コーパスに基づく単語多義性解消システムの評価法について*

藤井 敦　乾 健太郎　徳永 健伸　田中 穂積
東京工業大学大学院情報理工学研究科
{fujii,inui,take,tanaka}@cs.titech.ac.jp

1 はじめに

近年、コーパスに基づく単語多義性解消 (word sense disambiguation : WSD) が盛んに行われている。著者らは、その中のひとつである事例に基づく手法の提案及びシステムの開発を行ってきた [3, 4]。本論文は多義性解消システムを評価するための新しい枠組を提案し、提案した評価法に基づいて著者らのシステムの評価を行う。

多義性解消システムは、与えられた入力文中の多義語について、既存の辞書に分類された語義候補の中から最も適当と考えられる語義を選択する[1]。すなわちシステムのタスクは、いわゆるマッチングであり、その性能は正しく語義を特定できた出力の割合によって評価される。しかし、本論文では以下の問題を提起したい。語義分類は辞書によって分類の「視点 (viewpoint)」や「粒度 (granularity)」が異なるため、システムの評価は使用する辞書に強く依存する。これは形態素解析や統語解析の評価と大きく異なる点である。多義動詞「使う」を例に、この問題を検討してみよう。EDR 単語辞書 [21] は、「使う」を「人を使う」「材料を使う」「道具を使う」「消費する」など計7種類の語義に分類している。これらの語義は EDR シソーラスのノードにリンクされている。図1に「使う」の語義を含むシソーラスの一部を示す。図1から分かるように、「使う」の語義間に類似関係が存在し、アプリケーション (機械翻訳など) によっては「材料を使う」と「消費する」を必ずしも明確に区別する必要はないこともある。このような状況では、これら2つの語義は1つの語義 (例えば「使い果たす」) として見なすことが可能である。さらに「人を使う」以外の3つの語義は、例えば「物を使う」としてまとめることができるかもしれない。すなわち多義性解消システムの性能評価は、それが適用される個々のアプリケーションでの語義分類に基づいて行わなければならない。しかし他方において、システムを評価するためには一定量のテストデータが必要であり、異なる語義分類ごとに網羅的にテストデータを作成することは困難である。そこで本論文では、一つに固定した語義分類に基づく、より一般的な性能評価法を提案する。今、異なる多義性解消システム A と B の性能を比較する場合を考える。正解が「消費する」である入力「使う」に対して、A と B がそれぞれ「材料を使う」と「人を使う」を出力したとしよう。従来の二値評価 (正解/不正解) ではどちらの出力も不正解である。しかし、A の出力は B のそれに比べ正解により近く、その誤りはより許容できるものである。すなわち B よりも A に高い評価を与えるべきである。本論文では、誤りに対する「許容度 (acceptability)」の概念を導入し、許容度に応じて得点を与える評価法を提案する。システム評価のもう一つの重要な点は、上の例のように他の手法との比較を行うことである。本論文の評価実験では、著者らの事例に基づく手法を統計的手法 [16] と比較する。

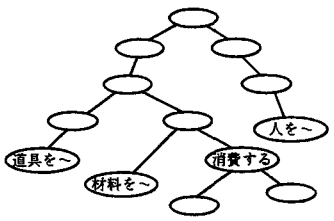

図 1: 動詞「使う」の語義を含む EDR シソーラスの一部

2 動詞多義性解消システム

2.1 事例に基づく手法

本システムの構成を図2に示す。本システムのタスクは入力に現れる動詞の多義性解消である。入力が与えられると、まず QJP [8] によって形態素/統語解析を行う。次に、QJP の解析結果から多義動詞とそれに係る格要素の構造 (V-C) を抽出して WSD に渡す。本手法は事例に基づく手法 [9, 11, 14] であり、システムにはデータベースすなわち動詞語義を付与した事例セットが与えら

*Evaluation of Word Sense Disambiguation Systems
[1]語義を自動的に分類するタスク [5, 13, 15, 18] とは異なる。

