信頼性を考慮した自律学習型単語境界推定方式の提案

柳原　正†　池田　和史†　松本　一則†　滝嶋　康弘†

株式会社ＫＤＤＩ研究所†

〒356-8502 埼玉県ふじみ野市大原 2-1-15

1. はじめに

一般的に形態素解析器[1)2)]は，文字列から単語境界と品詞を推定するための機能を提供する．これらのうち，単語境界は，一般的に文字列間における境界を表すラベルが付与されたコーパスを学習した上で，入力される文字列に対し，推定する．文字列の境界は与えられた辞書などのコーパスをもとに推測を行うため，コーパスが対応できない表現が出現した場合，推測精度が低下する．

このような表現については，境界を表すラベルをコーパスに登録することによって対応できる．しかし，この作業は人によって行われるため，登録対象となる表現が増大すると対応が困難となる．そこで，機械によって付与されたラベルを使用し再学習を行う半教師あり学習を用いた構成が考えられる．しかし，半教師あり学習を実行する際，機械によって付与されたラベルには精度が高いものと低いものが含まれており，これらを分離した上で再学習を行わなければならない．本論文では，このような自動付与された判定結果に対する信頼性を考慮した自律学習型の単語境界推定方式を提案する．

2. 従来手法

本論文で前提とする単語境界推定として，文字単位のn-gram統計情報をもととする単語境界推定方式[3)]とする．従来の辞書登録に基づいた形態素解析器とは異なり，単語の辞書への登録作業を不要とする代わりに，文字の連接に対する尺度のみからなるコーパスで動作するため，学習データが生成しやすいためである．また，文献3)では赤池情報量基準[4)]に基づくモデル検定から求めたスコア[5)]をもとに単語間の連接の強さを表すスコアを計算しており，確率に基づいた単語間の連接を表現した場合と比べ，単語境界推定の精度が高いことが確認されている[3)]．

本手法を用いて単語境界の推定を行うためには，以下の動作手順を行う．

1. ラベルなしデータに対し，文字列に隣接する特定の文字列およびその他の文字列との出現回数をもとに 2x2 分割表を用いて，文字列間の連接の強さを表すスコアを計算する．
2. 判定対象となる文字列に対し，文字列内の文字間におけるスコアを付与する．
3. 文字列間の文字列間の連節の強さを表すスコアとその前後の文字列間におけるスコアを比較する．特定の文字列とその隣接する文字列間のスコアにおいて相対的にスコアが下がっている場合では単語境界が存在する確率が高いと推定する．

3. 提案手法

本論文では，n-gram 統計情報に基づく単語境界推定方式によって推定された判定結果を用いて，学習データと併せた上で再帰的な学習を行う半教師あり学習の確立を目標としている．一般的な半教師あり学習を実施する場合では，主に以下の手順に基づく．

1. 人手などにより，学習データであるラベルありデータを生成する．これを教師データとする．
2. 学習データによって学習を行った判定器を用いて，ラベルなしデータに対して付与すべきラベルを判定し，付与する．
3. 付与されたラベルつきデータの中から一部のデータを選択し，1.の学習データと併せた上で再学習を行う．

この手順に従って付与されたラベルには，誤りが含まれているものと含まれていないものが混在することが考えられる．再学習を行う際に判定結果に誤りがあるものを使用した場合，精度が却って低下してしまうことを踏まえ，判定結果に誤りが含まれている可能性が低いデータのみを選定した上で再学習を行わなければならない．このため，機械によって付与されたラベルにおける信頼性を判定するための尺度を与えることが望ましい．

そこで，本論文では「特定の文字列内の区切りにおいて単語境界が存在することが確定したとき，同一の文字列を含む他事例においても同様の境界は発生しやすい」という仮説を立て，単語境界の判定結果が与えられた際に，他事例と比較し，他

"Confidence Based Active Learning for Word Segmentation"
Tadashi Yanagihara†, Kazushi Ikeda†, Kazunori Matsumoto†, Yasuhiro Takishima†
† KDDI R&D Laboratories Inc.

