中国版茶筌の開発

ゴーチュイリン　鄭育昌　浅原正幸　松本裕治
奈良先端科学技術大学院大学 情報科学研究科
{ling-g,yuchan-c,masayu-a,matsu}@is.naist.jp

1　はじめに

制限なく利用可能な中国語形態素解析システムは依然として少ない。本稿では、日本語用の形態素解析器である「茶筌」を基にして、大規模な中国語用の辞書を整備することにより、利用制限の少ない中国語形態素解析器を構成するプロジェクトの概略について述べる。まず、Penn Chinese Treebank（CTB）[8] のわかち書きと品詞に基づき、基本辞書 28,390 語（単語のみ）を構築する。語彙を増やすため、大量の生テキストデータから、未知語抽出器を用いて、未知語を抽出する。抽出された未知語は、人手によりチェックを行い、新しい語彙として辞書に登録する。構成された形態素解析器の評価実験において、わかち書きで 91.9 の F-値、品詞タグ付けで 85.7 の F-値が得られた。また、辞書を整備する際に用いた未知語抽出器の評価実験において、70.3% の再現率と 55.7% の精度、未知語に限定しての品詞タグ付けで 63.8% の精度が得られた。

2　システムのアーキテクチャ

「茶筌」[10] の日本語モデルと同様に中国語の形態素解析問題を Hidden Markov Model （HMM）を用いて解析する。HMM は辞書に登録されている語を認定する際、短い単位より長い単位を認定する傾向 (length-bias problem) [6] がある。しかし、長い単位を辞書に列挙することは困難である。そこで図 1 の左側に示すような多段の解析モデルを採用する。まずあらかじめ最小単位を定義し、次に HMM を用いて最小単位を認定し、最後にチャンカーを用いて、HMM の最小単位の出力を目標の単位にまとめあげる。

解析モデル作成に必要な単位の定義と訓練データの構成方法を図 1 の右側に示す。最終出力として認定したい単位は CTB コーパス [8] による。HMM が出力する最小単位は CTB コーパスを基に独自に定義する (2.1 節)。HMM の訓練には最小単位タグつきコーパスと最小単位辞書を必要とする。最小単位タグつきコーパスは、定義を基に CTB コーパスを変換して作成する。最小単位辞書は、最小単位タグつきコーパスから既知語彙を選別するとともに、未知語抽出器と品詞推定器を構成し大量の生テキスト [7] から獲得した語彙も登録する。チャンカーのモデルは、最小単位を入力とし CTB コーパスで定義される単位を出力とするように、最小単位タグつきコーパスと CTB コーパスの 2 つを訓練データとして構成する。

現在公開されている「茶筌」を中国語対応させるために、文字コード毎のトークナイザーモジュール、未知語処理モジュールを改変した。

2.1 節で、今回定義した最小単位について述べる。2.2 節で、最小単位から CTB コーパスの単位にまとめあげるためのチャンカーについて述べる。2.3 節で、語彙を増やすために導入した未知語抽出器と品詞推定器について述べる。

2.1　最小単位の定義

最小単位は独自に定義したものである。元の CTB では固有名詞は一つの品詞（NR）でまとめている。それは固有名詞の範囲が広過ぎるため、品詞細分類を定義し、新たに 7 種類の品詞を導入する。まず、人名を名字（NR-PER-FAM）、名前（NR-PER-GIV）、外国人名（NR-PER-FOR）と他の人名（NR-PER-OTH）の 4 種類の品詞に分類する。次に、地名（NR-LOC）、組織名（NR-ORG）と他の一般固有名詞（NR-OTH）の 3 種類を定義する。この定義に基づき、人手でコーパスの単語と品詞を変換した。