れている。WSD は入力と事例の間の類似度に基づいて動詞語義ごとにスコアを計算し、スコアを最大化する動詞語義を出力する。

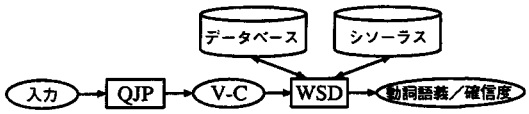

図 2: システム構成

スコアは、入力と事例に含まれる格要素間の類似度の総和によって計算される。動詞「使う」のデータベースの例を図 3 に示す。図 3 では、動詞「使う」の語義のうち「人を使う」「道具を使う」「消費する」の格要素の事例 (「計算機を」など) が記述されている。今、例として「助手が コピー機を 使う」という入力 (V-C) が与えられると、ガ格の「学生」と「助手」、ヲ格の「計算機」と「コピー機」の間の類似度が高いことから「道具を使う」を入力の「使う」の語義として出力する。格要素間の類似度は分類語彙表の名詞シソーラス [20] における二つの格要素間のパスの長さに基づいて計算される [9]。本システムではさらに、ガ格やヲ格などの格要素の動詞多義性解消における貢献度 (case contribution to disambiguation : CCD) を定量化し、貢献度が高い格要素間の類似度を優先的にスコアに反映させる [4]。図 3 のデータベースでは、ニ格やヲ格のように動詞語義ごとに事例が顕著に異なる格ほど多義性解消における貢献度が高くなる。語義候補 s のスコアは、動詞語義 s の格 c における事例と入力の格要素間の類似度をその格の貢献度 $CCD(c)$ で重み付けした加重和によって計算される (式 (1))。

$$Score(s) = \sum_{c} SIM(n_c, \mathcal{E}_{s,c}) \cdot CCD(c) \qquad (1)$$

$SIM(n_c, \mathcal{E}_{s,c})$ は格 c における入力中の格要素 n_c と動詞語義 s の格要素の事例セット $\mathcal{E}_{s,c}$ との間の類似度を表す。

さらに、システムは多義性解消の確信度を定量化して出力する。確信度は、1 位の (選択した) 動詞語義のスコアが大きく、1 位と 2 位の動詞語義のスコアの差が大きいほど高くなる [3]。多義性解消の確信度 C を式 (2) に示す。ここで $Score_1$ と $Score_2$ はそれぞれ 1 位と 2 位の動詞語義のスコアを表す。λ はパラメータであり、本論文の実験では $\lambda = 0.5$ とした。

$$C = \lambda \cdot Score_1 + (1 - \lambda) \cdot (Score_1 - Score_2) \qquad (2)$$

2.2 比較対象としての統計的手法

本論文の実験では、Yarowsky によって提案された統計的な多義性解消法を比較対象とした [16]。本手法は、

{彼, 企業}	が	{企画}	に	{従業員, 卒業生}	を	使う (人を〜)
{彼女, 学生}	が	{仕事, 研究}	に	{計算機, 機械}	を	使う (道具を〜)
{彼, 政府}	が	{車, 福祉}	に	{金, 燃料, 税金}	を	使う (消費する)

図 3: 「使う」に関する事例を含むデータベースの一部

動詞語義の前後に共起する単語 (文脈) の頻度情報をコーパスから抽出して統計ベースとして利用する[2]。スコアの計算はベイズ則に基づいており、動詞語義 s のスコアは入力に現れる共起語 w が与えられた場合の条件付き確率 $P(s|w)$ の総和として計算される (式 (3))。

