E-027

情報量基準に基づく形態素解析用辞書の自動獲得方式

Knowledge Acquisition for Morphological Lexicons based on Information Criteria

柳原 正† 池田 和史† 松本 一則† 滝嶋 康弘†
Tadashi Yanagihara Kazushi Ikeda Kazunori Matsumoto Yasuhiro Takishima

概要

一般的に用いられる形態素解析器では、単語境界や品詞の推定を行う際に形態素解析用辞書が必要である。新しい単語に対応するため、それらの単語を辞書に追加しなければならないが、この作業は人手によって行わなければならないため、コストがかかることが大きな問題となっている。本論文では、情報量基準に基づく形態素解析用辞書の自動獲得方式を提案する。提案内容では、情報量基準に基づくモデル検定によって、単語境界及び品詞を自動推定する。これにより、人手を借りずに形態素解析用辞書を自動的に更新可能となる。

1. 背景

形態素解析器では、形態素解析を行うために単語境界と品詞を推定するための機能を提供する。一般的に用いられるものとして、ChaSen[1]、MeCab[2]などが挙げられる。これらの形態素解析器では、形態素解析を行う際に単語のコストや品詞に関する項目が記された形態素解析用辞書をコーパスとして用いる。

これらの方式による形態素解析を行う場合、推定対象となる文字列が辞書内の単語で該当するものが存在しない場合、それらの文字列の前後の品詞などの特徴に基づいて品詞推定が行えるが、この方法による推定は辞書に登録されている単語と比較したとき、単語境界および品詞推定の精度が悪い。一方、これらの文字列を未知語として分類し、人手により未知語に対して適切な単語境界と品詞を付与し、辞書に登録することで、未知語と判定された文字列に対して正しい単語境界および品詞が判定されるようになる。しかし、この方法では人手による登録コストが高いことが欠点として挙げられる。

2. 提案方式

本論文では、形態素解析器によって未知語と判定される文字列に対し、形態素解析器から独立した機構によって、単語境界および品詞のタグ付けを行う形態素解析用辞書の自動獲得方式を提案する。本方式にもとづくシステム構成図を以下の図 1 に示す。動作手順は以下のとおりである。

1. 形態素解析器から品詞つきの単語とは別で得られる未知語を入力データとする。
2. 未知語に対し、単語境界推定機能によって、未知語内の単語境界が推定され、品詞なし単語が得られる。
3. 品詞なし単語に対し、品詞推定機能により、品詞が付与され、品詞あり単語が得られる。
4. 品詞あり単語は形態素解析用辞書の仕様に基づいて変換され、辞書へ登録される。

† 株式会社 KDDI 研究所, KDDI R&D Laboratories Inc.

本提案により、既存の形態素解析器に対し、従来人手によって行われていた辞書の項目を自動獲得することが可能になり、形態素解析用辞書の保守で必要となったコストを削減できる。また、一般的に用いられる形態素解析器を使用したシステムにおいて、形態素解析器を新たに置き換える必要がなく、本手法を実装した機構を付け足す形で導入できることが利点として挙げられる。

なお、形態素解析用辞書の知識獲得を実現するために必須となる単語境界の推定および品詞の推定の具体的な実現方法として、単語境界の推定を 3 章、品詞の推定を 4 章で詳細を述べる。

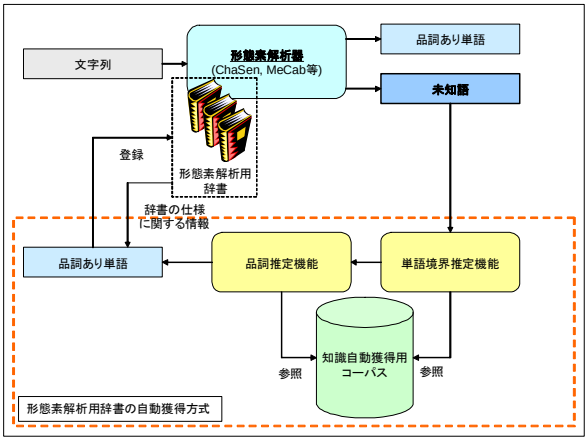

図 1 形態素解析用辞書の自動獲得方式に関するシステム構成図

3. 単語境界の推定方式

単語境界の推定方式として、文献[3]に掲載された方式を用いる。文献[3]は赤池情報量基準[4]に基づくモデル検定[5]を用いて、文字間における単語境界の存在の有無を判定する方式である。なお、本手法を採用する理由としては、以下の通りである。

