語義曖昧性解消におけるシソーラス利用の問題分析

新納 浩幸　佐々木 稔　古宮 嘉那子
茨城大学 工学部情報工学科
{shinnou, msasaki, kkomiya}@mx.ibaraki.ac.jp

1　はじめに

Project Next NLP[1] の「語義曖昧性解消・新語義発見」チーム[2]は、語義曖昧性解消（Word Sense Disambiguation; WSD）の誤り分析を通して、WSD の本質的問題や、WSD に関して今後研究すべき重要事項を明らかにしようとしている。我々はこのチームのメンバーであり、上記の活動の一環として、WSD におけるシソーラスの利用について分析した。ここではその分析結果と考察を述べる。

WSD を教師付き学習のアプローチで解決する場合、単語の上位概念を素性として用いることは一般に行われている。シソーラスはその上位概念を得るために利用される。しかし以下の点が未だ明らかになってはいない。

- どのレベルの上位概念（シソーラスの粒度）を用いればよいか
- 複数の意味を持つ単語の上位概念は複数存在するが、この曖昧性は無視して良いのか
- コーパスを利用して単語をクラスタリングすることで、シソーラスを自動構築できるが、このようなシソーラスと手作業で作成されたシソーラスとでは WSD での利用に差はあるのか

当然、考えられる結論として、上記の問題は対象単語に依存するはずである。つまり適切なシソーラスは対象単語毎に異なる。ここではこの問題に対して、シソーラスをアンサンブルして利用することも行った。また交差検定により対象単語毎に適切なシソーラスを選択・利用することも行い、これらのアプローチが有効かどうかを確かめた。

[1] https://sites.google.com/site/projectnextnlp/
[2] http://nlp.dse.ibaraki.ac.jp/~shinnou/ProjectNextNLP/

2　評価データと WSD システム

WSD に利用するシソーラスの違いによる精度の比較を行うためには、評価データと基本となる WSD システムが必要になる。

評価データは SemEval-2 の日本語 WSD タスクから作成した [2]。SemEval-2 のデータは対象単語が 50 単語あり、各対象単語に対して 50 個の訓練用例と 50 個のテスト用例が存在する。

基本システムは SemEval-2 のコンペの際に baseline とされたシステムを実装した。学習アルゴリズムは SVM であり、以下の 20 種類の素性を利用する。

e1= 二つ前の単語
e2= 二つ前の品詞
e3= その細分類
e4= 一つ前の単語
e5= 一つ前の品詞
e6= その細分類
e7= 問題の単語
e8= 問題の単語の品詞
e9= その細分類
e10= ひとつ後の単語
e11= ひとつ後の品詞
e12= その細分類
e13= 二つ後の単語
e14= 二つ後の品詞
e15= その細分類
e16= 係り受け
e17= ふたつ前の分類語彙表の値（5桁）
e18= ひとつ前の分類語彙表の値（5桁）
e19= ひとつ後の分類語彙表の値（5桁）
e20= ふたつ後の分類語彙表の値（5桁）

本来の baseline のシステムでは分類語彙表 ID の 4 桁と 5 桁を同時に使う形になっていたが、ここでのシステムでは 5 桁のみとした。また一般に一つの単語に対しては複数の分類語彙表 ID が存在するので、e17,e18,e19,e20　に対する素性は複数になる場合もある。SVM の学習は libsvm の線形カーネルを用いた。指定できるパラメータは全て default のままである。

Copyright(C) 2015 The Association for Natural Language Processing.
All Rights Reserved.

3　シソーラスの利用における問題点

3.1　シソーラスの粒度

ここでの WSD システムは分類語彙表 ID の 5 桁を利用している。分類語彙表では桁数が少ないほどより上位の概念を表している。

例えば、「側面」の分類語彙表での ID は 1.1750,1,1,4 である。分類語彙表 ID の 5 桁というのは、第 1 フィールドの 1 桁の数値と第 2 フィールドの 4 桁の数値とを合わせたものであり、「側面」の場合 11750 となる。分類語彙表の第 3 フィールドまで見ると 117501 となり 6 桁の数値となる[3]。

6 桁が 117501 であるものは以下の 4 つであり、これらの単語の上位概念のコードが 117501 と見なせる。

```
四面, しめん,1.1750,1,1,5
側面, そくめん,1.1750,1,1,4
部面, ぶめん,1.1750,1,1,3
面, めん,1.1750,1,1,2
```