$$Score(s) = \sum_{w \ in \ input} P(s|w) \qquad (3)$$

データスパースネスを避けるため、共起語は分類語彙表を用いてあらかじめ意味クラスに抽象化する。

3 許容度を用いた評価

3.1 許容度の計算

正解の動詞語義に (意味的に) 近い出力ほど許容度は高くなる。正解と出力の間の意味的距離は動詞シソーラスにおけるパスの長さによって計算し、パスが短いほど意味的に近くなる。正解 s に対して動詞語義 x を出力した場合の許容度 $A(x,s)$ を式 (4) に示す。ここで $LEN(x,s)$ はシソーラスにおける二つの動詞語義 x と s の間のパスの長さであり、$MAXLEN$ は (与えられた動詞に関して) 最も離れた動詞語義間のパスの長さである。図 1 の例では $MAXLEN = 7$ である。α はパラメータであり、この値を大きくするほど誤りに対する許容度が低くなり、二値評価に近くなる。

$$A(x,s) = \left(\frac{MAXLEN - LEN(x,s)}{MAXLEN} \right)^{\alpha} \qquad (4)$$

$A(x,s)$ は、0 (x と s が最も遠い) から 1 (x と s が一致) までの値を取る。

3.2 比較実験

本実験では EDR の語義分類及びシソーラスを用いて、(1) 事例に基づく手法、(2) Yarowsky の統計的手法、(3) 最頻出の動詞語義を常に選択する手法 (システムに要求される性能の下限 [6]) を比較した[3]。手法 (1) において、格要素の省略によって多義性解消が行えない場合は手法 (3) を用いる。EDR 日本語コーパス [21] から、比較的出

[2]本論文の実験では助詞などの機能語は共起語から除外した。

[3]許容度は、EDR 以外のシソーラス (WordNet [12] など) にも容易に適用できる。

現頻度の高い動詞を含む例文10880文を抽出して実験に使用した[4]。それぞれの動詞について例文を訓練用/テスト用のセットに分割し、10-fold クロスバリデーションによって評価を行った。表1に、いくつかのαのもとでの手法 (1)〜(3) の許容度を示す。precision はシステムの出力と正解動詞語義が完全に一致した割合であり、二値評価に相当する。動詞「書く/描く」以外は、手法 (1) が他の手法よりも許容度が高いことが分かる。

手法 (1) と (2) をさらに比較してみよう。コーパスに基づく手法では、訓練データのサイズと許容度との関係が重要である。図4に訓練データのサイズと許容度 ($\alpha = 0.5$) の関係を示す[5]。図4から、どちらの手法も訓練データを増やすことで許容度が向上するものの、手法 (1) の方が手法 (2) よりも全般的に許容度が高いことが分かる。次に、適応性 (applicability) と許容度との関係を調べる。適応性とは、確信度が閾値以上の (自信のある) 出力が全体に占める割合である。言い替えると、自信のある出力の許容度が高いシステムほど性能が良いという観点からの評価である。結果を図5に示す。確信度の閾値を上げる (自信のあるものしか出力しない) ことによって、適応性が下がり許容度が上がることが分かる。同じ適応性における許容度は手法 (1) の方が高いことが分かる。最後に、スコアが高い上位 k個の動詞語義を出力した場合の recall/precision のトレードオフによってシステムの性能を比較してみよう。recall は k個の出力の中に正解と完全一致する動詞語義を含む割合である (本比較では許容度は用いない)。kの値を変化させながら評価した結果を図6に示す。手法 (1) のグラフが手法 (2) よりも右上に位置しており、より良い結果であることを示している[6]。

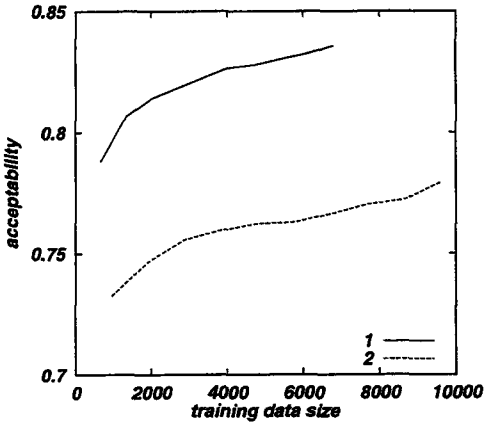

図 4: 訓練データサイズと許容度 ($\alpha = 0.5$)

