntactic patterns

patten x0: "以 < *CN* or *TFN*> {title} 为 <title> "

(10)　以潘杜泉为团长的香港代表团 ...

The fragment "潘杜泉为" in (10) will create four *CN* candidates "潘杜""潘杜泉""杜泉""杜泉为", but only "潘杜泉" passes under the constraint of pattern x0.

3.2. The Competition Agent at the Intermediate Level for Resolving Conflicts among Low Level Agents

The candidates given independently by three agents may contradict each other on some occasions (see Fig.2). We observe from 497 randomly selected sentences that low level agents generate multiple(>=2) unknown word candidates in 17.7% of them(Fig.5), and, the probability of conflicting is about 88% if candidate number is 2 and 100% if it is greater than 2(Fig.6).

A *competition agent* is established to deal with such conflicts. The evaluation is based on all information from various resources, that is:

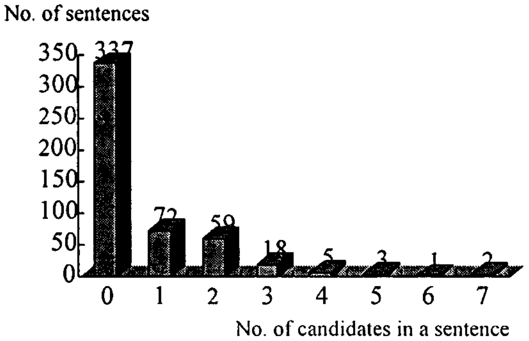

Fig.5 The distribution of candidates in sentences

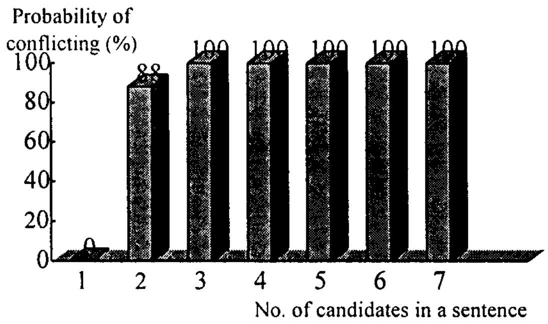

Fig.6 The probability of conflicting among unknown word candidates

$$Eval(candidate) = f_{\sum}(InterStatisInfo, InterStrucInfo, ExterStatisInfo, ExterStrucInfo)$$

About 77% conflicts can be solved by this agent. The output of it, including correct candidates and some unsolved conflicts, are then sent to a high level agent for further processing.

3.3. The Bigram-based Agent at the High Level for Coping with all the Remaining Ambiguities

The conventional POS bigram model and a dynamic programming algorithm are used in this high level agent. The searching space of the algorithm is the complete combination of all possible word and tag sequences, and the complexity of it can be theoretically and experimentally proved still polynomial.

3.4. Global Statistics & Local Statistics

Global statistics are referred to statistical data derived from very large corpora, as mutual information and t-test in Cseg&Tag1.0, whereas local statistics to those derived from the article in which the input sentence stands — like a chche. Both of them take characters as basic unit of computation, because any Chinese word is exactly a combination of characters in one way or another. Experiments by us reveal that they(especially the latter) are quite important in the resolution of ambiguities and unknown words. Refer back to "张玉" and "张玉爱" in (8a) and (8b) as an example. The both *CN* candidates are reasonable given the isolated sentence only, but by cache, it is in fact a collection of ambiguous entities unsolved so far in the current input article, the algorithm will have more evidence to make decision. We will discuss this in depth in another paper.

3.5. Rule Base

It contains knowledge in rule form, including almost all word formation rules in Chinese, a number of simple but very reliable syntactic rules, and some heuristic rules.

4. Experimental Results

Cseg&Tag1.0 is implemented in Windows environment with Visual C++1.0 programming language. The dictionary supporting it contains 60,133 word entries along with word frequencies, parts of speech, and various types of information necessary for the purpose of segmentation and tagging. The size of manually tagged corpus for training the bigram model is about 0.4 M words, and that of the raw corpus for achieving global statistics is 20M characters.

We define:

$$Seg.precision = \frac{\#\,words-correctly-segmented}{\#\,words-in-input-texts}$$

$$Tag.precision = \frac{\#\,words-correctly-tagged}{\#\,words-in-input-texts}$$

The preliminary open tests show that for CSeg&Tag1.0, the word segmentation precision is ranging from 98.0% to 99.3%, POS tagging precision from 91.0 to 97.1%, and the recall and precision for unknown words are from 95.0% to 99.0% and from 87.6% to 95.3% respectively. The speed is about 100 characters per second on Pentium 133. A running sample of Cseg&Tag1.0 is demonstrated as follows(tokens underlined in the output are unknown words successfully identified while those in bold are words wrongly tagged):

[input text]
<全国政协举行新年茶话会江泽民发表讲话>（记者邹爱国、何平） 政协全国委员会今天上午在政协礼堂举行新年茶话会，迎接九十年代的到来。江泽民、李鹏、万里、姚依林、宋平、李瑞环、王震、田纪云、李锡铭、丁关根、温家宝等国家领导人，...