分類語彙表 ID の 5 桁が 11750 となる単語を集めると、以下のようになり、6 桁の 117501 から 1175010 までの上位概念のコードが 11750 と見なせる。

```
四面, しめん,1.1750,1,1,5
側面, そくめん,1.1750,1,1,4
部面, ぶめん,1.1750,1,1,3
面, めん,1.1750,1,1,2
月面, げつめん,1.1750,10,1,4
地面, じめん,1.1750,10,1,2
路面, ろめん,1.1750,10,1,3
....
スロープ, すろおぷ,1.1750,8,2,2
凹面, おうめん,1.1750,8,3,1
凸面, とつめん,1.1750,8,3,2
クラインの壷, くらいんのつぼ,1.1750,8,4,2
メビウスの帯, めびうすのおび,1.1750,8,4,1
新生面, しんせいめん,1.1750,9,1,2
生面, せいめん,1.1750,9,1,3
```

このように分類語彙表では ID を数値と見なしたときの左から数える桁数が上位概念のレベルに対応している。

ここでの WSD システムは単語の上位概念のコードとして、分類語彙表 ID の 5 桁を使ったが、4 桁（上記例では 1175）あるいは 3 桁（上記例では 117）のコードを使うことも可能である。どのレベルの上位概念を使うのが良いのか、つまりどの粒度のシソーラスを使うのが良いのかは明かではない。

本論文の実験では、この桁数を 3 桁あるいは 4 桁にした場合の評価データにおける正解率を調べた。また分類語彙表を用いない場合の正解率も調べた。

[3]第 3 フィールドの数値は 10 以上の数もあるので、結果的には 6 桁以上の数値になる場合もある。

3.2　上位概念の曖昧性

複数の意味を持つ単語では、複数の上位概念が存在する。

例えば「頭」には以下の 5 つの分類語彙表の ID が存在する。

```
頭, あたま,1.1404,2A,1,2
頭, あたま,1.1960,54,1,2
頭, あたま,1.1960,64,2,1
頭, あたま,1.2430,8,1,2
頭, あたま,1.5710,1,1,3
```

ここでの WSD システムでは「頭」の上位概念を使う場合、それらを単に並べて利用している。例えば、直後の単語が「頭」であった場合は、以下の 4 つの素性が生成される[4]。

```
e19=11404, e19=11960, e19=12430, e19=15710
```

しかしこのような処理には妥当性がない。理想的にはその文での「頭」の語義を決定し、その語義に対応する上位概念のコードのみを使うべきである。

本論文では、簡易な手法により、曖昧な分類語彙表のコードを一意に決めることにした。一意に決めた場合の評価データにおける正解率を調べた。

上記の「簡易な手法」[7] を概説する。BCCWJ のコアデータに含まれる単語を分類語彙表の 5 桁を使って、全て上位概念に置き換える。その際、曖昧なものも全て列挙する。次に置き換えられた上位概念の頻度を調べておく。システムにおいて、曖昧な分類語彙表のコードが生じた際には、この頻度表を参照し、最も頻度の高いものを利用することにした。

3.3　名詞クラスタリングとの差

WSD で利用するシソーラスとしては、通常、分類語彙表か日本語語彙体系が使われる。ただしシソーラスは単語の上位概念のコードを与えるだけに利用されるので、名詞をクラスタリングした結果がシソーラスの代用になる。名詞のクラスタリングは大規模なコーパスがあれば可能であり、利用したコーパスの分野に適したシソーラスが構築できる、単語のカバレッジが大きい、などの長所がある。

ここでは新聞記事 4 年分（毎日新聞 '95 ～ '98) から頻度上位 10000 個の名詞のクラスタリングを行った。クラスタ数は 1000 のもの NC-1K と 2000 のもの NC-2K の 2 種類を作成した[5]。分類語彙表での上

[4]1.1960,54,1,2 と 1.1960,64,2,1 はマージされる

[5]文内での単語間の共起頻度を調べ、10000 個の各名詞を 10000 次元のベクトルで表現し、それをもとに bayon によりクラスタリングを行った。

Copyright(C) 2015 The Association for Natural Language Processing.
All Rights Reserved.