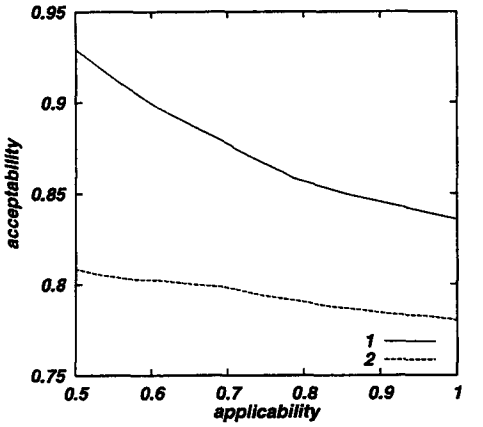

図 5: 適応性 (applicability) と許容度 ($\alpha = 0.5$)

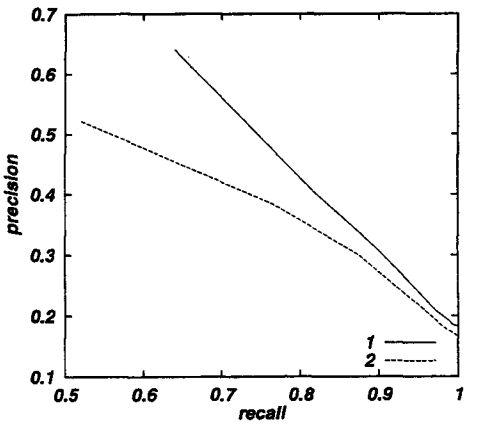

図 6: Recall と Precision

4 おわりに

本論文は、誤りに対する許容度を用いた多義性解消システムの評価法を提案し、著者らが開発した事例に基づく手法を統計的手法の代表的なものと比較する評価実験を行った。(a) 訓練データと許容度、(b) 適応性と許容度、(c) recall と precision、という観点から評価した結果、著者らの手法は比較対象よりも優れていることが確認された[7]。その他の多義性解消法 [1, 7, 19] との比較は今後の課題である。

最後に、多義性解消システム評価におけるその他の重

[4]EDR コーパスでは、各単語に EDR 辞書に基づいた語義が付与されており、本実験では動詞語義の情報を使用した。
[5]格要素省略のため、手法 (1) と (2) は訓練データ総数が異なる。
[6]図5、図6の実験では訓練データを全てシステムに与えている。
[7]図4、図5の実験では$\alpha = 0.5$以外でも同様の傾向が確認された。

表 1: 各手法の性能比較 (動詞ごとの許容度)

動詞	例文数	EDR 語義数	$\alpha = 0.5$			$\alpha = 1$			$\alpha = 2$			precision ($\alpha = \infty$)		
			(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
使う	2117	7	.908	.894	.877	.843	.819	.787	.757	.718	.665	.583	.506	.394
受ける	1709	10	.912	.828	.733	.864	.736	.594	.824	.658	.482	.796	.609	.416
持つ	1636	12	.810	.686	.653	.740	.575	.512	.697	.503	.416	.677	.471	.373
見る	1476	17	.837	.798	.791	.732	.667	.655	.612	.516	.496	.437	.291	.411
求める	1123	5	.773	.722	.699	.721	.654	.610	.700	.626	.578	.698	.621	.571
出す	967	5	.740	.633	.579	.687	.568	.515	.660	.538	.486	.647	.529	.482
加える	516	4	.781	.772	.753	.728	.717	.696	.695	.684	.662	.671	.659	.636
書く/描く	503	2	.712	.728	.577	.712	.728	.577	.712	.728	.577	.712	.728	.577
送る	434	9	.876	.829	.809	.789	.728	.676	.685	.577	.518	.548	.424	.359
設ける	399	3	.818	.815	.805	.742	.728	.707	.686	.663	.633	.667	.642	.609
合計	10880	—	.836	.779	.738	.772	.693	.634	.714	.615	.540	.640	.521	.449