数値表現の線形結合を解析するため、全ての数字の組み合わせを辞書に登録することは困難である。提案する解析モデルでは、CTB に出現する数値表

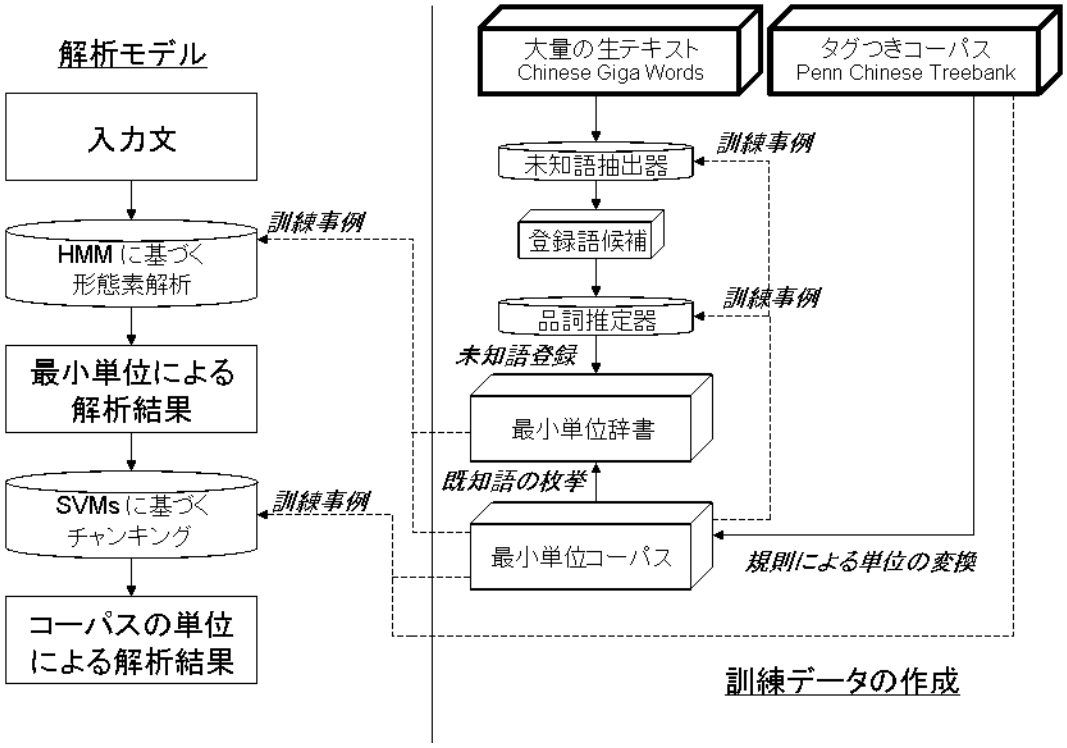

図 1: 中国語形態素解析器の作成

現を我々が定義する最小単位に分割し、数値表現の線形結合をチャンカーに行う。数値表現を構成する品詞は 3 種類ある。時間名詞（NT：三月）、基数詞（CD：十多）と序数詞（OD：第九）である。これらの 3 種類の品詞をさらに細分類する。一般の数字（アラビア数字、漢数字）は、全て品詞 CD-NOR を付与する。その他の時間名詞と序数詞には、それぞれ NT-NOR と OD-NOR を付与する。さらに助数詞に対して、新しい品詞（NT-AFF、CD-AFF、OD-AFF）を導入する。先に示した例は、三/CD-NOR 月/NT-AFF、十/CD-NOR 多/CD-AFF、第/OD-AFF 九/CD-NOR のように分割される。

アルファベットを含む単語についても、数値表現と同様により短い単位を導入する。アルファベットを含む単語を文字単位に分割し、辞書にはａ, Aからｚ, Ｚまでの単語のみを品詞 FW として登録する。実際には、アルファベットを含む単語の品詞は様々であり、新しい単位を導入することで元の品詞情報は失なわれてしまう。例えば、”Ｐ／Ｅ値 ”（NN）を最小単位に変換すると、”Ｐ/FW ／/PU Ｅ/FW 値/NN ”になる。現時点では長い単位として ”Ｐ／Ｅ値 ”を辞書に登録しないが、もしアルファベットを含む長い単位がよく利用されるのであれば長い単位を辞書に登録することも可能である。

