Achilles: NiCT/ATR Chinese Morphological Analyzer for the Fourth Sighan Bakeoff

Ruiqiang Zhang[1,2] and **Eiichiro Sumita**[1,2]
[1]National Institute of Information and Communications Technology
[2]ATR Spoken Language Communication Research Laboratories
2-2-2 Hikaridai, Seiika-cho, Soraku-gun, Kyoto, 619-0288, Japan
{ruiqiang.zhang,eiichiro.sumita}@atr.jp

Abstract

We created a new Chinese morphological analyzer, Achilles, by integrating rule-based, dictionary-based, and statistical machine learning method, conditional random fields (CRF). The rule-based method is used to recognize regular expressions: numbers, time and alphabets. The dictionary-based method is used to find in-vocabulary (IV) words while out-of-vocabulary (OOV) words are detected by the CRFs. At last, confidence measure based approach is used to weigh all the results and output the best ones. Achilles was used and evaluated in the bakeoff. We participated the closed tracks of word segmentation and part-of-speech tagging for all the provided corpus. In spite of an unexpected file encoding errors, the system exhibited a top level performance. A higher word segmentation accuracy for the corpus ckip and ncc were achieved. We are ranked at the fifth and eighth position out of all 19 and 26 submissions respectively for the two corpus. Achilles uses a feature combined approach for part-of-speech tagging. Our post-evaluation results prove the effectiveness of this approach for POS tagging.

1 Introduction

Many approaches have been proposed in Chinese word segmentation in the past decades. Segmentation performance has been improved significantly, from the earliest maximal match (dictionary-based) approaches to HMM-based (Zhang et al., 2003) approaches and recent state-of-the-art machine learning approaches such as maximum entropy (MaxEnt) (Xue and Shen, 2003), support vector machine (SVM) (Kudo and Matsumoto, 2001), conditional random fields (CRF) (Peng and McCallum, 2004), and minimum error rate training (Gao et al., 2004). After analyzing the results presented in the first and second Bakeoffs, (Sproat and Emerson, 2003) and (Emerson, 2005), we created a new Chinese word segmentation system named as "Achilles" that consists of four modules mainly: Regular expression extractor, dictionary-based Ngram segmentation, CRF-based subword tagging (Zhang et al., 2006), and confidence-based segmentation. Of the four modules, the subword-based tagging, differing from the existing character-based tagging, was proposed in our work recently. We will give a detail description to this approach in the following sections.

In the followings, we illustrate our word segmentation process in Section 2, where the subword-based tagging is implemented by the CRFs method. Section 3 illustrates our feature-based part-of-speech tagging approach. Section 4 presents our experimental results. Section 5 describes current state-of-the-art methods for Chinese word segmentation. Section 6 provides the concluding remarks.

2 Introduction of main modules in Achilles

The process of Achilles is illustrated in Fig. 1, where three modules of Achilles are shown: a dictionary-

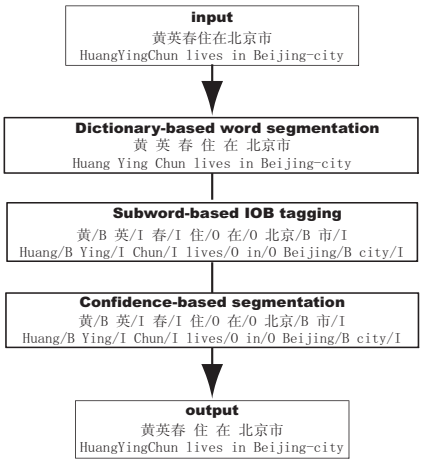

Figure 1: Outline of word segmentation process

based N-gram word segmentation for segmenting IV words, a subword-based tagging by the CRF for recognizing OOVs, and a confidence-dependent word segmentation used for merging the results of both the dictionary-based and the IOB tagging. An example exhibiting each step's results is also given in the figure.

The rule-based regular expression is not shown in the figure because this module interweaves with the other modules. This module can be called if needed at any time. The function of this module is to recognize numerical, temporal expression and others like product number, telephone number, credit number or alphabets. For example, "三万五千(35,000)", "八月(August)", "0774731301", "George Bush".

