# 学習器の組み合わせによる中国語わかち書きシステム

**浅原 正幸**　　**福岡 健太**　　**東 藍**　　**續木 貴史**
**ゴー チュイリン**　　**渡邉 陽太郎**　　**松本 裕治**
奈良先端科学技術大学院大学 情報科学研究科　　松下電器産業株式会社

## Abstract

中国語のわかち書きにおける未知語 (辞書にない語) の問題を解決するために、未知語処理とわかち書きを同時に行う手法と未知語抽出とわかち書きを分割して行う手法の 2 つがある。本稿では後者の手法に基づき、Support Vector Machines (SVMs)、最大エントロピー法 (MaxEnt)、条件付確率場 (CRFs) の 3 つの学習器で未知語抽出器を構成する。それらの出力の多数決もしくは和集合を用いてテストデータ中に出現する未知語を予め抽出する。訓練データ中の単語と未知語抽出器が出力した単語とを用いて、条件付確率場 (CRFs) に基づくわかち書きシステムを構成し、評価実験を行った。

## 1　はじめに

本稿では中国語のわかち書きにおける未知語（辞書にない語）の問題を扱う。未知語の問題を解くための 1 つの手法は、単語単位のマルコフモデルのシステムに未知語モデルを埋め込む手法である。内元ら (Uchimoto et al., 2001) は、最大エントロピーマルコフモデルに基づく形態素解析器（わかち書きと同時に品詞タグづけを行う）を構成する際に、既知語（辞書にある語）だけではなく全ての 5 文字以下の単語を未知語候補としてラティス上に展開する手法を提案した。中川ら (Nakagawa, 2004) は、隠れマルコフモデルに基づく形態素解析器を構成する際に、既知語と文字単位に未知語境界を表現するノードをラティス上に展開する手法を提案した。

上記手法が単語単位のマルコフモデルを基にしているのに対して、既知語部分と未知語部分の両方を文字単位に処理する手法が提案されている。Xue ら (Xue and Converse, 2002) は最大エントロピー法によるタグづけ器と誤り駆動によるタグづけ器を用いて、文字単位に語境界を表現するタグづけする手法を提案した。また浅原らは (Asahara et al., 2003)、単純に文字単位に処理することにより単語間の関係が失われるのを避けるため、単語単位のマルコフモデルによるわかち書き器の冗長解析結果を基に文字単位に語境界を Support Vector Machines を用いてタグづけする手法を提案した。

Goh ら (Goh et al., 2004) は未知語処理と既知語処理を分割する手法を提案した。未知語抽出器の出力の中から人手による規則を用いて語として認定しにくいものを取り除いて候補語集合を構成し、その候補語集合と既知語集合を基に既知語処理ではわかち書きの曖昧性解消のみを行うように分割した。

本稿ではこの未知語処理と既知語処理を分割する枠組を基にして中国語わかち書き器を構成する。未知語抽出器の出力の選別を自動化するために 3 つの未知語抽出器を構成し、その多数決を用いる。既知語処理部分は、この単語集合を候補語として条件付確率場 (Lafferty et al., 2001) を用いたわかち書き器 (Kudo and Matsumoto, 2004) により行う。開発したシステムは、2005 年の SIGHAN Workshop で行われた 2nd International Chinese Word Segmentation Bakeoff (以下 “Bakeoff”) (Emerson, 2005) に参加し、他のシステムとの比較を行った。その結果についても報告する。

