E-019

Achilles: A Chinese Morphological Analyzer

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. We tested Achilles using data from Sighan Bakeoff 2005. Achilles outperforms the best contester in CITYU, PKU and MSR corpora, achieving the highest F-scores.

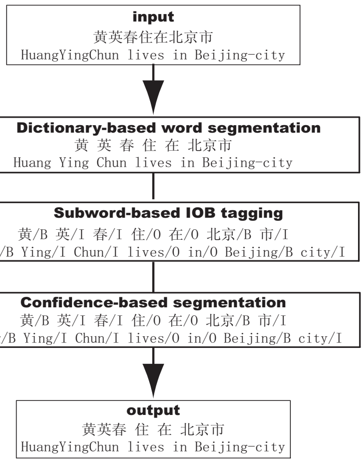
input
黄英春住在北京市
HuangYingChun lives in Beijing-city
Dictionary-based word segmentation
黄 英 春 住 在 北京市
Huang Ying Chun lives in Beijing-city
Subword-based IOB tagging
黄/B 英/I 春/I 住/O 在/O 北京/B 市/I
Huang/B Ying/I Chun/I lives/O in/O Beijing/B city/I
Confidence-based segmentation
黄/B 英/I 春/I 住/O 在/O 北京/B 市/I
Huang/B Ying/I Chun/I lives/O in/O Beijing/B city/I
output
黄英春 住 在 北京市
HuangYingChun lives in Beijing-city

Figure 1: Outline of word segmentation process

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 presents our experimental results. Section 4 describes current state-of-the-art methods for Chinese word segmentation, with which our results were compared. Section 5 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-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, $C = c_0c_1c_2 \dots c_N$, 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

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

such that

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

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/t̃aku/software." According to the CRFs, the probability of an IOB tag sequence, $T = t_0t_1 \cdots t_M$, given the word sequence, $W = w_0w_1 \cdots w_M$, is defined by

$$\begin{aligned} & p(T|W) = \\ & \exp\left(\sum_{i=1}^{M}\left(\sum_k \lambda_k f_k(t_{i-1}, t_i, W) + \sum_k \mu_k g_k(t_i, W)\right)\right)/Z, \qquad (2) \\ & Z = \sum_{T=t_0t_1\cdots t_M} p(T|W) \end{aligned}$$

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:

$$w_0, w_{-1}, w_1, w_{-2}, w_2, w_0w_{-1}, w_0w_1, w_{-1}w_1, w_{-2}w_{-1}, w_2w_0$$

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:

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

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

$$CM_{iob}(t|w_i) = \frac{\sum_{T=t_0t_1\cdots t_M, t_i=t} P(T|W, w_i)}{\sum_{T=t_0t_1\cdots t_M} P(T|W)}$$

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

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

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. In Section 3.2 we will present the experimental segmentation results of the confidence measure approach.

3 Experiments

We used the data provided by Sighan Bakeoff 2005 to test Achilles described in the previous sections. The data contain four corpora from different sources: Academia Sinica (AS), City University of Hong Kong (CITYU), Peking University (PKU) and Microsoft Research in Beijing (MSR). Since this work was to evaluate the proposed subword-based IOB tagging, we carried out the closed test only. Five metrics were used to evaluate segmentation results: recall(R), precision(P), F-score(F), OOV rate(R-oov) and IV rate(R-iv). For detailed info. of the corpora and these scores, refer to (Emerson, 2005).

For the dictionary-based approach, we extracted a word list from the training data as the vocabulary. Trigram LMs were generated using the SRI LM toolkit for disambiguation. Table 1 shows the performance of the dictionary-based segmentation. Since there were some single-character words present in the test data but not in the training data, the R-oov rates were not zero in this experiment. In fact, there were no OOV recognition. Hence, this approach produced lower F-scores. However, the R-ivs were very high.