2.1 Dictionary-based N-gram word segmentation

Dictionary-based N-gram word segmentation is an important module for Achilles. This module can achieve a very high R-iv, but no OOV detection. We combined with it the N-gram language model (LM) to solve segmentation ambiguities. For a given Chinese character sequence, $C = c_0 c_1 c_2 \dots c_N$, the problem of word segmentation can be formalized as finding a word sequence, $W = w_{t_0} w_{t_1} w_{t_2} \dots w_{t_M}$, which satisfies

$$
\begin{aligned}
& w_{t_0} = c_0 \dots c_{t_0}, \quad w_{t_1} = c_{t_0+1} \dots c_{t_1} \\
& w_{t_i} = c_{t_{i-1}+1} \dots c_{t_i}, \quad w_{t_M} = c_{t_{M-1}+1} \dots c_{t_M} \\
& t_i > t_{i-1}, \quad 0 \le t_i \le N, \quad 0 \le i \le M
\end{aligned}
$$

such that

$$
\begin{aligned}
W &= \arg\max_W P(W|C) = \arg\max_W P(W)P(C|W) \\
&= \arg\max_W P(w_{t_0} w_{t_1} \dots w_{t_M}) \delta(c_0 \dots c_{t_0}, w_{t_0}) \\
&\quad \delta(c_{t_0+1} \dots c_{t_1}, w_{t_1}) \dots \delta(c_{t_{M-1}+1} \dots c_M, w_{t_M})
\end{aligned} \tag{1}
$$

We applied Bayes' law in the above derivation. Because the word sequence must keep consistent with the character sequence, $P(C|W)$ is expanded to be a multiplication of a Kronecker delta function series, $\delta(u, v)$, equal to 1 if both arguments are the same and 0 otherwise.

Equation 1 indicates the process of dictionary-based word segmentation. We looked up the lexicon to find all the IVs, and evaluated the word sequences with the LMs.

2.2 Subword-based IOB tagging using CRFs

If dictionary-based module recognizes IVs successfully, the subword-based IOB tagging can recognize OOVs. Before the subword-based tagging, the character-based "IOB" tagging approach has been widely used in Chinese word segmentation recently (Xue and Shen, 2003; Peng and McCallum, 2004; Tseng et al., 2005). Under the scheme, each character of a word is labeled as 'B' if it is the first character of a multiple-character word, or 'O' if the character functions as an independent word, or 'I' otherwise." For example, "全(whole) 北京市(Beijing city)" is labeled as "全(whole)/O 北(north)/B 京(capital)/I 市(city)/I".

We proposed the subword-based tagging (Zhang et al., 2006) to improve the existing character-based tagging. The subword-based IOB tagging assigns tags to a pre-defined lexicon subset consisting of the most frequent multiple-character words in addition to single Chinese characters. If only Chinese characters are used, the subword-based IOB tagging is downgraded into a character-based one. Taking the same example mentioned above, "全(whole) 北京市(Beijing city)" is labeled as "全(whole)/O 北京(Beijing)/B 市(city)/I" in the subword-based tagging, where "北京(Beijing)/B" is labeled as one unit.

We used the CRFs approach to train the IOB tagger (Lafferty et al., 2001) on the training data. We

downloaded and used the package "CRF++" from the site "http://www.chasen.org/taku/software." According to the CRFs, the probability of an IOB tag sequence, $T = t_0 t_1 \cdots t_M$, given the word sequence, $W = w_0 w_1 \cdots w_M$, is defined by

$$p(T|W) = \exp\left(\sum_{i=1}^{M}\left(\sum_k \lambda_k f_k(t_{i-1}, t_i, W) + \sum_k \mu_k g_k(t_i, W)\right)\right)/Z,$$
$$Z = \sum_{T=t_0 t_1 \cdots t_M} p(T|W) \quad (2)$$

where we call $f_k(t_{i-1}, t_i, W)$ bigram feature functions because the features trigger the previous observation t_{i-1} and current observation t_i simultaneously; $g_k(t_i, W)$, the unigram feature functions because they trigger only current observation t_i. λ_k and μ_k are the model parameters corresponding to feature functions f_k and g_k respectively.

The model parameters were trained by maximizing the log-likelihood of the training data using L-BFGS gradient descent optimization method. In order to overcome overfitting, a gaussian prior was imposed in the training.

The types of unigram features used in our experiments included the following types:

$w_0, w_{-1}, w_1, w_{-2}, w_2, w_0 w_{-1}, w_0 w_1, w_{-1} w_1,$
$w_{-2} w_{-1}, w_2 w_0$

where w stands for word. The subscripts are position indicators. 0 means the current word; $-1, -2$, the first or second word to the left; $1, 2$, the first or second word to the right.

For the bigram features, we only used the previous and the current observations, $t_{-1} t_0$.

As to feature selection, we simply used absolute counts for each feature in the training data. We defined a cutoff value for each feature type and selected the features with occurrence counts over the cutoff.

A forward-backward algorithm was used in the training and viterbi algorithm was used in the decoding.

