文脈類似度と表記類似度を用いた教師あり同義語抽出

森本 康嗣　　柳井 孝介　　岩山 真
yasutsugu.morimoto.zf@hitachi.com　kohsuke.yanai.cs@hitachi.com　makoto.iwayama.nw@hitachi.com
(株)日立製作所 中央研究所

1. はじめに

同義語辞書は，文書における表現の揺れを吸収するための言語資源であり，文書検索での質問拡張等を代表に多くの応用先がある。特に，コールセンタでのFAQ検索や特許の公知例検索等の業務向けの検索では，再現率が高い，すなわち検索漏れが少ないことが望まれる。そのため，同義語辞書のような検索漏れを防止するための言語資源に対するニーズが大きい。

同義語辞書は，価値が大きいデータであることから，人手によって多くの辞書が編纂されている。一方で，同義語辞書を自動的に作成しようとする試みが古くから行われている。本報告では，人手作成の同義語書を教師データとして用いることにより，教師あり学習技術を適用した同義語抽出に関して検討した結果について述べる。提案方式では，教師あり同義語抽出の枠組みによって，類似度のパラメータを適切にチューニングすることと，単語の出現文脈の類似性と表記の類似性という全く異なる尺度を統合することで高い精度を実現できる。

2. 従来研究と課題

2.1 従来研究

上位・下位語，兄弟語などを含む，広義の同義語を抽出する方式には数多くのアプローチが存在する。以下では，代表的なものについて２つの観点から整理する。

(1) 特徴抽出

どのような情報を特徴として用いるかに関しては多くの研究が行われており，最近ではWeb上の資源を活用する研究が盛んであるが[1][2][3]，以下では代表的な方式である単語の出現文脈を用いる方式，および単語の表記を用いる方式について説明する。

(a) 文脈ベース方式

単語の出現文脈を単語の「意味」だと考え，出現文脈間の類似度により，単語間の意味的類似度を定義する[4]。出現文脈の定義，文脈間の類似度の定義による数多くのバリエーションが存在し得る。対象となる同義語のタイプを選ばないのが長所である。相澤らは，データを大規模化した際の類似度計算方式，特にノイズの除去方式について提案している[5]。王らは，出現文脈の定義が抽出精度にどのように影響するかを詳細に検討し，また同義語関係が対称な関係であることを利用したランキング方式を提案している[6]。

(b) 表記ベース方式

代表的なものとしては，「コンピュータ」と「コンピューター」のような表記揺れ(異表記語)の抽出がある[7]。いわゆる同義語ではないが，"computer"と「コンピュータ」のような翻字(transliteration)の抽出も発音に関する情報を用いる点で同種のものである[8]。また，「回路」と「電子回路」のような単語レベルでの包含関係にある上位・下位語を抽出する試みも存在する[9]。

(2) 学習方式

(a) 教師なし学習

従来研究の多くは，同義語抽出を教師なし学習の問題として扱う。上で述べた特徴に基づく類似度を定義し，類似度順にランキングを行い，ある閾値以上の語を同義語として抽出する。

(b) 教師あり学習

同義語抽出を教師あり学習の問題として定式化する研究がいくつか知られている。渡部らは，既存の同義語辞書とWebの検索エンジンを使用して，類義語を抽出する方法を提案している[10]。

萩原[11]は，文脈ベース方式の同義語抽出を教師あり学習によって行う方式を提案している。人手作成の同義語辞書を用いてラベル付けを行うことにより教師データを作成し，類似度を示すスコア関数を直接学習することで非常に高精度な同義語抽出を実現している。

2.2 課題と本研究のアプローチ

教師あり学習による同義語抽出は，1)既存の同義語辞書を教師データとして用いることで高い精度が実現可能であり，2)異なる種類の特徴を自然に統合できるという優れた特徴を持ち，本報告でも教師あり学習のアプローチを採用する。本報告と最も類似した先行研究は萩原によるものであり，異なる点は以下の通りである。