[output]
<\xg 全国\sd 政协\j 举行\vgd 新年\td 茶话会\ngd 江泽民\np 发表\vgd 讲话\ngd >\ xg （\xp 记者\ngd 邹爱国\np 、\xp 何平\np ）\xp 政协\j 全国\sd 委员

会\ngd 今天\td 上午\td **在\vgm** 政协\j 礼堂\ngd 举行\vgd 新年\td 茶话会\ngd ，\xp 迎接\vgd 九十\mx 年代\ngd 的\ed 到来\vgd 。\xs 江泽民\np 、\xp 李鹏\np 、\xp 万里\np 、\xp 姚依林\np 、\xp 宋平\np 、\xp 李瑞环\np 、\xp 王震\np 、\xp 田纪云\np 、\xp 李锡铭\np 、\xp 丁关根\np 、\xp 温家宝\np 等\ egm 国家\ngd 领导人\ngd ，\xp …

It should be pointed out that Cseg&Tag1.0 is just the result of the first round of our investigation. To get our goal, i.e., developing a system with approximately 99% segmentation precision and 95% tagging precision for any running Chinese texts in any cases, quite a lot of work is still waiting there to be done. What we can say now is that we believe it is possible to reach this destination in a not very far future, and we know more than before about how to approach it. The second round work is ongoing currently, with emphasis on two aspects: (1) to promote the algorithm, particularly those associated with agents and cache, carefully; (2) to improve the quality of knowledge base by both enlarging the size of the relevant resources(textual corpora, unknown word banks, etc.) and refining the lexicon, tagged corpus and the rule base.

Acknowledgment

This research is supported by the National Natural Science Foundation of China and by the Youth Science Foundation of Tsinghua University, Beijing, P.R.China.

References

[1] N.Y. Liang, "Automatic Chinese Text Word Segmentation System — CDWS", *Journal of Chinese Information Processing* , Vol.1, No.2, 1987

[2] C.K. Fan, W.H. Tsai, "Automatic Word Identification in Chinese Sentences by the Relaxation Technique", *Computer Processing of Chinese and Oriental Languages*, Vol.1, No.1, 1988

[3] C. Kit, Y. Liu, N. Liang, "On Methods of Chinese Automatic Word Segmentation", *Journal of Chinese Information Processing*, Vol.3, No.1, 1989

[4] J.S. Zhang, Z.D. Chen, S.D. Chen, "A Method of Word Identification for Chinese by Constraint Satisfaction and Statistical Optimization Techniques", *Proc. of ROCLING-IV*, Kenting, 1991

[5] J.S. Chang, S. Chen, Y. Zheng, X.Z. Liu, S.J. Ke, "A Multiple-Corpus Approach to Identification of Chinese Surname-names", *Proc. of Natural Language Processing Pacific Rim Symposium*, Singapore, 1991

[6] B.Y. Lai, S. Lun, C.F. Sun, M.S. Sun, "A Tagging-Based First Order Markov Model Approach to Chinese Word Identification", *Proc. of ICCPCOL-92,* Florida, 1992

[7] K.J. Chan, S.H. Liu, "Word Identification for Mandarin Chinese Sentences", *Proc. of COLING-92*, Nantes, 1992

[8] L.J. Wang, et al. "Recognizing Unregistered Names for Mandarin Word Identification", *Proc. of COLING-92,* Nantes, 1992

[9] M.S. Sun, B.Y. Lai, S. Lun, C.F. Sun, "Some Issues on Statistical Approach to Chinese Word Identification", *Proc. of 3rd International Conference on Chinese Information Processing*, Beijing, 1992

[10] C.H. Chang, C.D. Chen, "HMM-based Part-of-Speech Tagging for Chinese Corpora", *Proc. of the Workshop on Very Large Corpora*, Ohio, 1993

[11] C.H. Chang, C.D. Chen, "A Study on Integrating Chinese Word Segmentation and Part-of-Speech Tagging", *Communications of COLIPS*, Vol.3, No.2, 1993

[12] M.S. Sun and W.J. Zhang, "Transliterated English Name Identification in Chinese Texts",*Computational Linguistics: Research & Application*, Beijing Language Institute Press, Beijing, 1993

[13] M.S. Sun, C.N. Huang, H.Y. Gao, J. Fang, "Identifying Chinese Names in Unrestricted Texts", *Communications of COLIPS*, Vol.4, No.2, 1994

[14] R. Sproat, C. Shih, W. Gale, N. Chang, "A Stochastic Finite-State Word Segmentation Algorithm for Chinese", *Proc. of 32nd Annual Meeting of ACL*, New Mexico, 1994

[15] D.Y. Shen, M.S. Sun and C.N. Huang, "Identifying Chinese Place Names in Unrestricted Texts", *Computational Linguistics: Research & Development*, Tsinghua University Press, Beijing, 1995

[16] J.Y. Nie, M.L. Hannan, W. Jin, "Unknown Word Detection and Segmentation of Chinese Using Statistical and Heuristic Knowledge", *Communications of COLIPS*, Vol.5, No.1 , 1995

[17] M.S. Sun, B.K. T.sou, "Resolving Ambiguities in Chinese Word Segmentation", *Proc. of PACLIC-10*, Hong Kong, 1995

[18] M.S. Sun, C.N. Huang, "Word Segmentation and Part-of-speech Tagging for Unrestricted Chinese Texts", *A Tutorial* on the *International Conference on Chinese Computing'96*, Singapore, 1996