位概念をこのクラスタ番号に置き換えて、評価データにおける正解率を調べた。

4 実験

4.1 各シソーラスの比較

実験の結果を表 1 に示す。表中の NO はシソーラスを利用しなかったことを示し、B3、 B4、 B5 はそれぞれ分類語彙表の 3 桁、4 桁、5 桁を利用したことを示す。AMB は前述した分類語彙表の曖昧性を除去した場合を示す。また NC-1K と NC-2K はシソーラスの代わりに、名詞クラスタリングの 1000 クラスターのものと 2000 クラスターのものとで代用したことを示す。表中の値は評価データ (SemEval-2 日本語辞書タスク）における正解率である。

NO	B3	B4	B5 (base)	AMB	NC -1K	NC -2K
75.72	77.08	77.04	76.92	76.44	76.60	76.52

表 1: 各シソーラス利用時の正解率 (%)

どのシソーラスを使ってもほとんど差がない。ただし、使わない場合 (NO) は明らかに他のものよりも正解率が悪く、シソーラスは使わないよりも使った方がよいということは言える。また NC-1K と NC-2K の差もほとんどないことから、WSD で必要とされるシソーラスの粒度はかなり粗いものでもよいと考えられる。

4.2 シソーラスのアンサンブル

従来より粒度が異なるシソーラスはそれらを並べて利用することが行われてきた。例えば SemEval-2 の baseline のシステムでも分類語彙表の 4 桁と 5 桁を同時に利用している。例えば `e19=11750` は 5 桁、`e19=1175` は 4 桁であるが、4 桁と 5 桁を同時に使うというのは `e19=11750, e19=1175` と素性を並べて利用することを意味する。これはシソーラスをアンサンブルしていることに対応する。

ここでは分類語彙表の 4 桁と 5 桁のアンサンブルの他、B5 と NC-1K あるいは B5 と NC-2K のアンサンブルも試した、結果を表 2 に示す。表中の B45 が分類語彙表の 4 桁と 5 桁のアンサンブル、B5-NC-1K と B5-NC-2K が B5 と NC-1K あるいは B5 と NC-2K のアンサンブルを示す。

シソーラスをアンサンブルすることは精度向上に効果があることがわかる。

B5 (base)	B45	B5-NC-1K	B5-NC-2k
76.92	76.96	77.28	77.24

表 2: シソーラスのアンサンブルの正解率 (%)

4.3 交差検定によるシソーラスの選択

ここまでの実験で 10 種類のシソーラス（NO, B3, B4, B5, AMB, NC-1K, NC-2K, B45, B5-NC-1K, B5-NC-2k）を利用した。WSD では対象単語毎に最適なシソーラスは異なると考えられる。そこで 10 分割交差検定を利用して、対象単語毎に最適なシソーラスを選択、利用することで評価データの正解率を求めた。なお交差検定で優劣が付かない場合、先の実験の正解率順の以下で選択を行った。

```
NO < AMB < NC-2K < NC-1K < B5 < B45 <
  < B4 < B3 <  B5-NC-2K <  B5-NC-1k
```

仮に最良の選択が行えたとしたら、正解率は 79.68% になるが、交差検定での選択では 76.28% となった。効果はなかったと言える。

5 考察

シソーラスの粒度に関してはここでの実験では差が出なかった。わずかではあるが 3 桁が最も正解率が高く、次いで 4 桁、5 桁である。分類語彙表の 3 桁の名詞の種類数は 43、4 桁は 303、5 桁は 544 である。また 2 桁の実験も行ったが、その正解率は 75.60% であり、これは分類語彙表を利用しないものよりも低かった。ここからシソーラスの粒度に関してはかなり粗い（100 種類程度の上位概念）ものでも WSD には十分と考えられる。

上位概念の曖昧性に関しても、解消することで効果が出るのかどうかは疑わしい。おそらく解消できたとしても、有意差が出るような大きな差は生じないと予想する。そのため上位概念の曖昧性への対処は現実的には無用だと予想する。ただし平仮名で記された単語に対しては、シソーラスを使わないという対処はあり得る。例えば「こと」のように平仮名で記された単語は、一般に、複数の上位概念が生じるが、それらは悪影響の方が大きく、利用しない方がよいと思われる（例えば [4] などはこの処理を行っている）。

また上位概念の曖昧性を解消するのは WSD そのものであり、WSD のための素性を作る段階で WSD を行うことには、通常の WSD の枠組みでは難しい。これは all words の WSD [8] と問題的には等価となっ

Copyright(C) 2015 The Association for Natural Language Processing.
All Rights Reserved.