3.1 Effects of the Character-based and the subword-based tagger

The main difference between the character-based and the word-based is the contents of the lexicon subset used for re-segmentation. For the character-based tagging, we used all the Chinese characters. For the subword-based tagging, we added another 2000 most frequent multiple-character words to the lexicons for tagging. The segmentation results of the dictionary-based were re-segmented using the FMM, and then labeled with "IOB" tags by the CRFs. The segmentation results using CRF tagging are shown in Table 2, where the upper numbers of each slot were produced by the character-based approach while the lower numbers were of the subword-based. We found that the proposed subword-based approaches were effective in CITYU and MSR corpora, raising the F-scores from 0.941 to 0.946 for CITYU corpus, 0.959 to 0.964 for MSR corpus. There were no F-score changes for AS and PKU corpora, but the recall rates were improved. Comparing Table 1 and 2, we found the CRF-modeled IOB tagging yielded better segmentation than the dictionary-based approach. However, the R-iv rates were getting worse in return for higher R-oov rates. We will tackle this problem by the confidence measure approach.

	R	P	F	R-oov	R-iv
AS	0.941	0.881	0.910	0.038	0.982
CITYU	0.928	0.851	0.888	0.164	0.989
PKU	0.948	0.912	0.930	0.408	0.981
MSR	0.968	0.927	0.947	0.048	0.993

Table 1: Our segmentation results by the dictionary-based approach for the closed test of Bakeoff 2005, very low R-oov rates due to no OOV recognition applied.

	R	P	F	R-oov	R-iv
AS	0.951 0.953	0.942 0.940	0.947 0.947	0.678 0.647	0.964 0.967
CITYU	0.939 0.950	0.943 0.942	0.941 0.946	0.700 0.736	0.958 0.967
PKU	0.940 0.943	0.950 0.946	0.945 0.945	0.783 0.754	0.949 0.955
MSR	0.957 0.965	0.960 0.963	0.959 0.964	0.710 0.716	0.964 0.972

Table 2: Segmentation results by a pure subword-based IOB tagging. The upper numbers are of the character-based and the lower ones, the subword-based.

3.2 Effects of the confidence measure

In section 2.3, we proposed a confidence measure approach to re-evaluate the results of IOB tagging by combinations of the results of the dictionary-based segmentation. The effect of the confidence measure is shown in Table 3, where we used $\alpha = 0.7$ and confidence threshold $t = 0.8$. In each slot, the numbers on the top were of the character-based approach while the numbers on the bottom were the subword-based. We found the results in Table 3 were better than those in Table 2 and Table 1, which prove that using confidence measure approach achieved the best performance over the dictionary-based segmentation and the IOB tagging approach. The act of confidence measure made a tradeoff between R-ivs and R-oovs, yielding higher R-oovs than Table 1 and higher R-ivs than Table 2.

Even with the use of confidence measure, the word-based IOB tagging still outperformed the character-based IOB tag-

	R	P	F	R-oov	R-iv
AS	0.953	0.944	0.948	0.607	0.969
	0.956	0.947	0.951	0.649	0.969
CITYU	0.943	0.948	0.946	0.682	0.964
	0.952	0.949	0.951	0.741	0.969
PKU	0.942	0.957	0.949	0.775	0.952
	0.947	0.955	0.951	0.748	0.959
MSR	0.960	0.966	0.963	0.674	0.967
	0.972	0.969	0.971	0.712	0.976

Table 3: Effects of combination using the confidence measure. The upper numbers and the lower numbers are of the character-based and the subword-based, respectively

	AS	CITYU	MSR	PKU
Bakeoff-best	0.952	0.943	0.964	0.950
Achilles	0.951	0.951	0.971	0.951

Table 4: Comparison our results with the best ones from Sighan Bakeoff 2005 in terms of F-score

ging. It proves the proposed word-based IOB tagging was very effective.

4 Discussion

Achilles achieved excellent word segmentation results as shown in Table 4 , where the results of Achilles are listed together with the best results from Bakeoff 2005 in terms of F-scores. Since it was a closed test, we locked the function of regular expression. We could yield a better results than those shown in Table 4 using the regular expression. Achilles beat the best competitor in CITYU, PKU and MSR corpora. All the technologies except the subword-based tagging used in Achilles have existed, however, Achilles integrated these techniques seamlessly.

Achilles was designed through 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.

5 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. The current functions of Achilles are available for word segmentation only. We will expand its function into part-of-speech and semantic tagging in the future work.

References

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.

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.

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*.