読解支援システムのための語義曖昧性解消に関する研究

玉垣 隆幸　　白井 清昭
北陸先端科学技術大学院大学 情報科学研究科

1 はじめに

本研究における読解支援システムとは、オンラインで文章を読む際に、意味のわからない単語があれば辞書を引いて、その単語の意味を表示するシステムを指す。このようなシステムはいくつか提案されている [7] が、単語の意味が複数あるときはそれらを全て表示する場合が多い。ユーザは、ときには数十個の語義の定義文を読み、その中から文脈に合う正しい語義を選択しなければならない。したがって、システムが文脈を考慮して語義曖昧性解消を行い、正しい語義のみを提示することができれば、ユーザにとってより使いやすいシステムになると考えられる。

本研究では、正しい単語の意味だけをユーザに提示する読解支援システムの作成を目的とし、国語辞典を語義の定義とした語義曖昧性解消手法を提案する。読解支援システムでは、ユーザが意味を調べたい単語は多様であるので、様々な単語に対して語義の曖昧性を解消することが要求される。本研究では、複数の手法に基づく語義曖昧性解消システム (ここでは分類器と呼ぶ) を組み合わせて用いることにより、このことを実現する。具体的には以下の3つの分類器を用いる。

- 国語辞典の定義文を用いる方法
  1 辞書の用例を用いた分類器
  2 語義の出現条件を用いた分類器
- 語義タグ付きコーパスを用いる方法
  3 Support Vector Machine(SVM) による分類器

そして、この3つの分類器の出力 (語義) から最終的な出力を決定する。

2 関連研究

語義曖昧性解消の先行研究として、辞書の定義文を用いる手法 [2]、用例ベースの手法 [3]、語義タグ付きコーパスを用いた機械学習による手法 [5, 9, 12] などがある。これらのうち、近年は機械学習に基づく手法が有望であるとされており、75%から90%程度の正解率で曖昧性の解消ができると報告されている。しかし、これらの手法は訓練コーパスの頻出語についてのみしか語義の曖昧性を解消できない。また、語義タグ付きコーパスは作成コストが高く、全ての単語に対して十分な量の訓練コーパスが得られるわけではない。したがって、様々な単語に対して語義曖昧性解消を行うためには、機械学習による分類器とそれ以外の分類器を組み合わせて用いることが有効であると考えられる。本研究では、特に語義タグ付きコーパス以外の知識源を利用した分類器との組み合わせを試みる。

複数の分類器を組み合わせるアルゴリズムは過去にもいくつか提案されている。まず、語義タグ付きコーパスのみを知識源とし、学習アルゴリズムや学習に用いる素性集合を変えた分類器を組み合わせる手法がある [9, 5, 12]。これらの手法は、精度 (適合率) の向上は望めるが、適用率 (出力を返す単語の割合) の向上は望めない。一方、語義タグ付きコーパス以外の知識源を用いた分類器と機械学習による分類器を組み合わせる手法も提案されている。Agirre らは WordNet を用いた分類器と機械学習による分類器を組み合わせる手法を提案した [1]。これに対し、我々はコーパス以外の知識源として、シソーラスではなく国語辞典を用いる。一方、Litkowski は国語辞典を用いた分類器と機械学習による分類器を組み合わせる方法を提案している [6]。この研究では、辞書による分類器が出力する語義 (辞書の語義立て) と機械学習による分類器が出力する語義 (WordNet の意味クラス) が異なり、前者の語義を後者の語義に変換してから両者を混合している。これに対し、本研究では、各分類器は同じ語義を出力し、このような語義の変換を必要としない。

3 提案手法

本研究では、読解支援システムに用いる辞書として岩波国語辞典 [8] を用いる。すなわち、語義の定義は岩波国語辞典の語義立てに従うものとする。

3.1 辞書の用例を用いた分類器

岩波国語辞典では、定義文中に用例が記述されていることがある。動詞「愛する」の語釈文を図1に示す。図1において「子を—」、「国を—」、「酒を—」等、括弧で囲まれた部分が用例である。

用例を用いた分類器は、入力文と語釈文中の用例の類似度を計算し、最も類似度の高い用例を持つ語義を選択する。例えば、入力文が「彼は娘を愛している」のとき、図1中の3つの用例との類似度を計算する。その結果、「子を愛する」との類似度が高ければ、入

【愛する】
それに対し愛をそそぐ。
(1) かわいがり、いつくしむ。「子を—」。心から大切に思う。「国を—」
(2) 異性を恋い慕う。
(3) 物事を強く好む。「酒を—」