判定対象：
　マジ|で
事例：
　1: |マジ|で|ヤバ|い|　○
　2: |マジで|ヤバイ|　×
　3: |マジ|で|ヤバ|そう|　○
　…
マジ|で：100件中、90回出現→p_1: 0.9

図1. 判定対象および事例との比較

事例においても同一の区切りに単語境界の存在の有無をもとに信頼性を判定する方式を提案する．

このときの信頼性の基準としては，「単語境界が存在した事例」「単語境界が存在しなかった事例」のそれぞれの事例数を確率として表現する．例えば，図 1.における例では「マジ」と「で」の間に単語境界が存在すると判定された場合に，n-gram 統計情報内の同一の文字列を含む事例を抽出し，「同一の文字列内の同一の区切りにおいて，単語境界が存在した確率」および「「同一の文字列内の同一の区切りにおいて，単語境界が存在しなかった確率」を求める．(図 1.中では 100 件中に 90 件単語境界が存在したことを想定している)前者を p_1，後者は$(1-p_1)$として表すことができる．

次に，これらの確率をもとに判定結果の信頼性を表す尺度に変換する．本論文ではConfidence Based Active Learning[6]と呼ばれる手法に着目している．文献6)の特徴としては，判定結果の信頼性の判定としてエントロピを尺度としている．これは，半教師あり学習において，学習した際に得られる情報量が多いものを優先的に抽出できるためである．上記で述べたp_1および$(1-p_1)$の値に対する不確定性(Uncertainty)をエントロピで計算するためには以下の公式を用いる．これによって与えられた単語境界推定結果に対する不確定性を表すIが求まる．

不確定性 I をもとに，与えられた閾値 t と比較を

$$I = -p_1 \log p_1 - (1-p_1)\log(1-p_1)$$

行い，閾値 t を満たすものに対し，信頼性があると判定する．これにより，判定結果のうち，機械によって付与されたラベルにおいて信頼性があるデータのみが抽出でき，抽出されたデータを学習時に用いた教師データと混在したのち，再学習を行う．この過程を繰り返すことにより，精度が向上できることが期待できる．

4. 評価方針

本提案の評価方法としては，ウェブから取得したデータに対し，人手によってラベル付与を行ったものを用いる．本提案は半教師あり学習であるため，3 種類のデータが必要となる：開始時に利用する教師データ，再帰的な学習を行う際に用いる強化学習用データ，精度評価を計測するために判定する評価用データ，である．3 種類のデータの生成方法として，n-fold cross validation を用いてランダムにデータを n 等分し，生成された n 等分のデータを 3 種類のデータにランダムに仕分ける．このとき，強化学習用データおよび評価用データについては，すでに人手によりラベルが付与されていた場合はラベルを取り除く．これらのデータ実験を行う手順は以下の通りである．

1. 教師データを単語境界推定方式に入力し，学習を行う．
2. 単語境界推定方式を用いて，評価用データに対し，単語境界を推定したのち，精度を計測する．
3. 単語境界推定方式を用いて，強化学習用データに対し，単語境界を推定したのち，信頼性が高いもののみを選定する．
4. 3.で選定対象となったデータを 1.の学習データと併せた上で再学習を行う．2.と同様に評価用データに対し，単語境界を推定したのちに精度を再計測する．2.の精度および4.の精度を比較する．

5. おわりに

本論文では，単語境界推定方式における半教師あり学習を実現するため，エントロピによって表される不確定性を信頼性の尺度として取り組んだ方式を提案した．また，本手法の有効性を検証する際の評価方針について延べた．今後は本手法を実際のデータを使用して評価を行い，本手法を適用した場合における判定結果の精度への影響について調査していく予定である．

謝辞

本研究は独立行政法人情報通信研究機構(NiCT)より受託された「インターネット上の違法・有害情報の検出技術の研究開発」の研究支援によって行われた．

参考文献

1) MeCab: Yet Another Part-of-Speech and Morphological Analyzer. http://mecab.sourceforget.net
2) 日本語形態素解析システム JUMAN: http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html
3) 柳原，松本，池田，滝嶋．情報量基準に基づいた単語境界推定方式の提案．情報処理学会研究報告 2009-NL-190，pp. 43-48, 2009.
4) 鈴木義一郎, 情報量基準による統計解析入門, 講談社サイエンティフィック, 1995
5) Kazunori Matsumoto and Kazuo Hashimoto, "Schema Design for Causal Law Mining from Incomplete Database, Discovery Science, Second International Conference, 1999, Lecture Notes in Computer Science 1721 Springer, pp. 92-102
6) Li, Mingkun and Sethi. "Confidence-Based Active Learning", IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 28, Num. 8, pp. 1251-1261. 2006.