以上の最小単位定義に基づいてコーパスを変換する。元のコーパス約 42 万語を変換した結果 45 万語になった。品詞数は 34 個から 41 個に増加した。最小単位のコーパスから全部の単語を抽出して、最小単位の基本辞書 33,438 語（単語と品詞のペア）を構築した。

構成された最小単位のコーパスと基本辞書（訓練データ中にのみ出現する単語による）を用いて、単純 bi-gram HMM のモデルを構成した。コーパスを訓練データ 90%/ テストデータ 10%に分割して評価実験を行なった。テストデータの中には 4.2%の未知語が存在し、わかち書きで 91.9 の F-値、品詞タグ付けで 85.7 の F-値が得られた。

2.2 チャンカーによる CTB 単位同定

HMM から出力される最小単位をチャンカーを用いて CTB の元の単位に復元する。チャンカーは Support Vector Machines に基づく Yamcha [5] を用いる。素性として、HMM から出力される単語の前後 2 単語とその品詞である。実験結果では、HMM から出力された結果を入力とした場合の CTB 単位

のわかち書きで 90.5 の F-値、品詞タグ付けで 84.0 の F-値が得られた。

2.3 未知語抽出とその品詞推定

最小単位のコーパスから基本辞書を構築したが、このままでは語彙が少ない。語彙が少ないと、実際の場面では未知語に遭遇することが多く、現在のモデルは未知語モデルを組み込んでいないため解析結果が悪くなってしまう。そこで語彙を増やすため、大量の生テキストデータから、未知語抽出器を用いて未知語を抽出する。抽出された未知語は人手によりチェックを行い、新しい語彙として辞書に登録する。

未知語抽出器は [2] に示されるモデルを使う。Maximum Entropy Models を用いて文字単位に position tagging を行ない、単語のわかち書きを行なう。素性としては前後 2 文字と文字タイプ（NUMber、ALPhabet、SYMbol、HANzi）を用いる。わかち書きの出力中で、基本辞書に登録されていない単語を新規に登録する候補とする。

抽出された未知語を辞書に登録するためには、未知語の品詞を推定する必要がある。品詞推定も Maximum Entropy Models を用いた [3]。コーパス中で頻度 1 の単語を未知語として、訓練データを構成した。素性は前後 2 単語とその品詞である。テストの際には、テストデータを形態素解析器に入力して、未知語の前後文脈の品詞を展開し、これを素性として用いた。

CTB から構成された最小単位コーパスにおける未知語抽出器の評価実験を行なった。訓練データ 90%、テストデータ 10% の実験設定で、70.3%の再現率と 55.7%の精度、未知語に限定しての品詞タグ付けで 63.8%の精度が得られた。

語彙を増やすために、実際に大量の生テキストから未知語を抽出する。生テキストとして、LDC が配布している Chinese Gigawords (1118380K 文字) [7] (CGW) を用いた。以下、CGW における未知語抽出精度について述べる。

最小単位コーパスにより訓練した未知語抽出器を用いることにより、CGW コーパスの一部[1]である生テキスト（45836 文字/625 文[2]）から 2258 語の未知語が抽出された。抽出される未知語の精度を評価するために、抽出された未知語を人手で次に示す 3 つに分類した：

- 抽出単位と品詞がともに正しい 40.7%(921/2258)
- 抽出単位は正しいが品詞が誤り 34.1%(772/2258)
- 抽出単位が誤り 24.6%（566/2258）

CTB コーパスにおける実験とほぼ同等の精度が得られた。品詞を無視して単位のみを判定すると CGW コーパスにおける精度は 74.9% であり、CTB コーパスにおける抽出精度 (55.7%) より良かった。なお、人手による判定基準は、抽出された単位が出現した文脈では正しくない単位であっても、辞書に単語として登録すべきものであれば正解とした。例えば、出現する文脈中で正しいわかち書きは ”手榴弾/NN ” だが、抽出された単語が ”手/NN、榴弾/NN ”であった場合、”榴弾/NN ”も登録すべき語彙として正解とした。「抽出単位は正しいが品詞が誤り」の約 50%は固有名詞を一般名詞とする誤りであった。特に、組織名（NR-ORG）、外国人名（NR-PER-FOR）、名字（NR-PER-FAM）、名前（NR-PER-GIV）においてこのような誤りが多かった。

