QA レポート

システム情報工学研究科コンピュータサイエンス専攻 1 年
201220629 大西 健志

研究題目：辞書の定義文に関する用例の分布を用いた未知語義の検出

指導教官：山本幹雄・乾孝司

発表日時：2012 年 10 月 25 日

質問 1

なぜ予備実験と提案手法とで対象語を変えたのか。

発表時の回答

予備実験では「キリン」を対象語としたが、発表では分かりやすくするため「ドライバー」を対象語に選んだ。

改善した回答

「キリン」は新聞記事に出現する頻度が多いと思い、予備実験の対象とした。

「ドライバー」は岩波国語辞典に定義されており、それぞれ語義の違いがはっきりしているため例として選んだ。

しかし、確かに実験の対象語と例の語は一致させた方が良く、例に用いる語でも実験をする必要があったと思う。

質問 2

この手法は有用なのか。

発表時の回答

提案手法の実験をしていないので分からないが、有用だと思う。

改善した回答

提案手法の実験をしていないので分からないが、有用だと思う。

現在は語義定義文中の語は全てその語義に関係があると仮定しているが、この手法が有用でなければ別のアプローチを考えなければならない。

質問 3

このような手法の定量的な評価法はどのようなものが一般的なのか。

発表時の回答

田中の手法では精度で評価していたので、本研究も精度で評価したいと思う。

改善した回答

田中[1]は新語義発見の評価として精度（Precision）だけでなく、Recall、F 値を用いていた。また語義曖昧性解消の評価としては正解率を用いていた。

また、Semeval-2[2]では、語義曖昧性解消の評価については Precision、未知語義検出の評価については Accuracy、Precision、Recall を用いていた。本研究では Semeval で用いられたこちらの評価法を用いたほうが妥当かと思う。

質問 4

提案手法で、未知語義の判定はどうするのか。

発表時の回答

全ての語義について同様の分布を作り、判定したい用例がどの語義に対してもその語義である確率が低ければ、その用例を新語義の用例とする。また、このことは発表の中で説明するのを忘れていた。

改善した回答

すべての語義について同様の分布を作り、判定したい用例を特徴空間に配置した時に、どの語義の用例の分布に対しても密度の低いところに配置されればそれを新語義の用例とする。

質問 5

判定に使う「距離」はどうするのか。

発表時の回答

特に考えていない。提案手法を定式化するところで考えていきたい。

改善した回答

田中の手法では特徴ベクトル同士の近さを測るために「類似度」という言葉を用い、コサイン類似度を用いている。本手法ではそもそも距離を用いるかどうかも考えていきたい。

質問 6

語義定義分から単語を取り出して分布を作成しているが、何を基準に単語を取り出しているのか。

発表時の回答

名詞、動詞、形容詞などの自立語を分布の作成に用いようと思う。

改善した回答

自立語は、助詞などとは違い何らかの意味を表していると考えられるため、自立語を用いようと思う。また、田中[1]も自立語を特徴ベクトルの素性として用いていた。

質問 7

予備実験として田中の手法を選んだ理由は何か。また、結局は挙げた手法を全て実装して提案手法との比較を行うのか。

発表時の回答

Semeval-2 において田中の手法が良い結果を出しており、また直感的だと思ったためである。また、関連研究はできるだけ多く実装して比較したい。

改善した回答

上記の回答に「しかし、田中の手法を基にしたシステムよりも良いシステムがある。また、どのシステムも絶対的には良いとはいえないため、どのシステムも伸びしろがあると言える。色々なシステムを実装してそれらの分析も行いたい。」を加える。

参考文献：

[1] 田中博貴, 2009 年.「用例のクラスタリングに基づく単語の新語義の発見」，修士論文，北陸先端科学技術大学院大学情報科学研究科.

[2] Manabu Okumura, Kiyoaki Shirai, Kanako Komiya, and Hikaru Yokono. 2010. Semeval-2010 task:Japanese WSD. In Proceedings of the SemEval-2010: 5th International Workshop on Semantic Evaluations.