## 2　提案手法

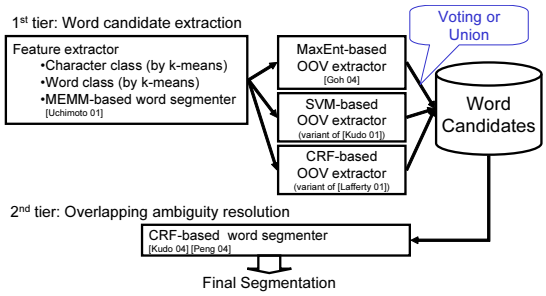


Figure 1: 提案手法の全体像

提案する手法は未知語処理とわかち書きの曖昧性解消処理の 2 つの部分からなる (Figure 1)。未知語処理部分として、3 つの未知語抽出器を構成した。1 つ目と 2 つ目の抽出器は、わかち書きとチャンキングの 2 段階の解析を行う。まず最大エントロピーマルコフモデル (MEMMs) を用いたわかち書きを行う。このわかち書き結果を素性として、未知語出現箇所を文字単位にチャンキングすることにより未知語抽出器を行う。1 つ目の抽出器ではサポートベクトルマシン (SVMs) を、2 つ目の抽出器では条件付確率場 (CRFs) を用いてチャンカーを構成した。3 つ目の抽出器は、既発表 (Goh et al., 2004) の手法で、文字単位に最大エントロピー法の分類器を用いて語境界を推定し、未知語と既知語 (訓練データもしくは辞書にある語) を同時に抽出する。これらの未知語抽出器を用いてテストデータを解析して生成される 3 つの出力を基に多数決 (voting) もしくは和集合 (union) により 2 種類の未知語集合を得る。

未知語集合と訓練データ中の語を合わせて単語集合を作り、この単語集合を辞書として条件付確率場を用いてわかち書きの曖昧性を解消する。

上記学習器は周囲の単語や文字の出現を文脈素性として用いる。これらの素性は細かく、全く同じ単語や文字の出現がない場合がある。予めクラスタリング (K-means 法) を行い、単語や文字にクラスを付与し、これらを素性として用いることにより、データスパースネスの問題に対応する。単語のクラスは、学習器 (MEMMs や CRFs) の隠れ状態としても用いる。

### 2.1 文字・単語のクラスタリング

訓練データ中の全ての単語と全ての文字について、K-means アルゴリズムを用いて、ハードクラスタリングを行う。クラスタリングを行うために R 2.2.1 (http://www.r-project.org/) を用いる。

前後の単語を素性として用いて、単語を 20 クラスに分類する。さらに、“未知語” (未知語抽出器によって得られた語)、文頭、文末の 3 クラスを別途定義し、計 23 クラスを定義する。

前後の文字とその文字を含む単語中の相対位置を BIES タグ[1]で表現したものを素性として用いて、文字を 20 クラスに分類する。さらに、テストデータにしか出現しない文字 1 クラスを別途定義し、計 21 クラスを定義する。

### 2.2 未知語抽出器

構成した 3 種類の未知語抽出器を順に説明する。

1 つ目と 2 つ目の未知語抽出器は、未知語出現箇所を系列タギングで推定することによる。仮想的に未知語を生成するために、訓練データを 80 % と 20 % に分割し、前者に出現する単語のみを、わかち書き器を構成するための候補語として用いる。後者にのみ出現する単語は未知語とみなし、この出現箇所を学習することにより未知語抽出器を構成する。

最初にわかち書きを行うモデルとして、最大エントロピーマルコフモデル (MEMMs) (McCallum et al., 2000) に基づくわかち書き器 (Uchimoto et al., 2001) の出力を用いる。わかち書き器の内部では、1 文中の全ての候補語をラティス上に展開される。各単語は前節で付与された単語クラスを持ち、それをラティス上の隠れ状態として持つ。訓練時には、先行する単語クラスを前状態素性として、現在の単語の最初の文字とそのクラスおよび最後の文字とそのクラスを観測素性として用い、状態遷移確率を最大エントロピー法で推定する。解析の際には、求められた状態遷移確率に基づく Viterbi アルゴリズムによる。

このわかち書き器による出力を基にして、文字単位に BIO タグ[2]を付与することにより、未知語の出現箇所を同定する。1 つ目と 2 つ目の未知語抽出器の違いは、この未知語出現箇所同定に用いる学習器にある。前者はサポートベクトルマシン (SVMs) に基づく手法 (Kudo and Matsumoto, 2001) を用い、後者は条件付確率場 (CRFs) に基づく手法 (Lafferty et al., 2001) (Peng and McCallum, 2004) を用いた。各学習器に与える素性として、前後 2 文字ずつ含む 5 文字窓幅にある文字と文字クラスを用いる。MEMMs によるわかち書き器によって生成された単語境界は、BIES タグにより文字単位の情報に分割し、単語と BIES タグの 2 つ組を素性として用いる。また、単語クラスと BIES タグの 2 つ組についても素性として用いる。尚、SVMs に基づく手法は、工藤らの YamCha (http://chasen.org/~taku/software/yamcha/) を用いた。