図 1: 「愛する」の語釈文

【慕う】
(1) 愛着の心をいだいてあとを追う。「母を—って三千里」。恋しく思って (心の中で) 追い求める。「故国を—」。「彼女がひそかに—青年」。
(2) 徳や学問・技量を敬い、これにならおうとする。「徳を—って集まる」。

図 2: 「慕う」の語釈文

力文「愛する」の語義として (1) を選択する。これは、過去に提案された用例に基づく手法 [3] と同じだが、用例データベースを作成する代わりに、辞書中の用例をそのまま用いる点が異なる。

次に、入力文と用例の類似度を計算する方法を説明する。類似度は、同じ格に立つ名詞の意味的類似度から求める。まず、各語義毎に用例から格 c の格要素となる名詞の集合 NE_c を抽出する。図 1 からは次のような格要素が抽出される。

【愛する】(1) $NE_{ヲ}$={ 子, 国 }
【愛する】(3) $NE_{ヲ}$={ 酒 }

岩波国語辞典では全ての語義に用例があるわけでなく、また用例から得られる格要素の数も十分ではない。そこで、用例から得られる格要素の数を増やすため、語釈文中の最後の動詞を上位語とみなし、上位語と元の語とでは似ている名詞が格要素として現れると仮定して、上位語の語釈文から格 c の格要素の集合 $NE_{c'}$ を抽出する。例えば、「愛する」の (2) の語義の上位語を「慕う」とし、「慕う」の語釈文の用例 (図 1) から格要素を抽出する。

【愛する】(2) $NE_{ヲ'}$ = { 母, 故国, 徳 }
$NE_{ガ'}$ = { 彼女 }

他の語義についても同様に格要素を抽出する。

入力文 s と用例文 e の類似度 $Sim(s,e)$ は式 (1) のように定義した。

$$Sim(s,e) = \sum_c w_c s_c(ns_c, NE_c) \quad (1)$$

$$s_c(ns_c, NE_c) = max_{ne_c \in NE_c} s(ns_c, ne_c) \quad (2)$$

$$s(w_i, w_j) = \frac{2 \times d_k}{d_i + d_j} \quad (3)$$

表 1: 格要素が得られた語義数と格要素数

	見出しのみ	上位語使用
多義である動詞の語義数	9,967	
格要素を獲得できた語義数	3,371	4,227
獲得できた格要素数	5,645	30,148

【さらに】
(1) その上に。重ねて。「—懇願する」「—は増援部隊も加わった」ますます。もっと。「—上達する」。
(2) 《あとに打消しを伴って》少しも。いっこうに。さらさら。「—反省の色がない」。

図 3: 「さらに」の語釈文 (抜粋)

式 (1) において、$s_c(ns_c, NE_c)$ は格 c に対する入力文の格要素 ns_c と用例 (上位語の用例を含む) から得られた格要素の集合 NE_c の類似度である。また、w_c はその重みである。w_c は経験的に定めた。特に、上位語の用例から抽出された格要素への重み $w_{c'}$ は、元の単語の用例から抽出された格要素の重み w_c よりも低くなるようにした。式 (2) において、$s_c(ns_c, ne_c)$ は二単語間の類似度で、式 (3) で定義される。式 (3) における d_i,d_j は単語 w_i,w_j のシソーラスにおける深さ、d_k は w_i と w_j の共通上位ノードのシソーラスにおける深さを表す。シソーラスは日本語語彙体系 [4] を使用した。

岩波国語辞典から得られた格要素の数を表 1 に示す。

多義である動詞のうち、用例から格要素を獲得できた語義の割合は 33.8%、上位語の用例も使用したときには 42.4%である。したがって、この分類器の適用率は低いが、他の分類器と組み合わせることにより、システム全体の適用率の向上が期待できる。

3.2 語義の出現条件を用いた分類器

岩波国語辞典では、ある語義が出現する条件が但し書きとして記述されていることがある。例を図 3 に示す。「さらに」の (2) の語義には、《あとに打消しを伴って》という但し書きの後に語義の定義が記述されている。したがって、入力文が「後悔しているようすなどさらにない」のとき、「さらに」の後に打ち消しの表現があるので、この語義は (2) であると推測できる。このように、岩波国語辞典に記述された語義の出現条件は、語義曖昧性解消の有効な手がかりとなる。

そこで、語義の出現条件を用いて語義曖昧性解消を行う分類器を作成した。この分類器は、候補となる全

ての語義について、入力文がその語義の出現条件を満たすかどうかを調べる。そして、出現条件を満たす語義があれば、これを正しい語義として出力する。また、複数の語義が出現条件を満たすときには、その全ての語義を出力する。出現条件を満たす語義がなければ、出力なしとする。