萩原が，類似度に関する前提知識を用いないのに対し，本研究では教師なし学習によるアプローチにおいて利用されている類似度の計算式のチューニングを教師あり学習によって行うというアプローチを採る。これは，スコア関数そのものを学習するには，教師データの量，スパースネスが問題になることが予想されるためである。

また，文脈類似度と表記類似度という全く異なる種類の類似度を教師あり学習によって統合することで，同義語抽出を行う点が異なる。

近年，文書検索において注目されている“Learning

to Rank”[12] では，従来知られている各種のスコア関数を素性として用い，人手によって作成された正解付きの検索結果リストによってチューニングする。本研究は，同義語抽出を，ある入力単語に対して類似した単語を検索する問題だと解釈して，“Learning to Rank”の枠組みを同義語抽出に適用するものである。

3. 提案方式

3.1 教師あり学習による同義語抽出

本研究では，2.2 節で述べたアプローチに従い，同義語抽出を教師あり学習の問題として定式化する。

(1) 単語ペアのベクトル表現

2 個の単語ペア p=<q,b>を考える。q は見出しに相当する単語であり，以下クエリと呼ぶ。b は q に対する同義語候補であり，以下ターゲットと呼ぶ。両者をあわせて処理対象単語と呼ぶ。この単語ペア p を素性ベクトルで表現するためには，各種アプローチによる類似度を素性であると考えるのが自然である。i 番目の類似度計算方式を i 番目の素性であると考えると，i 番目の類似度計算方式によるスコア s_i を用いて，$(s_1, s_2, s_3,\cdots,s_N)$という素性ベクトルが得られる。この素性ベクトルが，単語ペア p の素性ベクトル表現となる。

(2) ラベルの付与

ベクトル表現された単語ペアに対し，同義語辞書を用いてラベルを付与する。ラベルが付与された教師データが得られることによって，同義語抽出を識別問題として解くことができる。本報告では，同義語抽出を２値分類問題として扱う，即ち単語ペア p が同義語であれば正例(+1)，同義語でなければ負例(-1)であると考える。

(3) ラベル付与における負例の扱い

ラベル付与の際，基本的には，同義語辞書に含まれる単語ペアは同義語，そうでない単語ペアは同義語ではないと考えてラベルを付与する。ただし，以下の点を考慮すべきである。同義語辞書中に記載がある単語ペアについては，その記載に従ってラベルを付与することに問題はない。しかしながら，記載がない単語ペアに対しては注意が必要である。辞書に含まれていない単語ペアが同義語である可能性が常に存在するためである。この問題について，本研究では以下のように取り扱う。

同義語辞書は，単語のペアの集合だと考えることができ，この集合を SP とする。また，SP 中に含まれている単語全体を W とする。このとき，ある単語ペアが同義語辞書 SP に含まれていない場合に，以下の２種類を区別する。

(a) 個々の単語は両方ともWに含まれる場合

(b) 個々の単語の少なくとも一方がWに含まれない場合

前者に関しては，同義語ではないものとして扱い，後者については，同義語であるかないかが不明であると考える。

3.2 素性

単語ペアに関する素性としては，従来研究で提案されている各種のアプローチに基づいて定めた類似度を用いることができる。本報告では，汎用性が高く，テキストの大規模化によって高い精度が期待できる文脈類似度を中心に，表記類似度を併用した場合について検討する。

(1) **文脈類似度**

文脈類似度は，文脈の定義(近傍単語の抽出方法)と類似度の定義によって様々なバリエーションがあり得るが，本報告では以下のように設定した。

(a) 文脈の定義

文脈としては，以下の定義が考えられる。個々の定義による文脈を別の素性にすることも考えられるが，予備評価の結果効果が見られなかったため，下記の定義にしたがって抽出した結果をマージしたものを文脈として用いた。