3 つ目の未知語抽出器は、決定的に最大エントロピー法により BIES タグを付与していく (Goh et al., 2004) ことによる。

以降の節では、各未知語抽出器をそれぞれ “MEMMs+SVMs”、 “MEMMs+CRFs”、 “MaxEnt” と略記する。

未知語抽出器から 3 つの未知語集合を得る。最後のわかち書き器に与える候補語集合を作成するために、多数決と和集合の 2 種類の手法を用いる。1 つ目の手法は、2 つ以上の未知語抽出器が認定した語のみを候補語に追加する手法で、以降 “voting” 法と略記する。2 つ目の手法は、各未知語抽出器が認定した語全てを候補語に追加する手法で、以降 “union” 法と略記する。

### 2.3 条件付確率場に基づくわかち書き

最終的なわかち書きは、条件付確率場 (CRFs) に基づく手法 (Kudo and Matsumoto, 2004) による。未知語抽出器からの候補語集合と訓練データ中の単語集合を用いて、候補語集合を構成する。MEMMs に基づく手法と同様に、1 文中の単語候補全てをラティス上に展開する。MEMMs と異なるのは、状態遷移確率を最大エントロピー法を用いて各点で求めるのではなく、条件付確率場を用いて文全体で確率値を正規化することにより求める。これにより、他の単語単位のマルコフモデルに基づく手法に起きうる、より長い単語を選択しやすいという、長さバイアス問題 (Kudo and Matsumoto, 2004) を回避することができる。条件付確率場に与える素性として、単語、単語クラス、接頭文字、接尾文字とそれらの文字クラスなどを用いた。

## 3 評価: SIGHAN Bakeoff 2005

本システムは 2005 年に開かれた SIGHAN Workshop の Bakeoff (中国語のわかち書きコンテスト) に参加した。その結果を報告する。SIGHAN Workshop は、2002 年より開かれている、中国語の言語処理を研究対象とする ACL (Association for Computational Linguistics) の分科会である。2003 年の会議で第 1 回目の Bakeoff (Sproat and Emerson, 2003) が開かれ、2005 年に 2 回目 (Emerson, 2005) が開かれた。まず訓練データ（わかち書き済みテキスト）が参加者に配布され、約 2 週間の後テストデー

[1]“B” は単語の先頭文字、“I” は単語内の文字で先頭でも末尾でもない文字、“E” は単語の末尾文字、“S” は 1 文字で 1 単語を表現する文字。

[2]“B” は未知語の先頭文字、“I” は未知語内の文字で先頭でない文字、“O” は既知語内の文字。

Table 1: SIGHAN Bakeoff 2005 における結果

| | AS | CITYU | MSR | PKU |
|---|---|---|---|---|
| CRFs + Voted Unk. | 0.947/0.606/0.971 (2/11) | 0.942/0.629/0.967 (2/15) | 0.971/0.570/0.988 (1/29) | 0.934/0.521/0.955 (10/23) |
| CRFs + Union Unk. | 0.939/0.445/0.967 (7/11) | 0.928/0.598/0.940 (8/15) | 0.966/0.571/0.974 (2/29) | 0.917/0.325/0.940 (14/23) |
| Char.-based tagging (Goh et al., 2004) | 0.952/0.696/0.963 (1/11) | 0.941/0.736/0.953 (3/15) | 0.958/0.718/0.958 (6/29) | 0.941/0.760/0.941 (7/23) |