- 情報量基準に基づくモデル検定では、他手法では必須である外的パラメータが不要である。
- 単語境界を推定するためのコーパスは文字の出現頻度をもとにしており、計算量とコーパスは小さく抑えられる。

本提案では、文献[3]に掲載された手法をもとに、汎用性能が高いSupport Vector Machine (以降、SVM)上で実装した。SVM上で実装する際の手順は以下のとおりである。

3.1. 単語境界推定用コーパスの構築

文献[3]では、任意の文字または文字列を対象に、その前後に他の文字や文字列が出現したとき、対象となる文字や文字列と前後に出現した文字や文字列と関連性があるもの

であれば単語境界が存在せず、関連性のあるものであれば単語境界が存在する、という仮定に基づいて、単語境界を推定する。このためには、文の先頭からn番目に存在する文字l_nや、文字l_nを先頭とする文字列l_nに対し、文字l_{n+1}や文字列s_{n+1}が後続したとき、各種変数であるa, b, c, d, AIC(IM), AIC(DM), Eを含む特徴量が必要となる。以下に、ケース別において、変数として割り当てるべき値を掲載する。

n 文字目の文字 l_n に対し、文字 l_{n+1} が直後に出現した場合

- a = N_{11}:n 文字目が文字 l_n であるとき、$n+1$ 文字目に文字 l_{n+1} が直後に出現した事例数
- b = N_{12}:n 文字目が文字 l_n であるとき、$n+1$ 文字目に文字 l_{n+1} が直後に出現しない事例数
- c = N_{21}:n 文字目が文字 l_n でないが、$n+1$ 文字目に文字 l_{n+1} が直後に出現した事例数
- d = N_{22}: n 文字目が文字 l_n でない上、$n+1$ 文字目に 文字 l_{n+1} が直後に出現しない事例数
- AIC(IM): n 文字目に文字 l_n が出現する現象と $n+1$ 文字目に 文字 l_{n+1} が出現する現象において、独立関係であると仮定したときの値
- AIC(DM): n 文字目に文字 l_n が出現する現象と $n+1$ 文字目に 文字 l_{n+1} が出現する現象において、従属関係であると仮定したときの値
- E: AIC(IM)および AIC(DM)の差分から求められる値

n 文字目の文字 l_n に対し、文字列 s_{n+1} が直後に出現した場合

- a = N_{11}:n 文字目が文字 l_n であるとき、$n+1$ 文字目に文字列 s_{n+1} が直後に出現した事例数
- b = N_{12}:n 文字目が文字 l_n であるとき、$n+1$ 文字目に文字 s_{n+1} が直後に出現しない事例数
- c = N_{21}:n 文字目が文字 l_n でないが、$n+1$ 文字目に文字列 s_{n+1} が直後に出現した事例数
- d = N_{22}: n 文字目が文字 l_n でない上、$n+1$ 文字目に 文字列 s_{n+1} が直後に出現しない事例数
- AIC(IM): n 文字目に文字 l_n が出現する現象と $n+1$ 文字目に 文字列 s_{n+1} が出現する現象において、独立関係であると仮定したときの値
- AIC(DM): n 文字目に文字 l_n が出現する現象と $n+1$ 文字目に 文字列 s_{n+1} が出現する現象において、従属関係であると仮定したときの値
- E: AIC(IM)および AIC(DM)の差分から求められる値