岩波国語辞典では、語義の出現条件は図3のように二重角括弧で囲まれて記述されていることが多い。そこで、岩波国語辞典おける多義の単語の語釈文から、二重角括弧で囲まれた記述を語義毎に取り出し、入力文がその条件を満たしているかどうかを判定するプログラムを作成した。プログラムとして実装した語義の出現条件の例を以下に挙げる。

- 単語の活用形に関する条件
  【快い】《主に連用形で》
  【眈む】《受身の形で》
- 前後の単語の表記、品詞、活用形に関する条件
  【くれる】《動詞連用形+「て」を受けて》
  【あがり】《名詞のあとに付く》
  【さっぱり】《多く「と」を伴って》
- 語義が出現する定型表現
  【いっぺん】《「いっぺんに」の形で》
  【否や】《「…や否や」「…と否や」の形で》
- 後に打ち消しの表現を伴うか否か
  【一切】《下に打消しを伴って、副詞的に》
  【てんで】《俗に、打消しを伴わずに》
- 文中での位置に関する条件
  【頂戴】《文末で》

出現条件を取り出すことのできた語義の数は973であった。このうち、582の語義について、条件を満たすか否かを判断するプログラムを実装した。岩波国語辞典における多義語の語義の総数は37,908なので、出現条件を判定するプログラムを実装できた語義の割合はわずか1.5%である。しかし、出現条件を取り出すことのできた語義は頻出単語の語義が多く、実際にこの分類器を用いるときの適用率はもっと大きくなると予想される。また、出現条件を満たすときには高い精度で正しい語義を選択できると期待される。

3.3 SVMによる分類器

語義タグ付きコーパスを利用してSVMを学習し、分類器を作成した。学習には以下の素性を用いた。

- 対象語の直前、直後にある単語の品詞
- 対象語の直前、直後にある単語の表記
- 対象語の前後n語以内にある自立語の基本形
  本研究では$n=20$とした。

SVMの学習にはLIBSVMを用いた[1]。ν–SVM [10]によって学習を行い、カーネルは線形カーネル、$\nu=0.0001$とした。SVMは二値分類器であるのに対し、本研究における語義曖昧性解消問題は多値問題である。そこで、pairwise法を用いてSVMを多値問題に適用した。

訓練コーパスとしてRWCコーパス [11]を使用した。RWCコーパスは、毎日新聞の3,000記事の単語に対して、岩波国語辞典の正しい語義が付与された語義タグ付きコーパスである。全3,000記事のうち、2,400記事を訓練データとして使用した。また、訓練データにおける出現頻度が10以上である1,585個の単語についてのみSVMの学習を行った。それ以外の単語については、SVMによる分類器は語義を出力しないとした。

3.4 混合モデル

本項では、3.1,3.2,3.3項で作成した分類器を組み合わせる方法について調べる。最初に、共通のテストデータ(ヘルドアウトデータ)を用意し、それぞれの分類器単体の正解含有率(式(4))を調べる。

$$\text{正解含有率} = \frac{\text{出力した語義に正解が含まれる単語数}}{\text{分類器が語義を一つ以上出力した単語数}} \quad (4)$$

そして、ヘルドアウトデータにおける正解含有率の一番高い分類器の出力を最終的な出力として選択する。但し、SVMについては単語毎に正解含有率を測定し、他の分類器の正解含有率との比較を行った。さらに、ヘルドアウトデータにおける頻度がOh以下の単語については、正解含有率の信頼性が低いので、全単語の平均の正解含有率をSVMの正解含有率とした。4節の実験ではOh=10とした。なお、用例に基づく分類器や語義の出現条件を用いた分類器についても単語毎に正解含有率を計算して比較するべきであるが、本研究では行わなかった。

4 評価実験

提案手法を評価する実験を行った。RWCコーパスのうち、300記事をヘルドアウトデータ、300記事をテストデータとした。これらはSVMの訓練データとは異なる記事である。ヘルドアウトデータ、テストデータにおける評価単語数はそれぞれ15,194、14,523であった。

ヘルドアウトデータにおける各分類器の正解含有率を表2に示す。出現条件を用いた分類器が最も正解

[1] http://www.csie.ntu.edu.tw/%7Ecjlin/libsvm/
[2] F値は$2PR/(P+R)$とした。(Pは精度、Rは再現率)

表 2: ヘルドアウトデータにおける正解含有率

	SVM	用例	出現条件
正解含有率	0.78	0.49	0.81

表 3: テストデータにおける各手法の評価

	精度	再現率	F 値[2]	適用率
混合モデル	0.7118	0.7286	0.7201	0.9007
用例	0.4081	0.0555	0.0977	0.1102
出現条件	0.5396	0.1875	0.2783	0.2272
SVM	0.7844	0.6955	0.7373	0.8867