値は F-値/未知語再現率/既知語再現率 (F 値による順位/参加団システム数) を表す。下線部は実際のトラックでは間に合わなかった参考値。

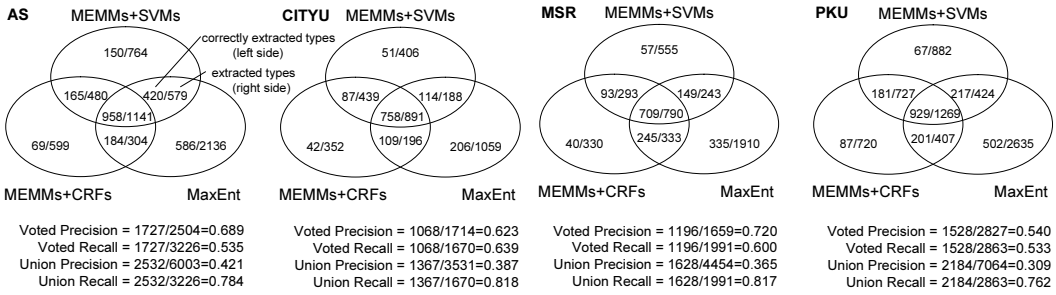


Figure 2: 各未知語抽出器の出力

タ（わかち書きされていないテキスト）が配布され、参加者がそのデータをわかち書きし 48 時間以内に開催者にわかち書き済みテキストを返す。開催者は全ての参加者の再現率、精度、F-値、未知語の再現率、既知語の再現率を公開する。4 組織 – Academia Sinica (AS)、香港城市大 (CITYU)、Microsoft Research(MSR)、北京大 (PKU)–からデータが提供され、参加者は自分が提供したデータ以外のトラックに参加することが可能である。外部データの利用の有無で、さらにトラックが分割される。Open Track では、配布された以外の辞書、コーパス、文字に関する情報を自由に用いて良い。Closed Track では、配布されたデータのみを使ってシステムを作成する。提案手法は、Closed Track の規則にのっとり、純粋にコーパスに基づく手法を用い、学習器に与える素性として、アラビア数字・漢数字の情報、人名に利用されやすい文字情報、品詞情報、その他人手による規則などは用いていない。

我々は、4 組織のデータ全てについての Closed Track に参加した。1 団体で複数のシステムを参加させて良いため、前節で述べた提案手法の未知語抽出器の結果を多数決 (voting) したものと和集合 (union) をとったものを参加させた。さらに、以前提案したシステムである Goh らのシステム (Goh et al., 2004) を我々の開発時のベースラインとして参加させた。Table 1 に、結果 (F 値、未知語の再現率、既知語の再現率) と順位を示す。表の各数字の意味を Table 1 の下部に示す。残念ながら実際のトラックでは、時間的制約のために下線が引かれている MSR データについて本稿の提案手法によるシステムの解析結果を返すことができなかった。Table 1 の数値は、締切後評価実験を行い、他のシステムと比較した結果である[3]。

3 つのデータ (AS, CITYU, MSR) については良い結果を収めたが、PKU データでは良い結果が得られなかった。結果と、未知語率、未知文字（テストデータにしか現われない文字）率、平均単語長などとの関連を調査したが、特別な相関は得られなかった。

Figure 2 に、各未知語抽出器が出力した候補語の異なりの包含関係を示す。実際にテストデータに出現した候補語を正解として評価する。未知語抽出器間で出力を比較すると “MaxEnt” が一番出力数が多く、“MEMMs+SVMs” と “MEMMs+CRFs” のほぼ 2 倍あった。未知語抽出の時点では、多数決 (voting) をとったモデルの方が精度がよく、和集合 (union) をとったモデルの方が再現率がよくなる。当初、未知語抽出において和集合をとったモデルの方が、多数決をとったモデルに比べて、最終的なわかち書きの結果の未知語の再現率も高いだろうと予測していた。しかし、結果は逆であった。和集合をとったモデルの場合、未知語抽出の時点での精度が 30 〜 42% しかなく、この誤りが、条件付確率場に基づくわかち書き器に対して悪影響を及ぼし、未知語、

[3]他団体のシステムの結果は、文献 (Emerson, 2005) もしくは http://www.sighan.org/bakeoff2005/results.php を参照すること。Web page 中の NAIST a が提案手法 voting, NAIST b が Goh らのシステム、NAIST c が提案手法 union である。間に合わなかった MSR データは、NAIST a システムとして最後のわかち書き推定に CRFs ではなく MEMM を用いたものを、NAIST c システムとして最後のわかち書き推定に未知語を全く加えなかった CRFs を用いたものの結果である。

既知語両方の再現率を下げる結果となった。

Closed Track で良い結果を収めた他団体のシステムについて言及する。Tseng ら (Tseng et al., 2005)(AS 3 位、CITYU 1 位、MSR 1 位、PKU 2 位) は、文字単位の条件付確率場による Peng らの手法 (Peng and McCallum, 2004) を基にしている。主な工夫として、畳語（同じ語を重ねて形成される語）に対応した素性の追加があげられる。Chen ら (Chen et al., 2005)(AS 4 位、CITYU 4 位、MSR 3 位、PKU 1 位) は、文字単位の最大エントロピーマルコフモデルによる Xue ら (Xue and Converse, 2002) の手法を基にしている。主な工夫は、予め訓練データに対して、人海戦術で人名、地名、数値表現をタグづけし、そのタグを学習する点にあるが、厳密には Closed Track の規則を違反している。Fu ら (Fu et al., 2005)(AS 6 位、CITYU 5 位、MSR 2 位、PKU 6 位) は、既知語部分に対し、単語単位の隠れマルコフモデルを作成したあと、その出力を用いて既知語が正しいか未知語境界があるかを文字単位に再タグづけする手法を用いている。

