単語多義性解消における漸進的なコーパス自動タグ付け

1E－1

鶴岡 慶雅　近山 隆
（東京大学 工学系研究科）

1. はじめに

複数の意味をもつ単語の意味を文脈から判定する単語多義性解消は、数多くの自然言語処理アプリケーションにおいて重要な役割をはたすと考えられている。近年は、計算機で利用可能なテキストの大幅な増加を反映して、コーパスベースの手法が注目を集めている。しかし、コーパスベースの手法は、一般に、意味のタグ付けがなされたコーパスを用意しなければならないが、大きなコーパスに人手で意味のタグ付けをするのは、非常にコストがかかる。そのため、意味のタグ付けがなされたコーパスの質・量ともに不足しているのが現状である。

そこで本稿では、Yarowsky によって提案されたコーパスへの自動タグ付け手法[1]を、形態素解析のみがなされた日本語テキストへの適用を試みる。

2. 決定リストによる多義性解消

複数の意味をもつ単語の意味は、文脈によってどの意味かが判定されるが、ここでは、その文脈というものを、*Evidence* の集まりという形で表現する。*Evidence* にはさまざまなタイプが考えられるが、本稿では、以下のような *Evidence* を用いる。

Wn ：ターゲットから距離 n 単語以内の名詞
Ln ：ターゲットの左側で距離 n 以内の名詞
Rn ：ターゲットの右側で距離 n 以内の名詞
（本稿では n は 7 以内に制限）

たとえば、ターゲットが

「環境」：意味 A. 自然環境
　　意味 B. 周囲の状況

という単語であるとして、

「自然環境を破壊から保護する」

という文からは、

L1~7 自然、W1~7 自然、R2~7 破壊、W2~7 破壊、R4~7 保護、W4~7 保護

という *Evidence* が得られる。

次に、このような *Evidence* を意味のタグ付けがなされたコーパスから大量に抽出し、それぞれの *Evidence* が、どれだけ意味の判定に有用であるかを次の式によって評価する。

$$likelihood = \frac{P(Sense_A|Evidence_i)}{P(Sense_B|Evidence_i)} \qquad (1)$$

この式は、$Evidence_i$ が、ターゲットの単語の意味を意味 A だと判定することのもっともらしさを表しているといえる。

実際に語義の判定をする際には、上記の *Evidence* を、*likelihood* の大きい順にならべて決定リストをつくり、それをターゲットの文脈に適用することによって判定を行う。

3. 漸進的なコーパス自動タグ付け

(1)式は、コーパス中の全文がタグ付けされている状態ならば計算は簡単である。しかし、漸進的にタグ付けを行う場合、タグ付け途中での *likelihood* は、どうやって計算すればよいだろうか？

まず、ベイズの定理より、

$$P(A|B) = \frac{P(A \cdot B)}{P(B)} = \frac{P(A) \cdot P(B|A)}{P(B)}$$

であるから、

$$likelihood = \frac{P(Sense_A) \cdot P(Evidence_i|Sense_A)}{P(Sense_B) \cdot P(Evidence_i|Sense_B)}$$

となる。本稿ではここで、

Automatically Sense-tagging the Corpus for Word Sense Disambiguation
Yoshimasa Tsuruoka and Takashi Chikayama
School of Engineering, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan

$$\frac{P(Sense_A)}{P(Sense_B)} \approx 1$$

および

$$\frac{P(Evidence_i|Sense_A)}{P(Evidence_i|Sense_B)} \approx \frac{P(Evidence_i|Tagged_A)}{P(Evidence_i|Tagged_B)}$$

の仮定をおく。そうすると、

$$likelihood \approx \frac{P(Evidence_i|Tagged_A)}{P(Evidence_i|Tagged_B)}$$

となり、タグ付けの途中でも計算が可能となる。

次に全体の流れを以下に示す。

STEP1

決定リストを用いて、*likelihood* が閾値以上の文をタグ付けする。

STEP2

タグ付けされた文から、*Evidence* を抽出し、新たな決定リストを作成する。

STEP3

STEP1 に戻り繰り返す。ただし、新たにタグ付けされた文がない場合は、閾値を下げて STEP1 に戻る。 閾値がある値以下になったら終了。

4. 評価

疑似多義語「政策/テレビ」を用いて評価を行った。疑似多義語とは、タグ付けをするシステムの側からは、「政策」と「テレビ」が同一の単語にしか見えないようにしておき、文脈から、どちらの単語であるかを決定させようとするものである。この方法をもちいることで、タグ付けの正誤の判定を人手を借りずに行うことができる。

コーパスとして「CD-毎日新聞 97 版」から「政策/テレビ」を含む 2000 文を抽出し、それを形態素解析システム「茶筌」で形態素解析したものをタグ付けシステムへの入力とした。

最初の種となる *Evidence* としては、それぞれ次のようなものを与えた。

「政策」 ： L1 金融

「テレビ」： R1 番組

表 1 実験結果

	出現回数	タグ付け数	正解	正解率
政策	953	499	463	92.8%
テレビ	1047	578	552	95.5%

表 2 最終的な決定リスト

Likelihood	Type	単語	語義
474.3	W7	朝日	テレビ
281.6	W7	記者	テレビ
228.2	W7	系	テレビ
224.9	W7	経済	政策
187.4	W7	金融	政策
186.8	W7	番組	テレビ
174.9	W7	放送	テレビ
148.2	W7	朝	テレビ
146.2	W5	基本	政策
146.2	W7	演説	政策
…	…	…	…

結果を表 1 に示す。「政策」「テレビ」ともに高い正解率が得られている。表 2 に、タグ付け終了時の決定リストを示す。繰り返しの結果として、最初の種として与えた *Evidence*「金融」「番組」よりも有用な *Evidence* が得られていることがわかる。

5. まとめ

本稿では、Yarowsky によって提案されたコーパスへの自動タグ付け手法[1]を日本語テキストへ適用し、その有効性を示した。今後、シソーラスの利用や擬似的な構文解析などと組み合わせることによって、適用範囲・正解率の向上を行いたいと考えている。

謝辞

形態素解析システムとして、奈良先端科学技術大学院大学松本研究室の「茶筌」を使用させていただきました。ここに謝意を示します。

参考文献

[1] David Yarowsky : Unsupervised Word Sense Disambiguation Rivaling Supervised Methods, In Proceedings of the 33rd Annual Meeting of the ACL. Cambridge, MA, pp. 189-196 (1995)