BCCWJ を用いた新しい語義曖昧性解消タスク

奥村学
(東京工業大学 精密工学研究所)

白井清昭
(北陸先端科学技術大学院大学
情報科学研究科)

概 要

国立国語研究所の前川喜久雄氏を領域長として，現代日本語書き言葉の大規模な均衡コーパス (「現代日本語書き言葉均衡コーパス」，BCCWJ; Balanced Corpus of Complementary Written Japanese と呼ばれている) を構築するとともに，それを活用した研究によりコーパスを評価するプロジェクトが進められている (http://www.tokuteicorpus.jp/)．この中で我々は現在，代表性のある語義タグ付コーパスの構築を行っている．このコーパスを利用して，後述する 2 つの特徴を持つ語義曖昧性解消 (WSD; Word Sense Disambiguation) の評価型タスクを，SemEval-2(http://semeval2.fbk.eu/Semeval2.html) の下で開催予定である．本稿では，このタスクの紹介を行う．

1　はじめに

語義曖昧性解消は，意味解析技術の一つとして，古くから自然言語処理分野で研究が進められている技術である．語義曖昧性解消では，複数の語義をもつ単語を対象に，与えられた文脈中で，辞書中のその単語の語義区分に基づき, どの語義で用いられているかを自動判定する．この技術の水準向上を目的とした評価型ワークショップが過去何度か開催されている (Senseval-1/2/3, SemEval-2007[1])．その中では，様々な言語における語義曖昧性解消タスクが設定されてきており，また，最近では，あらかじめ決められた語義区分を仮定することなく，与えられた用例集合をクラスタリング等することにより，単語の語義区分を同定するタスクも設定されていたりする．

語義曖昧性解消に関するこの評価型ワークショップは 3 年に 1 度のペースで開催されており，次回の Semeval-2 は 2010 年にワークショップが開催予定である (http://semeval2.fbk.eu/Semeval2.html)．この Semeval-2 に我々は後述する 2 つの特徴を持つ語義曖昧性解消の評価型タスクを提案し，無事採択された．本稿では，このタスクの背景, 狙う点, 課題の内容等について主に説明する．

2　BCCWJ コーパスを用いた意味解析

国立国語研究所の前川喜久雄氏を領域長として，文部科学省科学研究費補助金特定領域研究「日本語コーパス」プロジェクトが 2006 年 9 月にスタートしている (http://www.tokuteicorpus.jp/)．このプロジェクトでは，現代日本語書き言葉の大規模な均衡コーパス (「現代日本語書き言葉均衡コーパス」，BCCWJ; Balanced Corpus of Complementary Written Japanese と呼ばれている) を構築するとともに，それを活用した研究によりコーパスを評価することを目指している．このコーパスは，様々なジャンルのテキストからなり代表性があるという特徴をもつと同時に，10-20 年程度の時間幅を持ったテキストからなる部分を持ち，継時性があるという特徴も合わせ持っている．これらの特徴を持つコーパスはあまりなく，これらの特徴を持つコーパスを利用し言語処理研究を行うことには大きな意味がある．そこで，我々は，このプロジェクトの一貫として，以下の 2 つのテーマを選んで，上述したコーパスの特徴を活かした日本語の意味解析に関する研究を行っている[2]．

1. 意味解析では，単語の語義を自動的に同定すること (語義曖昧性解消) も 1 つのテーマであるが，この語義曖昧性解消を代表性のあるコーパスで行う場合には，どうすれば良いのだろうか．様々なジャンルのテキストでは，同じ単語についても，出現する語義の分布は異なるため，従来のように，特定のジャンルのテキスト (たとえば，新聞データ) を対象にした手法が同じようにうまく行くとは限らない．

2. 語義曖昧性解消は，あらかじめ仮定した辞書中の語義のどれであるかを決めようとするが，そもそも用例の語義が辞書中にない場合どうすれば良い

[1] http://nlp.cs.swarthmore.edu/semeval/index.php, http://www.senseval.org/senseval3/, http://193.133.140.102/senseval2/, http://www.itri.brighton.ac.uk/events/senseval/ARCHIVE/index.html.

[2] 詳細は，[5] を参照していただきたい．

のか．継時性があるコーパスでは，辞書に載っていない語義が時間経過を経て出現することはかなりあるように思われる．辞書中に載っていない，いわゆる新語義を発見できれば，辞書編集者に貢献することもできる．

3 代表性のある語義タグ付コーパスの構築

この中で我々は現在，代表性のある語義タグ付コーパスの構築を行っている．領域内で公開されているコアデータ (BCCWJ を構成するように，サンプリングされたデータ) に対して, 岩波国語辞典中の語義の区分に基づき, 人手で語義を付与する作業を行っている．過去のタグ付コーパス構築例にならい [6], タグ付けの際, 辞典中に該当の語義が見当たらない場合「該当なし」という判断を許し, また, 最下層の語義のどれかでは判断できない場合, より上位のラベルを付与することを許している．「該当なし」の場合，大辞林をひき，該当する語釈文があれば，それを明記し，該当するものがなければ，作業者自身が考えた語釈文を記載してもらうようしている．