ている[6]。ここで試した手法は非常に簡易なものであり、曖昧性が解消されているかどうかも怪しいが、all words の WSD システムのベースラインのシステムとして利用できると考えている。通常、all words の WSD はベースラインがランダムな語義の選択になっているが、ランダムな選択よりは多少は改善されていると思われる。

WSD にシソーラスを利用する場合、最適なシソーラスは対象単語に依存すると考えられる。実験では交差検定により、最適なシソーラスを選択しようとしたが、全く効果がなかった。WSD には本質的に領域適応 [5] の問題が生じていると考えている。領域適応が生じている場合、実験で行ったような交差検定は効果がないのは明かである。最適なシソーラスは、単語の言語的な特徴から予想するアプローチの方が優れている。また複数のシソーラスをアンサンブルすることは効果があった。この仕組みを突き詰めていくことが WSD の精度改善につながると予想する。この点が今後の課題である。

最後にシソーラスを利用して上位概念を素性として利用するというアプローチは、原理的には対象単語の周辺文脈をある空間に射影し、その空間上の点を素性として利用している形であることを注記しておく。通常のシソーラスの利用では、この空間が離散的であるために、粒度の問題が生じているが、連続的なものにすれば粒度の問題は生じない。同時に上位概念の曖昧性の問題も軽減される。これは名詞間距離を設定すること [3] でも可能であるが、より精緻に文脈を表現するためにトピックモデルを利用したり [1]、Deep Learning の手法を応用することも考えられる [6]。これらはシソーラスを自動構築するアプローチの発展形といえる。WSD システムの精度改善には、この方向でのアプローチが有望だと考えている。

6 おわりに

ここでは WSD にシソーラスを利用する際の曖昧になっている問題点をあげ、それに対して様々なシソーラスと SemEval-2 日本語辞書タスクのデータを利用した実験、考察を行った。得られた結果は以下にまとめられる。

- WSD ではシソーラスを使わないよりも使った方がよい
- シソーラスの粒度は精度にはあまり関係ない、おそらくかなり粗くてもよい
- 上位概念の曖昧性を解消することは困難であり、また解消しても WSD の精度が向上するかどうかは疑問である
- 異なるタイプのシソーラスをアンサンブルすることは効果がある
- 交差検定で単語毎に最適なシソーラスを選択するというアプローチはうまくいかない

今後は上記の 4 番目の項目を深めたい。その仕組みを考えることで、有効な素性を構築できる可能性がある。また対象単語の周辺文脈を適切な空間に射影し、その点を素性とするアプローチがシソーラスを利用するアプローチの発展形と見なせるため、このアプローチも追求していきたい。

[6] 概念を語義 ID に対応させる必要があるので、厳密には異なる。

参考文献

[1] Jun Fu Cai, Wee Sun Lee, and Yee Whye Teh. Improving Word Sense Disambiguation using Topic Features. In *the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 1015–1023, 2007.

[2] Manabu Okumura, Kiyoaki Shirai, Kanako Komiya, and Hikaru Yokono. On SemEval-2010 Japanese WSD Task. 自然言語処理, Vol. 18, No. 3, pp. 293–307, 2011.

[3] Hiroyuki Shinnou. Redefining similarity in a thesaurus by using corpora. In *COLING-96*, pp. 1131–1135, 1996.

[4] Hiroyuki Shinnou and Minoru Sasaki. Unsupervised learning of word sense disambiguation rules by estimating an optimum iteration number in the em algorithm. In *Seventh Conference on Natural Language Learning (CoNLL-2003)*, pp. 41–48, 2003.

[5] Anders Sogaard. *Semi-Supervised Learning and Domain Adaptation in Natural Language Processing*. Morgan & Claypool, 2013.

[6] 河野和平, 新納浩幸, 佐々木稔, 古宮嘉那子. Stacked denoising autoencoder を利用した語義曖昧性解消の領域適応. 言語処理学会第 21 回年次大会, p. (to appear), 2015.

[7] 江口晃, 新納浩幸, 佐々木稔. 名詞の主要語義の推定と語義識別への応用. 言語処理学会第 16 回年次大会, pp. PB1–4, 2010.

[8] 佐々木悠人, 古宮嘉那子, 森田一, 小谷善行. 周辺語義モデルによる日本語の教師無し語義曖昧性解消. 情報処理学会第 218 回自然言語処理研究会, pp. NL–218–3, 2014.

Copyright(C) 2015 The Association for Natural Language Processing.
All Rights Reserved.