・動詞句

単語 t に対し，「t を開発する」，「t がダウンする」のような，「名詞(句)＋助詞＋述語」のような動詞句のパターンを抽出し，「助詞＋述語」を文脈単語として用いる。

・「の」を含む名詞句

単語 a,b に対し，「a の b」のようなパターンを抽出し，「a」に対して「の b」を，「b」に対しては「a の」を文脈と単語して用いる。

・修飾語を含む名詞句

単語 t に対し，形容詞や形容動詞に対し，「新しい t」，「高速な t」のようなパターンを抽出し，t に対する文脈単語として，「新しい」，「高速な」を用いる。

(b) 類似度の定義

クエリ q とターゲット b の類似度に関しては，連想検索エンジン MANTA[13]で用いられている以下の式を用いた。

$$s(b \mid q) = \frac{1}{L + \kappa * [dlen(b) - L]} * \frac{1}{n} \sum_i w(t_i \mid S) * \nu(t_i \mid b)$$

ただし，

$$w(t_i \mid S) = \log\left(1 + \frac{\#D}{df(t)}\right) * \frac{1}{\#S} \sum_{d \in S} \nu(t \mid d)$$

$$\nu(t \mid d) = \frac{1 + \log(tf(t \mid d))}{1 + \log(tf(\bullet \mid d))}$$

であり，t はクエリ q の文脈単語，$\#D$ は処理対象単語数，$df(t)$は文脈単語 t を含む処理対象単語数，$tf(t|d)$は処理対象単語 d に関する文脈単語 t の出現回数，$tf(\bullet|d)$は処理対象単語 d に現れる文脈単語の出現回数の平均，$dlen(b)$はターゲット b の異なり語数，L は処理対象単語毎の文脈単語数の平均値，κ は定

数である。

上の式は，素性ベクトル間の内積による類似度を，文書長の分布により補正した形となっている。文書長による補正パラメータ κ は，MANTA のデフォルトでは 0.4 と定められているが，これを最適化することを考える。具体的には，κ を適宜変化させ，それぞれの値を用いて計算した類似度を素性として使用する。

(2) 表記類似度

表記ベース方式では，重複して出現する文字が同義語判定の手がかりとなる。例えば，「コンピュータ」と「コンピューター」のようなカタカナ語の異表記では，多くの文字が重複している。漢字で構成される単語についても，「分析」と「解析」，「気温」と「温度」のように同じ文字を含む同義語は多い。よって，単語ペアに対し，重複して出現する文字数に基づいた類似度を設定し，素性として用いる。具体的には，以下の通りである。

(a) 重複文字による表記類似度

基本的な考え方としては，2 つの単語について，一致している文字数を取得し，それぞれの単語の文字数の大きい方で正規化したものを類似度とする。ただし，後述する文字の重み付け，文字の類似度を考慮した補正を行う。本報告では，DP マッチングのような文字の順序や位置を反映させた処理は行わなかった。

(b) 文字の重み付けによる文字数の補正

「ー」のような頻出する文字の場合は，同義語ではないにもかかわらず，偶然合致する可能性が高い。そのため，文字に出現し易さに応じて重み付けを行う。文書検索における単語の IDF (Inversed Document Frequency)と同様の考え方で，文字が出現する単語数と全単語数から文字の重みを定義し，文字数を重みで補正して表記類似度を計算する。

(c) 文字の類似度

漢字は表意文字であり，意味が類似した文字が存在する。同義語辞書を用いることで，文字の類似度を計算し，表記類似度の計算に用いる。具体的には，同義語ペアに含まれる任意の文字のペアの頻度をカウントし，それぞれの文字の出現頻度で正規化したものを文字の類似度とする。