日本語の語義タグ付コーパスには，EDR コーパス (20 万文)，RWC コーパス (3000 記事) があるが，いずれも代表性のあるコーパスを元にしていない．海外では，代表性のあるコーパスの上にタグ付けを行うことで，代表性のある語義タグ付コーパスの構築が進んでおり，日本語においての構築は急務であると考える．

4 BCCWJ を用いた新しい語義曖昧性解消タスク

このコーパスが利用できるようになると，以下のような特徴を持つ語義曖昧性解消 (WSD; Word Sense Disambiguation) の評価型タスクが設定できる．

1. 日本語の語義タグ付コーパスはこれまですべて新聞データを元にしていたが，日本語で最初の代表性のある語義タグ付コーパスを用いた WSD タスクとなる．

2. これまでの WSD タスクでは，あらかじめ仮定した辞書中の語義セットから語義を選択する必要があったが，実際には辞書中に該当する語義がない用例も多数存在する．そのような，辞書中に語義がない用例も対象とする，初めての WSD タスクとなる．

4.1 代表性のあるコーパスを用いた語義曖昧性解消

代表性のあるコーパス中には, 複数のジャンルのテキストが混在していることになる．したがって, コーパスは, いくつかのジャンルごとのサブコーパスに分割できることになる．近年の語義曖昧性解消研究では, 訓練コーパスを用いて分類器を学習し, その分類器により語義を同定する (ある単語の出現がその単語の語義のうちどの語義の出現であるか分類する) 手法が採用されることが多く, また, より良い性能を得られている．この時，単語によっては，サブコーパスごとに，出現する語義の頻度分布が異なる場合が存在する．すると, あるジャンルのテキスト中の用例を対象に語義曖昧性解消しようとする時，同一ジャンルのサブコーパスを学習に利用するのが良さそうであるのは言うまでもないが，それ以外に，コーパス中のどのサブコーパスをどのように学習に利用するのが良いのかは自明な問題ではない．これはある種の領域適応 (domain adaptation) の問題であるが，これまでのように単一ジャンルのテキスト (たとえば，新聞データ) を利用していた場合にはさほど顕在化していない問題である．

なお，語義曖昧性解消における領域適応に関する先行研究には，たとえば，[2, 1] などがある．

4.2 新語義の発見

従来の語義曖昧性解消では，単語の語義を辞書などによってあらかじめ定義し，これらの語義の中からテキスト中の単語に対する適切な意味を選択する．ところが，単語の意味は年月とともに変化し，新しい語義や用法も日々生まれている．そのため，単語の語義をあらかじめ定義するのは必ずしも適切であるとは言えない．そこで，ある用例における単語の意味が既存の辞書に定義された意味に該当するのか，あるいは辞書の意味のいずれにも該当しない新語義なのかを判定することにより，単語の新語義を発見するという必要が生じる．

例を上げよう．岩波国語辞典で単語「ネタ」には，(1) 新聞記事などの材料, (2) 手品の仕掛け, (3) 証拠, (4) (料理の材料としての）食物，の 4 つの語義が与えられている．このとき，以下の用例 (a) の語義は，(2) に近いが,「何かの大事な部分」を指しており，また，(b) の語義は，4 つの語義のうちにはなく,「作り話」のような意味で用いられている．

> ストーリー上、必ず使いますか？　「ネタ」ばれでも良いので教えてください。(a)
> まさにその通りですね。。。。。「ネタ」みた

いに見えるけどほんとそうだよ (b)

このような用例を発見し，これらの用例の語義が辞書中には与えられていないものであると判定することが新語義発見の目標である．

4.3 新語義発見を含む語義曖昧性解消へのいくつかのアプローチ

上述したように，近年の語義曖昧性解消では, 訓練コーパスを用いて分類器を学習し, その分類器により語義を同定する手法が採用されることが多い．では，新語義発見を含む語義曖昧性解消では，どのような手法を採ることができるだろうか．

新語義発見を含む語義曖昧性解消でも，単純には,「新語義」というクラスを，対象とする単語の (辞書中に列挙されている) 語義クラスの集合に追加し，その中から語義クラスを選択する分類器を学習することで，同様の手法が採れそうである (discrimination-based な手法)．ただ，この手法の場合,「新語義」クラスに対する訓練データが非常に少ない (あるいは仮定できない) という問題に対処する必要がある．

新語義発見を含む語義曖昧性解消では，これ以外にも，様々な手法が考えられそうである．これも素朴だが，コーパスに出現する単語の用例集合を，その単語が出現する文脈の類似性という観点からまずクラスタリングし，このとき，同じ意味を持つ用例集合は同じクラスタに分類されると考え，クラスタリングによって得られたクラスタが単語の語義に対応すると仮定する (induction-based な手法)．そして，得られたクラスタを辞書中の語義クラスと対応付け，対応付けできなかったクラスタを新語義に対応するクラスタと考える．