含有率が高く、ついで SVM による分類器、用例を用いた分類器の順になった。

表 3 は、テストデータに対する提案手法の評価結果である。"用例"、"出現条件"、"SVM" はそれぞれの分類器を単独に使用したときの結果で、"混合モデル" は 3.4 項で述べた方法で 3 つの分類器の出力を組み合わせたときの結果である。また、表 4 は、混合モデルによって選択された分類器の数である。混合モデルの再現率は、単独のどの分類器よりも高い。また適用率も最も高い値が得られている。本研究の目的は、読解支援システムでの使用を前提とし、複数の知識源を用いた分類器を組み合わせることにより、より多くの単語について語義の曖昧性を解消することにある。表 3 の結果から、この目的がある程度達成されたことが確認された。しかし、SVM を用いた分類器と比べると、再現率の差は 3 %程度であり、改善の余地がある。一方、精度を比較すると、混合モデルは SVM を用いた分類器を比べて 7 %程度劣る。また F 値も約 2 %低い。これは、用例を用いた分類器や出現条件を用いた分類器の精度が低いためと考えられる。

5 おわりに

本研究では、国語辞典を利用した 2 種類の分類器と語義タグ付きコーパスから学習された分類器を組み合わせて語義曖昧性解消を行う手法を提案した。実験の結果、複数の分類器を組み合わせることにより再現率が向上することが確認された。

今後の課題として、まず国語辞典を利用した分類器を改良することが挙げられる。用例文を用いた分類器は動詞の語義しか曖昧性を解消できず、また語義の出現条件を用いた分類器も適用率が低い。単独の分類器の適用率を向上させるか、あるいは辞書の定義文を利用した新たな分類器を組み合わせる必要がある。また、国語辞典を用いた 2 つの分類器は SVM に比べて精度が低いので、これを改善することも課題の一つである。次に、複数の分類器を組み合わせるアルゴリズムを改良することにより、再現率や精度がさらに向上する可能性がある。正解含有率による重み付き投票を行う手法などを検討したい。最後に、本手法を読解支援システムに組み込むことが挙げられる。現在、日本語学習者用の読解支援システム「あすなろ」[7] に本手法を組み込むことを検討している。

表 4: 混合モデルで選択された分類器の数

用例	出現条件	SVM	回答なし
363	2,873	9,846	1,442

参考文献

[1] E. Agirre et al. Combining supervised and unsupervised lexical knowledge methods for word sense disambiguation. *Computers and the Humanities*, Vol. 34, No. 1,2, pp. 103–108, 2000.

[2] Jim Cowie et al. Lexical disambiguation using simulated annealing. *Proceedings of COLING*, pp. 359–365, 1992.

[3] Atsushi Fujii et al. To what extent does case contribute to verb sense disambiguation? In *Proceedings of COLING*, pp. 59–64, 1996.

[4] 池原 悟ら. 日本語語彙体系－全五巻－. 1997.

[5] Dan Klein et al. Combining heterogeneous classifiers for word-sense disambiguation. *Proceedings of the SIGLEX/SENSEAVAL Workshop on Word Sense Disambiguation*, pp. 74–80, 2002.

[6] Kenneth C. Litkowski. Sense information for disambiguation: Confluence of supervised and unsupervised methods. *Proceedings of the SIGLEX/SENSEAVAL Workshop on Word Sense Disambiguation*, pp. 47–53, July 2002.

[7] 仁科 喜久子ら. 構文表示と多言語インターフェースを備えた日本語読解学習支援システムの開発. 言語処理学会第 8 回年次大会, pp. 228–231, 2002.

[8] 西尾実, 岩淵悦太郎, 水谷静夫. 岩波国語辞典 第五版. 1994.

[9] Ted Pedersen. A simple approach to building ensembles of naive baysian classifiers for word sense disambiguation. *Proceedings of NAACL*, pp. 63–69, 2000.

[10] Bernhard et al.x Schölkopf. New support vector algorithms. *Neural Computation*, Vol. 12, pp. 1083–1121, 2000.

[11] 白井 清昭ら. 岩波国語辞典を利用した語義タグ付きテキストデータベースの作成. 情報処理学会自然言語処理研究会, pp. 2000(9):117–122, 2001.

[12] Hiroya Takamura et al. Ensembling based on feature space restructuring with application to WSD. In *Proceedings of NLPRS*, pp. 41–48, 2001.