4. 評価

4.1 評価の方法

人手作成の同義語辞書とテキストデータから教師データを作成し，５分割の交差検定法によって評価を行った。

(1) 同義語辞書

NICT から公開されている日本語 WordNet[14]を２値識別タスクのための同義語辞書として用いた。日本語 WordNet から，同義語である単語ペアを抽出し，全ての語彙を抽出して評価対象の語彙セットを定義した。半角英数字と 1 文字のカタカナ，ひらがなを除き，78,946 語となった。また，語彙セット中の単語からなる同義語の組は 1,578,374 組となった。1 語あたりの平均同義語数は，約 20 個となる。

(2) テキストデータ

新聞データ(CD-毎日新聞'91-'05，日経全文記事データベース'91-96)を用いた。

(3) 評価用単語ペアとラベル

語彙セットから，任意の単語を２個抽出し，評価用単語ペアとする。ただし，語彙セット中の全ての単語の組からなる全単語ペアに対し，同義語の占める割合は，0.1%未満に過ぎない。そのため，単純にランダムな語の組み合わせを全て評価するのではなく，同義語であると識別器が判定する可能性が高い単語ペアのみを選び評価することとした。具体的には，語彙セット中の全ての語に対して上位 1,500 語の同義語候補を抽出して評価用単語ペアを生成した。同義語候補抽出には，後述する 6 種類の文脈類似度全てについて実施し，得られた単語ペアを全てマージしたものを評価対象とした。

(4) 素性

文脈類似度の計算式において，κ を 0.1 から 0.6 まで 0.1 刻みで変化させて，それぞれの値で計算することにより得られた 6 種類の文脈類似度，および文字の類似度を用いない場合／用いた場合の 2 種類の表記類似度を素性として用いる。

(5) 識別器

識別器としては，svm_light[15]を用いた。カーネルは，予備評価の結果，最も精度が良かった多項式カーネル(２次)を用いた。

4.2 実験結果

萩原では，同義語ではないペアの数を同義語ペア数の 5 倍としていたため，同様の割合にサンプリングした結果に基づいて評価を行った。結果を表 1 に示す。従来方式は，文脈類似度によるランキングによる方式（教師なし）と萩原の方法に準じ，単語の出現文脈からなる教師データに対し，直接 SVM を適用した結果である。提案方式については，方式１は文脈類似度による素性のみを用いて SVM を適用した場合，方式２は文脈類似度に加え，表記類似度を用いた場合，方式３は表記類似度において，類似文字の情報も利用した場合である。

表 1 実験結果(平均適合率[%])

従来方式		提案方式		
教師なし	教師あり	方式１	方式２	方式３
48.7	53.5	61.6	67.1	77.3

従来方式（教師なし）の場合と比較して，教師あ

りの場合には，精度が明らかに良くなっており，同義語辞書を用いた教師あり学習の効果が大きいことが分かる。また，従来方式（教師あり）と方式１については，方式１が 8.1 ポイント良い結果となった。また，表記類似度を追加した方式２では，約 5.5 ポイント精度が改善している。更に，文字類似度を用いた方式３では，更に 10.2 ポイント精度が改善している。識別の閾値を変化させた場合の適合率，再現率の変化を図１に示す。