このように，新語義発見を含むように語義曖昧性解消を拡張すると，採りうる手法もより多様になり，タスクもよりおもしろさを増していると言うことができよう．

なお，新語義発見に関する先行研究はあまり多くはないが，前者のアプローチを採る研究として [3] が，後者のアプローチを採る研究として [4, 7, 8] がある．

4.4 課題設定

課題の詳細は以下の通りである．なお，Semeval-2 傘下のタスクであるため，公式なタスク定義はすべて英語で記述されている．以下の英語の部分は，タスクの web ページ (`http://lr-www.pi.titech.ac.jp/wsd.html`) 中の記述の抜粋である．

Task description:

This task can be considered an extension of SENSEVAL-2 JAPANESE LEXICAL SAMPLE Monolingual dictionary-based task. Word senses are de ned according to the Iwanami Kokugo Jiten, a Japanese dictionary published by Iwanami Shoten. Please refer to that task for reference.

Input: Test documents with marked target words from the BCCWJ corpus, where the genre of documents is also provided, because of their diversity. Examples include books, newspaper articles, white papers, blogs, magazines, and documents from a Q&A site on the WWW.

Output: The sense ID of each target word in the Iwanami Kokugo Jiten if the sense is in the dictionary. If systems nd that the sense is not in the dictionary, say 'new sense.'

The evaluation methodology:

Organizers will return the evaluation in two ways:

a. evaluating the outputted sense IDs, assuming the 'new sense' as another sense ID. The outputted sense IDs will be compared to the given gold standard word senses, and the usual precision measure for supervised word sense disambiguation systems will be computed using the standard SENSEVAL scorer. The Iwanami Kokugo Jiten has three levels for sense IDs, and we use the middle-level sense in the task. Therefore, we can call the scoring in the task 'middle-grained scoring.'

b. evaluating the ability of nding the instances of new senses, assuming the task as classifying each instance into a 'known sense' or 'new sense' class. The outputted sense IDs (same as in a.) will be compared to the given

> gold standard word senses, and the usual accuracy for binary classi cation will be computed, assuming all sense IDs in the dictionary are in the 'known sense' class.

The availability of the resources:

> The Iwanami Kokugo Jiten will be available soon from GSK (`http://www.gsk.or.jp/`). A corpus annotated with sense IDs will also be distributed as training data. Each article will be assigned its genre code. Participants in this task are required to submit a copyright agreement form to the National Institute of Japanese Language.

4.5 今後の日程

残念ながら，最終のワークショップの日程が確定していないため，タスクの日程も確定していないが，現時点ではおよそ以下のような日程を予定している．

1. 訓練データセット，スコアラ，入出力フォーマット等を 2009 年夏にリリース予定
2. Formal run を 2009 年度後半実施予定
3. SemEval Workshop は ACL-2010 に併設予定[3]

また，ACL に併設されるワークショップは，Semeval-2 に採択されたすべてのタスクを対象にしたワークショップであるが，本タスク独自のワークショップも別途日本で開催する予定である．

5 おわりに

多くの研究者の方々の参加を期待したい．参加を検討されている場合，オーガナイザまでご連絡下されば幸いである．タスクのメイリングリストを作成する予定である．また，タスクの詳細については，今後も参加者等との議論により，必要に応じて改訂していきたいと考えているので，タスクについての要望，意見等も是非積極的にオーガナイザまでお寄せいただければ幸いである．なお，今後タスクについての情報は随時詳細が決まり次第，上述したタスクの web ページ上で告知する予定である．

[3] 2010 年の ACL は 7 月 11 日から 16 日まで Sweden の Uppsala で開催される予定である．

参考文献

[1] Eneko Agirre and Oier Lopez de Lacalle. On robustness and domain adaptation using svd for word sense disambiguation. In *Proc. of COLING'08*, 2008.

[2] Yee Seng Chang and Hwee Tou Ng. Estimating class priors in domain adaptation for wsd. In *Proc. of ACL'06*, 2006.

[3] 菊田 篤史, 白井 清昭. 未定義語義の判別を含む語義曖昧性解消. In **言語処理学会第** *12* **回年次大会発表論文集**, pages 636–639, 2006.

[4] 九岡 佑介, 白井 清昭, 中村 誠. 複数の特徴ベクトルのクラスタリングに基づく単語の意味の弁別. In **第** *14* **回言語処理学会年次大会**, pages 572–575, 2008.

[5] 奥村 学, 白井清昭. 現代日本語書き言葉均衡コーパスを用いた意味解析 – 語義の自動特定，新語義の発見 –. **言語**, 37(8):66–73, 2008.

[6] 白井 清昭. Senseval-2 日本語辞書タスク. **自然言語処理**, 10(3):3–24, 2003.

[7] 白井 清昭. コーパスにおける語の意味の自動識別. **国文学 解釈と鑑賞**, 74(1):61–69, 2009.

[8] 田中 博貴, 中村 誠, 白井 清昭. 新語義発見のための用例クラスタと辞書定義文の対応付け. In **第** *15* **回言語処理学会年次大会**, pages P2–31, 2009.