2.3 Confidence-dependent word segmentation

Before moving to this step in Figure 1, we produced two segmentation results: the one by the dictionary-based approach and the one by the IOB tagging. However, neither was perfect. The dictionary-based segmentation produced results with higher R-ivs but lower R-oovs while the IOB tagging yielded the contrary results. In this section we introduce a confidence measure approach to combine the two results. We define a confidence measure, $CM(t_{iob}|w)$, to measure the confidence of the results produced by the IOB tagging by using the results from the dictionary-based segmentation. The confidence measure comes from two sources: IOB tagging and dictionary-based word segmentation. Its calculation is defined as:

$$CM(t_{iob}|w) = \alpha CM_{iob}(t_{iob}|w) + (1-\alpha)\delta(t_w, t_{iob})_{ng} \quad (3)$$

where t_{iob} is the word w's IOB tag assigned by the IOB tagging; t_w, a prior IOB tag determined by the results of the dictionary-based segmentation. After the dictionary-based word segmentation, the words are re-segmented into subwords by FMM before being fed to IOB tagging. Each subword is given a prior IOB tag, t_w. $CM_{iob}(t|w)$, a confidence probability derived in the process of IOB tagging, is defined as

$$CM_{iob}(t|w_i) = \frac{\sum_{T=t_0 t_1 \cdots t_M, t_i=t} P(T|W, w_i)}{\sum_{T=t_0 t_1 \cdots t_M} P(T|W)}$$

where the numerator is a sum of all the observation sequences with word w_i labeled as t.

$\delta(t_w, t_{iob})_{ng}$ denotes the contribution of the dictionary-based segmentation. It is a Kronecker delta function defined as

$$\delta(t_w, t_{iob})_{ng} = \begin{cases} 1 & \texttt{if } t_w = t_{iob} \\ 0 & \texttt{otherwise} \end{cases}$$

In Eq. 3, α is a weighting between the IOB tagging and the dictionary-based word segmentation. We found the value 0.7 for α, empirically.

By Eq. 3 the results of IOB tagging were re-evaluated. A confidence measure threshold, t, was defined for making a decision based on the value. If the value was lower than t, the IOB tag was rejected and the dictionary-based segmentation was used; otherwise, the IOB tagging segmentation was used. A new OOV was thus created. For the two extreme cases, $t = 0$ is the case of the IOB tagging while $t = 1$ is that of the dictionary-based approach. In a real application, a satisfactory tradeoff between

R-ivs and R-oovs could find through tuning the confidence threshold.

3 Part-of-speech Tagging

Our POS tagging is a traditional maximum entropy tagging (A.Ratnaparkhi, 1996) as follows,

$$p(t|h) = \frac{1}{Z(h)} exp(\sum_{i=1}^{M} \lambda_i f_i(h,t)) \quad (4)$$

where $Z(h)$ is a normalizing factor determined by requirement $\Sigma_t p(t|h) = 1$ over all t:

$$Z(h) = \sum_{t} exp(\sum_{i=1}^{M} \lambda_i f_i(h,t)) \quad (5)$$

In the evaluation, 17 categories of triggers were used, which include:

(w,t) , $(w_{-2}w_{-1}w,t)$, $(w_{-1}ww_1,t)$, (ww_1w_2,t) , $(w_{-1}w,t)$, (ww_1,t) , (t_{-1},t) , $(t_{-2}t_{-1},t)$, $(t_{-1}w_1,t)$, $(t_{-1}ww_1,t)$, $(w_{-1}w_1,t)$, (w_{-1},t) , (w_1,t) , $(t_{-1}w,t)$, $(t_{-2}t_{-1}w,t)$, $(w_{-2}w_{-1},t)$, (w_1w_2,t)

where:

w is the word whose tag we are predicting; t is the tag we are predicting; t_{-1} is the tag to the left of tag t; t_{-2} is the tag to the left of tag t_{-1}; w_{-1} is the word to the left of word w; w_{-2} is the word to the left of word w_{-1}; w_1 is the word to the right of word w; w_2 is the word to the right of word w_1 ;

In addition to the ME based POS tagging approach, we also combined a N-gram based POS tagging.

N-gram tagger is the most widely used tagger in part-of-speech tagging methods. The basic idea is to maximize a posterior probability $p(T|W)$ given a word sequence in order to find its tag sequence. By using Bayes rule, this can be transformed as to maximize $p(T) * p(W|T)$. Prior probability $p(T)$ is a N-gram language model of tag sequence. $p(W|T)$ is thought as an unigram model. In this experiment we used trigram to model $p(T)$.

Differing from the interpolation smoothing algorithm used in(Merialdo, 1994), both $p(T)$ and $p(W|T)$ were smoothed by back-off methods(Katz, 1987). Because a N-gram backoff model $P(T)$ is well-known, a backoff implementation of $p(W|T)$ was given here only. It is of the following equation.