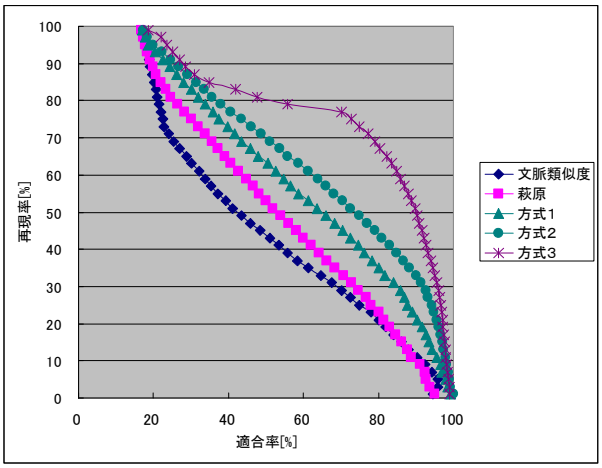

図 1　適合率と再現率の変化

方式３では，再現率が 80%以上の領域で急速に適合率が悪化しているが，これはカタカナ語の同義語の影響であると考えられる。

5. おわりに

既存の言語資源である同義語辞書を用いた，教師あり学習による同義語抽出方式を提案し，評価を行った。教師あり学習による同義語抽出は，同義語辞書の情報を有効に活用することにより高い精度が達成できるため，今後有望だと考えられる。本方式では，単語ペアを，テキストから得られる統計量や単語の文字列同士の類似度等から構成した素性のベクトルで表現し，同義語辞書を用いて正解を付与して，教師データを作成する。従来技術では，文脈となる単語そのものを素性として用いていたのに対し，提案方式では教師なし学習による同義語抽出で用いられる類似度計算式による結果を素性として用いる。予備評価の結果，従来方式と比較して，23.8 ポイントの精度向上が見られた。

今後，文書検索のような具体的なアプリケーションにおいて，どの程度の効果が得られるのかを明らかにしていく。

謝辞　日本語 WordNet を公開されている NICT の皆様，新聞記事を使用させて頂いた(株)毎日新聞社，(株)日本経済新聞社に謹んで感謝の意を表します。

参考文献

[1] 新里，鳥澤: HTML 文書からの単語間の上位下位関係の自動獲得，自然言語処理，12(1): 125-150, 2005.

[2] P. Turney: Mining the web for synonyms: PMI-IR versus LSA on TOEFL, Proc. of the 12th European Conference on Machine. Learning (ECML2001), pp.491-502, 2001.

[3] 大石他: 単語の共起関係と構文情報を利用した単語階層関係の統計的自動識別，情報処理学会研究報告.音声言語情報処理, SLP-61, pp.25-30, 2006.

[4] D. Lin: Automatic retrieval and clustering of similar words, Annual Meeting of the ACL archive, Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, pp.768-774, 1998.

[5] 相澤: 大規模テキストコーパスを用いた語の類似度計算に関する考察，情報処理学会論文誌，vol. 49-3, pp. 1426-1436 (2008).

[6] 王他: 単語類似度ネットワークを通じた自動同義語獲得，情報処理学会自然言語処理研究会報告，NL-185, pp.7-14 (2008).

[7] 曲，白井: 情報検索ための片仮名異表記処理，情報処理学会自然言語処理研究会報告，NL-162,pp.125-130,2004.

[8] K. Knight and J. Graehl: Machine Transliteration, Computational Linguistics, 24(4), pp. 599-612, 1998.

[9] 小山，竹内: 日本語複合語用語の入れ子関係に基づく階層的体系化，情報処理学会自然言語処理研究会報告, NL-180, pp.49-54, 2007.

[10] 渡部啓吾: 検索エンジンを用いた関連語の自動抽出，人工知能学会第 22 回全国大会，3B1-04, 2008.

[11] Masato Hagiwara: A Supervised Learning Approach to Automatic Synonym Identification based on Distributional Features, Proc. of ACL 2008 Student Research Workshop, pp. 1-6, 2008.

[12] Ramesh Nallapati: Discriminative models for information retrieval. SIGIR 2004: pp.64-71. 2003.

[13] 安田，今一，岩山，丹羽: 連想検索エンジンのスケーラビリティおよび障害耐性の向上，情報処理学会第 69 回全国大会, pp.383-384, 2007.

[14] Francis Bond, Hitoshi Isahara, Kyoko Kanzaki and Kiyotaka Uchimoto: Boot-strapping a WordNet using Multiple Existing WordNets, Proc. of the Sixth International Conference on Language Resources and Evaluation (LREC 2008), pp. 1619-1624, Marrakech, 2008.

[15] T. Joachims: Making large-Scale SVM Learning Practical, Advances in Kernel Methods - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.