## 4 おわりに

今回は SIGHAN Bakeoff のスケジュールおよび規則により、複数の未知語処理器を組み合わせて予め候補語集合を選別し、単語単位の条件付確率場を用いて最終的なわかち書きを取る手法を用いた。実用に有効な未知語抽出手法の他の手法として、Peng ら (Peng and McCallum, 2004) の手法がある。文字単位の条件付確率場を作成し、切り出されるトークンの周辺確率を用いて、そのトークンの尤度を順序づけし、ある尤度以上のもののみを候補語として利用する手法である。この手法の利点として、辞書を編纂する際には、人手で尤度順に従って選別作業を行う点があげられる。

現在は、単語単位のマルコフモデルのシステムに未知語モデルを埋め込む手法について研究を行っている。内元ら (Uchimoto et al., 2001) が提案した、既知語と未知語候補（全ての部分文字列）同時に展開されたラティス上で、最大エントロピーマルコフモデルではなく、条件付確率場により解析を行う手法について検討している。

## References


Masayuki Asahara, Chooi-Ling Goh, Xiaojie Wang, and Yuji Matsumoto. 2003. Combining Segmenter and Chunker for Chinese Word Segmentation. In *Proc. of Second SIGHAN Workshop*, pages 144–147.

Aitao Chen, Yiping Zhou, Anne Zhang, and Gordon Sun. 2005. Unigram Language Model for Chinese Word Segmentation. In *Proc. of Fourth SIGHAN Workshop*, pages 138–141.

Thomas Emerson. 2005. The Second International Chinese Word Segmentation Bakeoff. In *Proc. of Fourth SIGHAN Workshop*, pages 123–134.

Guohong Fu, Kang-Kwong Luke, and Percy Ping-Wai Wong. 2005. Description of the HKU Chinese Word Segmentation System for Sighan Bakeoff 2005. In *Proc. of Fourth SIGHAN Workshop*, pages 165–168.

Chooi-Ling Goh, Masayuki Asahara, and Yuji Matsumoto. 2004. Pruning False Unknown Words to Improve Chinese Word Segmentation. In *Proc. of PACLIC-18*, pages 139–149.

Taku Kudo and Yuji Matsumoto. 2001. Chunking with Support Vector Machines. In *Proc. of NAACL-2001*, pages 192–199.

Taku Kudo and Yuji Matsumoto. 2004. Applying Conditional Random Fields to Japanese Morphological Analysis. In *Proc. of EMNLP-2004*, pages 230–237.

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 282–289.

Andrew McCallum, Dayne Freitag, and Fernando Pereira. 2000. Maximum Entropy Markov Models for Information Extraction and Segmentation. In *Proc. of ICML-2000*, pages 591–598.

Tetsuji Nakagawa. 2004. Chinese and Japanese Word Segmentation Using Word-Level and Character-Level Information. In *Proc. of COLING-2004*, pages 466–472.

Fuchun Peng and Andrew McCallum. 2004. Chinese Segmentation and New Word Detection using Conditional Random Fields. In *Proc. of COLING-2004*, pages 562–568.

Richard Sproat and Thomas Emerson. 2003. The First International Chinese Word Segmentation Bakeoff. In *Proc. of Second SIGHAN Workshop*, pages 133–143.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 2005. A Conditional Random Field Word Segmenter. In *Proc. of Fourth SIGHAN Workshop*, pages 168–171.

Kiyotaka Uchimoto, Satoshi Sekine, and Hitoshi Isahara. 2001. The Unknown Word Problem: a Morphological Analysis of Japanese Using Maximum Entropy Aided by a Dictionary. In *Proc. of EMNLP-2001*, pages 91–99.

Nianwen Xue and Susan P. Converse. 2002. Combining Classifiers for Chinese Word Segmentation. In *Proc. of First SIGHAN Workshop*, pages 63–70.