3 関連研究

現在手に入る中国語形態素解析器はいくつかある。清華大学 [9] は CSeg&Tag1.0 というシステムを開発しており、60,133 語の見出し語を持つ。報告されている精度はわかち書きで 98.0%〜99.3%で、品詞付けで 91.0%〜97.1%である。北京大学 [4] は SLex1.1 を公開しており約 7 万語の見出し語を持つ。報告されている精度はわかち書きで 97.05%、品詞付けで 96.42%であった。市販されている形態素解析システムとして Basic Technology が構築した Chinese Morphological Analyzer（CMA）[1] がある。約 120 万語の見出し語を持っている。しかし、精度については報告していない。これらのシステムに比べると、現在の我々のシステムの語彙数は依然として規模が小さい。しかし今後、生テキストから抽出された語を人手でチェックして語彙を増やし、継続的に辞書のメンテナンスを行なうことにより、これら関連研究に匹敵するような形態素解析器を構成したいと考えている。

[1] xin200209 の先頭から 625 文。
[2] 最小単位コーパスに基づいた解析器により 29505 語に分割された。

表 1: 評価実験結果

	わかち書き			品詞付け		
	再現率	適合率	F-値	再現率	適合率	F-値
HMM による最小単位解析	94.2%	89.8%	91.9	87.8%	83.7%	85.7
チャンキング	93.0%	88.2%	90.5	86.3%	81.8%	84.0

	抽出			品詞推定
	再現率	適合率	F-値	正解率
未知語抽出	70.3%	55.7%	62.2	63.8%

4 おわりに

本稿では、現在奈良先端大で行なっている利用制限の少ない中国語形態素解析システムを構築するプロジェクトの概要について述べた。提案手法では、まず HMM によりあらかじめ定義した最小単位を同定し、次にチャンカーにより CTB コーパスの単位に戻すという、2 段階の解析モデルを導入した。また、今回採用した CTB コーパスは、利用制限は小さいが、コーパスの規模が小さく、そこから生成される基本辞書の規模も小さい。そこで、CTB コーパスで学習した未知語抽出器を用い、大量の生テキストから未知語を抽出し、人手でチェックして新たな語彙を獲得する方法を提案した。各要素技術の精度を表 1 にまとめる。現在の登録語彙数は依然少ないが、今後引き続き作業を続けていくことにより、大規模な辞書を構成していきたいと考えている。

参考文献

[1] Thomas Emerson. 2001. Segmenting Chinese Text. *MultiLingual Computing & Technology*, pages 43–??

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

[3] Chooi-Ling Goh. 2003. Chinese Unknown Word Identification by Combining Statistical Models. Master's thesis, Nara Institute of Science and Technology, Japan.

[4] Institute of Computational Linguistics, Peking University. 2001. Chinese Text Segmentation and POS Tagging. http://www.icl.pku.edu.cn/nlp-tools/segtagtest.htm.

[5] Taku Kudo and Yuji Matsumoto. 2001. Chunking with Support Vector Machines. In *Proceedings of NAACL*, pages 192–199.

[6] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying Conditional Random Fields to Japanese Morphological Analysis. In *Proceedings of EMNLP*, pages 230–237.

[7] Linguistic Data Consortium. 2003. Chinese Gigaword. http://www.ldc.upenn.edu/.

[8] Linguistic Data Consortium. 2004. The Penn Chinese Treebank v4.0. http://www.cis.upenn.edu/~chinese/.

[9] Maosung Sun, Dayang Shen, and Changning Huang. 1997. CSeg&Tag1.0: A Practical Word Segmentation and POS Tagger for Chinese Texts. In *fifth Conference on Applied Natural Language Processing*, pages 119–126.

[10] 松本ら, 2002. **形態素解析システム「茶筌」**. 奈良先端科学技術大学院大学. http://chasen.aist-nara.ac.jp/.