n 文字目の文字列 s_n に対し、文字 l_{n+1} が直後に出現した場合

- a = N_{11}: n 文字目に文字列 s_n であるとき、$n+1$ 文字目に文字 l_{n+1} が出現した事例数
- b = N_{12}: n 文字目に文字列 s_n であるとき、$n+1$ 文字目に文字 l_{n+1} が出現した事例数
- c = N_{21}: n 文字目に文字列 s_n でないが、$n+1$ 文字目に文字 l_{n+1} が出現した事例数
- d = N_{21}: n 文字目に文字列 s_n でない上、$n+1$ 文字目に文字 l_{n+1} が出現した事例数
- AIC(IM): n 文字目に文字列 s_n が出現する現象と $n+1$ 文字目に 文字 l_{n+1} が出現する現象において、独立関係であると仮定したときの値
- AIC(DM): n 文字目に文字列 s_n が出現する現象と $n+1$ 文字目に 文字 l_{n+1} が出現する現象において、従属関係であると仮定したときの値
- E: AIC(IM)および AIC(DM)から求められる差分

n 文字目の文字列 s_n に対し、文字列 s_{n+1} が直後に出現した場合

- a = N_{11}:n 文字目が文字列 s_n であるとき、$n+1$ 文字目に文字列 s_{n+1} が直後に出現した事例数
- b = N_{12}:n 文字目が文字列 s_n であるとき、$n+1$ 文字目に文字 s_{n+1} が直後に出現しない事例数
- c = N_{21}:n 文字目が文字列 s_n でないが、$n+1$ 文字目に文字列 s_{n+1} が直後に出現した事例数
- d = N_{22}: n 文字目が文字列 s_n でない上、$n+1$ 文字目に文字列 s_{n+1} が直後に出現しない事例数
- AIC(IM): n 文字目に文字列 s_n が出現する現象と $n+1$ 文字目に 文字列 s_{n+1} が出現する現象において、独立関係であると仮定したときの値
- AIC(DM): n 文字目に文字列 s_n で出現する現象と $n+1$ 文字目に 文字列 s_{n+1} が出現する現象において、従属関係であると仮定したときの値
- E: AIC(IM)および AIC(DM)の差分から求められる値

上記の例では文字 l_n や文字列 s_n に対し、文字や文字列が直後に存在する例であるが、文字 l_n や文字列 s_n の直前に存在する文字や文字列でも同様の処理を行う。

なお、このままでは文字列 s が長いものも計算することとなるため、文字列 s の長さの上限である文字数 L を定義して制約として加える。

最後に特徴量に対し、以下の変数も含める。

- 文字 l_n と文字 l_{n+1} の間にある区間における単語境界の有無

本手法のもとである文献[3]では教師なしデータによる学習を行っていたが、二値分類を解く場合におけるSVMの性質を踏まえ、事前にタグ付与を行った。

3.2. 単語境界の推定

上記で記載したとおり、文全体に対し、一括で単語境界を推定することが困難であるため、L 文字からなる文字列(以降、ウィンドウ)で単語境界を観測し、n+1 文字ずつずらしながら単語境界の推定を行う方式を取る。

具体的には説明は以下の図 2 で示す。図 2 では 6 文字からなる文字列を対象としており、ウィンドウの大きさは 3 文字である、とする。このとき、このウィンドウは図 2 のように、3 回ずらしながら、全文内における単語境界を推定する。

はじめに、ウィンドウが Win_1 の位置に存在したとき、l_1 と l_2、l_2 と l_3 の間に単語境界が存在すべきかについて推定する。次に、ウィンドウが Win_2 の位置に存在したとき、l_2 と l_3、l_3 と l_4 の間に単語境界が存在すべきかについて推定する。Win_3 や Win_4 でも同様のことを行う。

単語境界の判定結果を計算したあと、単語境界を推定した全結果に対し、同一の文字間における単語境界の推定を行った結果ごとにまとめる。例えば、図 2 の例では以下のようになる。

- l_1 と l_2 の間：Win_1
- l_2 と l_3 の間：Win_1、Win_2
- l_3 と l_4 の間：Win_2、Win_3
- l_4 と l_5 の間：Win_3、Win_4
- l_5 と l_6 の間：Win_4