要な要素について述べる。まず、タスクの上限の推定が挙げられる。一つの方法として、人間の多義性解消の正解率を用いた推定法が提案されている [6]。また、近年のコーパスに基づく手法の多くは、訓練用コーパスへの語義付与やコーパス検索のコストが大きいといった問題を抱えている。多義性解消の精度を落とすことなく、これらのコストをどの程度軽減できるかといった観点から評価を行うことも考えられる [2, 3, 10, 17]。

謝辞

QJP の使用に関して多大な援助を頂きましたリコーの亀田雅之氏に感謝致します。

参考文献

[1] Eugene Charniak. *Statistical Language Learning*. The MIT Press, 1993.

[2] Sean P. Engelson and Ido Dagan. Minimizing manual annotation cost in supervised training from corpora. In *Proceedings of ACL*, pp. 319–326, 1996.

[3] Atsushi Fujii, Kentaro Inui, Takenobu Tokunaga, and Hozumi Tanaka. Selective sampling of effective example sentence sets for word sense disambiguation. In *Proceedings of the Fourth Workshop on Very Large Corpora*, pp. 56–69, 1996. http://xxx.lanl.gov/ps/cmp-lg/9702010.

[4] Atsushi Fujii, Kentaro Inui, Takenobu Tokunaga, and Hozumi Tanaka. To what extent does case contribute to verb sense disambiguation? In *Proceedings of COLING*, pp. 59–64, 1996.

[5] Fumiyo Fukumoto and Jun'ichi Tsujii. Automatic recognition of verbal polysemy. In *Proceedings of COLING*, pp. 764–768, 1994.

[6] William Gale, Kenneth Ward Church, and David Yarowsky. Estimating upper and lower bounds on the performance of word-sense disambiguation programs. In *Proceedings of ACL*, pp. 249–256, 1992.

[7] Graeme Hirst. *Semantic Interpretation and the Resolution of Ambiguity*. Cambridge University Press, 1987.

[8] Masayuki Kameda. A portable & quick Japanese parser : QJP. In *Proceedings of COLING*, pp. 616–621, 1996.

[9] Sadao Kurohashi and Makoto Nagao. A method of case structure analysis for Japanese sentences based on examples in case frame dictionary. *IEICE TRANSACTIONS on Information and Systems*, Vol. E77-D, No. 2, pp. 227–239, 1994.

[10] David D. Lewis and William A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of ACM SIGIR*, pp. 3–12, 1994.

[11] Xiaobin Li, Stan Szpakowicz, and Stan Matwin. A WordNet-based algorithm for word sense disambiguation. In *Proceedings of IJCAI*, pp. 1368–1374, 1995.

[12] George A. Miller, et al. Five papers on WordNet. Technical report, Cognitive Science Laboratory, Princeton University, 1993.

[13] James Pustejovsky and Branimir Boguraev. Lexical knowledge representation and natural language processing. *Artificial Intelligence*, Vol. 63, No. 1–2, pp. 193–223, 1993.

[14] Naohiko Uramoto. Example-based word-sense disambiguation. *IEICE TRANSACTIONS on Information and Systems*, Vol. E77-D, No. 2, pp. 240–246, 1994.

[15] Takehito Utsuro. Sense classification of verbal polysemy based-on bilingual class/class association. In *Proceedings of COLING*, pp. 968–973, 1996.

[16] David Yarowsky. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of COLING*, pp. 454–460, 1992.

[17] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of ACL*, pp. 189–196, 1995.

[18] Uri Zernik. Lexicon acquisition: Learning from corpus by capitalizing on lexical categories. In *Proceedings of IJCAI*, pp. 1556–1562, 1989.

[19] 奥村学. 自然言語の意味的曖昧性の解消法. 人工知能学会学会誌, Vol. 10, No. 3, pp. 332–339, 1995.

[20] 国立国語研究所. 分類語彙表, 増補版, 1996.

[21] 日本電子化辞書研究所. EDR 電子化辞書仕様説明書, 1995.