	R	P	F	R-oov	R-iv
CKIP	0.938	0.931	0.935	0.640	0.966
CITYU	0.943	0.933	0.938	0.686	0.965
CTB	0.941	0.943	0.942	0.663	0.961
NCC	0.931	0.933	0.932	0.592	0.950
SXU	0.932	0.929	0.930	0.487	0.971

Table 1: Post evaluation of word segmentation.

$$p(w|t) = \begin{cases} \bar{p}(w|t) & \text{if } \bar{p}(w|t) \neq 0 \\ \beta(t)\bar{p}(w) & \text{otherwise} \end{cases} \quad (6)$$

where:

- $\bar{p}(w|t)$ and $\bar{p}(w)$ are discounting relative frequencies of $p(w|t)$ and $p(w)$, calculated by back-off discounting algorithm. The discount thresholds of $\bar{p}(w|t)$ and $\bar{p}(w)$ in present experiment were 12 and 1 respectively. A new word 'UNK' was added to the vocabulary, whose probability $\bar{p}(w)$ represents that of all the unseen words.
- $\beta(t)$ is a normalizing value to ensure $\Sigma_w p(w|t) = 1$.

4 Experiments

We participated all the closed evaluation of word segmentation and part-of-speech tagging. Our scores should have achieved better than the official numbers if we had submitted the results in the right format. Achilles outputs results in GBK/BIG5 format. However, the format determined by bakeoff organizers is Unicode-16. We made a lethal error when we converted the files from GBK/BIG5 to Unicode-16. Hence, the official results display wrong scores for our system's results.

We evaluated our results again in the post-evaluation. The results for word segmentation is shown in Table 1. The results for POS tagging is shown in Table 2.

Table 1 and Table 2 represent the real performance of Achilles in this evaluation. The official data do not.

5 Discussion

Achilles achieved good word segmentation results as shown in Table 1. Achilles was designed through

	Acc.	R-oov	R-iv
CKIP	0.913	0.530	0.946
CITYU	0.881	0.470	0.914
CTB	0.934	0.709	0.947
NCC	0.945	0.575	0.963
PKU	0.937	0.646	0.952

Table 2: Post evaluation of part-of-speech tagging.

three perspectives: IV recognition, OOV recognition and regular expression recognition. IV recognition can be solved at higher accuracy by dictionary-based approach. OOV recognition can be solved by IOB tagging. However, the flexible numerical and temporal expression cannot be solved by the above two methods. Hence, we used regular expression. Finally, the inconsistency of the above methods are resolved by confidence measure approach. These features causes higher performance achieved by Achilles.

6 Conclusions

This paper described systematically the main features of our Chinese morphological analyzer, Achilles. Because of its delicate design and state-of-the-art technological integration, Achilles achieved better or comparable segmentation results when it was compared with the world best segmenter.

You can get Achilles from the site "http://www.slc.atr.jp/~rzhang/Achilles.html".

References

A.Ratnaparkhi. 1996. A maximum entropy part-of-speech tagger. In *Proceedings of the Empirical Methods in Natural Language Processing Conference*.

Thomas Emerson. 2005. The second international chinese word segmentation bakeoff. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, Jeju, Korea.

Jianfeng Gao, Andi Wu, Mu Li, Chang-Ning Huang, Hongqiao Li, Xinsong Xia, and Haowei Qin. 2004. Adaptive chinese word segmentation. In *ACL-2004*, Barcelona, July.

S. Katz. 1987. Estimation of probabilities for sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics Speech and Signal Processing*, 35:400–401.

Taku Kudo and Yuji Matsumoto. 2001. Chunking with support vector machine. In *Proc. of NAACL-2001*, pages 192–199.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML-2001*, pages 591–598.

B. Merialdo. 1994. Tagging english text with a probabilistic model. *Computational Linguistics*, 20(2):155–172.

Fuchun Peng and Andrew McCallum. 2004. Chinese segmentation and new word detection using conditional random fields. In *Proc. of Coling-2004*, pages 562–568, Geneva, Switzerland.

Richard Sproat and Tom Emerson. 2003. The first international chinese word segmentation bakeoff. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, Sapporo, Japan, July.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 2005. A conditional random field word segmenter for Sighan bakeoff 2005. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, Jeju, Korea.

Nianwen Xue and Libin Shen. 2003. Chinese word segmentation as LMR tagging. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*.

Huaping Zhang, HongKui Yu, Deyi xiong, and Qun Liu. 2003. HHMM-based Chinese lexical analyzer ICTCLAS. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, pages 184–187.

Ruiqiang Zhang, Genichiro Kikui, and Eiichiro Sumita. 2006. Subword-based tagging by conditional random fields for chinese word segmentation. In *Proc. of HLT-NAACL*.