このとき、各文字間の判定結果において、判定結果が一つのみである場合と複数ある場合が存在する。

一つのみであった場合、判定結果をそのまま反映する。

複数であった場合、判定結果がすべて単語境界の存在を示している場合に限り、単語境界が存在すると判定する。

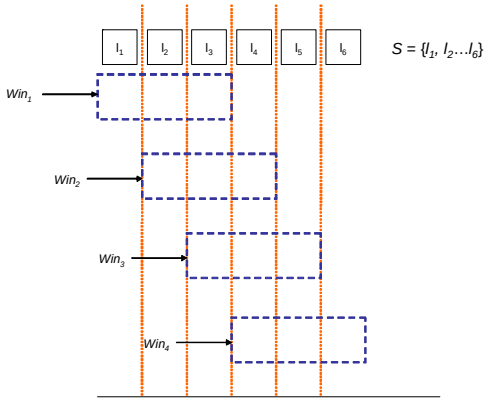

図 2 単語境界の判定に関する動作手順図

4. 品詞の推定方式

品詞の推定方式として、文献[6]の手法を採用する。文献[6]の手法は汎用性能が高いSVMを用いて品詞の推定を行っており、品詞の推定に対し、英語を対象とした場合では95%以上の精度を示している。品詞の推定のみを行う手法であるため、日本語に適用するためには、文書の分かち書きへの対応が必要となる。本提案では、文献[3]が提供する単語境界の推定機能によって、本課題を解決する。このため、以下の説明はすでに文書から単語境界が得られたことを前提とし、動作手順を説明する。

4.1. 品詞推定用コーパスの構築

品詞を推定するためのコーパスが必要となるが、基本的には単語境界の推定時に用いたコーパスに対し、情報をもとに拡張したものを使用する。具体的には、以下の情報を含めたトークンを生成し、コーパスとして用いる。

- 未知語の前後 3 つの品詞
- 未知語の前後 3 つの単語
- 未知語の前後 4 文字以下の語頭と語尾
- 未知語が数字・大文字・ハイフンの有無

4.2. 品詞の推定

品詞の推定として、文献[6]では前方の品詞のみを使用する方法と前後の品詞を使用する方法の 2 種類が掲載されている。これらのうち、精度がより良い後者の手法である、前後の品詞を使用する手法を採用する。

5. 評価方針

本方式の有効性を示すため、インターネットから取得したブログ記事を実験用データとして用いる。これらのデータを基に精度を計測する。

評価実験で利用するSVMとして、LIBSVM[7]を採用する。LIBSVMのSVMタイプとして標準のC-SVC、カーネル関数としては文献[6]と同じく、polynomial kernel(d = 2)を使用する。

学習データとして、実験用データから生成したコーパスから、1,000、10,000、100,000 トークンより構成されるデータセットから選択する。これらのトークンのデータセットを SVM に対して学習データとして入力する。一方、コーパスの全データのうち、学習データの対象に含まれないデータの残りのデータを評価データとして用いる。

まとめ

本論文では、形態素解析器における形態素解析用辞書の知識獲得手法について述べた。情報量基準にもとづく単語境界推定および品詞の推定によって、本手法が実現できることを提案した。今後はこの方式の有効性を評価し、形態素解析用辞書に対する影響を検証する。

参考文献

[1] ChaSen: http://chasen.naist.jp/
[2] MeCab: Yet Another Part-of-Speech and Morphological Analyzer. http://mecab.sourceforge.net/
[3] 柳原 正, 池田 和史, 松本 一則, 滝嶋 康弘, “情報量基準を用いた単語境界推定方式”, 第 190 回 情報処理学会自然言語処理研究会, 2009.
[4] H. Akaike. “Information Theory and an Extension of the Maximum Likelihood Principle.” In *2nd International Symposium on Information Theory*, 1973.
[5] K. Matsumoto and K. Hashimoto, "Schema Design for Causal Law Mining from Incomplete Database", In *Proceedings of Discovery Science, Second International Conference (DS '99)*, 1999.
[6] 中川 哲治, 工藤 拓, 松本 裕治, “Support Vector Machine を用いた未知語の品詞推定”, 第 141 回 情報処理学会自然言語処理研究会, 2001.
[7] LIBSVM – A Library for Support Vector Machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm/