語の大域的多義性解消に基づく省略解析の精度向上

河原 大輔　　黒橋 禎夫
東京大学 大学院情報理工学系研究科
{kawahara, kuro}@kc.t.u-tokyo.ac.jp

1　はじめに

長い間、言語解析研究の中心であった構文解析は、今や精度が 90%を越え、様々な応用研究の基礎技術として使うことができるレベルになった。それに伴い、構文解析の次のステップに位置づけられる省略・照応解析が最近活発に研究が行われるようになった [1, 2, 4, 10]。しかし、その精度はまだ満足の行くものではなく、高精度な省略・照応解析が要求される質問応答・自動要約・機械翻訳などの言語処理アプリケーションで実用的に使えるレベルにはなっていない。我々の省略・照応解析システム [4] の誤りの原因を調査したところ、語義の曖昧性の問題がひとつの原因になっていることがわかった。

我々の省略・照応解析システムは、用例のマッチングを行うために、汎用的なシソーラスである日本語語彙大系 [8] を用いている。日本語語彙大系にはひとつの語に多数の意味属性が定義されており、この多義性を解消しないと不適切なマッチングが生じ、省略・照応解析の精度を悪化させてしまう。例えば、日本語語彙大系において、「ごぼう」は<作物>, <野菜>, <僧侶>, <寺院>*という意味属性をもつ。料理の話をしているときに「ごぼう」に対して<僧侶>の意味属性でマッチングをとると、「ごぼう」が主格として解釈される可能性があり精度悪化の原因となる。

語義の曖昧性の問題は、言語処理における基本的な問題のひとつであり、近年、SENSEVAL[6, 7] のような語義曖昧性解消のコンテストなども行われているが、省略・照応解析のような高度な解析に統合的に用いられることはほとんどなかった。本論文では、格フレームに基づいて名詞と用言 (動詞、形容詞、名詞+判定詞) の多義性解消を行いつつ、省略・照応解析を行う手法を提案する。名詞の多義性解消は、名詞のもつ複数の意味属性の中から文脈において適切なものを選択する処理とする。同様に、用言の多義性解消は、用言のもつ複数の格フレームの中から適切なものを選択する処理とする。さらに、多義性を解消した結果を同文章の解析中は保持しておき、それ以降の解析で利用する。

*本論文では、意味属性を<>で表す。

2　省略・照応解析システムの概要

我々は、格フレーム辞書 [3]、先行詞選好順序および機械学習に基づく省略・照応解析システムを提案している [4]。その解析の手順を以下に示す。

1. 入力文を構文解析する。
2. 入力文中の各用言について、文頭から順番に以下の処理を行う。
   (a) 入力用言の述語項構造に基づいて格フレームを絞り込む。
   (b) 絞り込んだ結果の格フレームそれぞれについて以下の処理を行う。
      i. 格フレームと入力側の格要素との対応をとり、対応づけられていない格をゼロ代名詞と認識する。
      ii. 照応詞の先行詞を同定する。
   (c) もっともスコアが高い格フレームに決定し、その格フレームを用いたときの解析結果を出力する。

以下では、(2a) と (2b) の各処理について詳説する。

2.1　格フレームの絞り込み

用言の用法の決定に対して、用言の直前格要素が重要な役割を果たす。特に、直前格がヲ格、ニ格の場合はその傾向が強い。また、直前格要素が<主体>の場

表 1: 「擁立」の格フレームの例

	格	用例
擁立 (1)	ガ	<主体>, 派, 政党, ···
	ヲ	<主体>, 候補, 候補者
	ニ	<主体>, 選挙区, 選, ···
擁立 (2)	ガ	<主体>
	ヲ	<主体>, 議員, 外相, ···
	ニ	<主体>, 候補, 後継, ···
⋮	⋮	⋮

合、例えば「<主体>が　求める」という表現からは、用言の用法が決まらず、格フレームを選択することができない。これらの点を考慮して、格フレームを絞り込む条件を以下のように設定する。

1. 入力側の対象用言が直前格要素 C をもつ。
2. 直前格要素 C と直前格 cm が以下のいずれかの条件を満たす。
   - cm がヲ格、ニ格のいずれかである。
   - cm がヲ格、ニ格以外で、C が意味属性<主体>をもたない。
3. cm をもつ格フレームが存在し、cm の用例群と C の類似度が閾値以上ある。

条件 3 を満たす格フレームのなかで、もっとも類似度が高い格フレームのみに絞り込む。条件を満たさない場合は絞り込みを行わず、対象用言のもつすべての格フレームについて後続の処理を行う。ここで用いる類似度は、直前格要素と格フレームの直前格の各用例との類似度のうちもっとも高いものとする。用例間の類似度は最大 1.0 とし、日本語語彙大系を用いて計算する [9]。

例として、図 1 の「擁立する」を考える。「擁立」に対して表 1 のような格フレームがある。入力側の表現「候補を　擁立する」は上記の条件 1, 2 を満たし、格フレーム「擁立 (1)」が条件 3 を満たすので、格フレーム「擁立 (1)」が選択される。

2.2　入力側格要素と格フレームの格との対応づけ

選択された格フレームについて、入力側の格要素と格フレームの格との対応づけを行う。格要素に格助詞が付属している場合は、その格助詞の格に対応する格フレーム側の格に対応づける。被連体修飾詞や係助詞

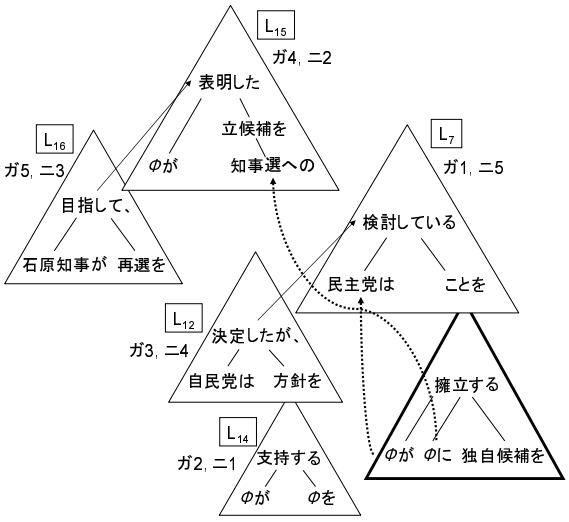

図 1: 省略・照応解析例

句のように、文中から格がわからない場合は、次表の格それぞれに対応させ、対応づけ全体のスコアがもっともよい対応を選択し、格を決定する。

係助詞句	：ガ, ヲ, ガ 2 †
被連体修飾詞	：ガ, ヲ, 外の関係

対応づけ全体のスコアは、各格の対応の類似度を足したものとする。各格の対応の類似度は「格フレームの絞り込み」で述べた類似度と同様である。

対応づけが終わったときに、格フレームに入力文の格要素と対応づけられていない格があり、それがガ格、ヲ格、ニ格のいずれかであれば、ゼロ代名詞であると認識する。図 1 の「候補を　擁立する」では、格フレーム「擁立 (1)」が選択されており、格フレームのガ格とニ格に対応する入力側の格要素がないことがわかる。従って、システムはガ格とニ格をゼロ代名詞として検出する。

2.3　先行詞の同定

前節で認識したゼロ代名詞および指示詞について先行詞の同定を行う。本手法は、格ごとに設定された先行詞選好順序 [4] に従って先行詞候補を調べ、格フレームの用例との類似度が閾値を越え、かつ、様々な素性を考慮した分類器によって正例と分類される候補を先行詞に決定する。

図 1 の「擁立」では、ガ格とニ格がゼロ代名詞として認識されている。例えば、ガ格の候補は、先行詞選

† 二重主語構文の外のガ格を「ガ 2」格と呼ぶ。

好順に、L_7:民主党, L_{14}:自民党 (ϕ ガ), L_{14}:石原知事 (ϕ ヲ), $\cdots$ となっている。1 番目の「民主党 (類似度: 0.73)」は分類器によって正例と分類され、格フレームの用例との類似度が 0.73 と閾値 (0.60) を越えているので先行詞に決定される。

3 語の大域的多義性解消

本研究では、上記の省略・照応解析システムに語の多義性解消処理を統合する。多義性解消は格フレームに基づいて用言と名詞に対して行う。さらに、1 文章内では、用言や名詞が複数の意味で使われることはほとんどない (one sense per discourse) と考え、多義性解消結果を保持しておくことによって後の解析で利用するという大域的手法をとる。以下では、料理教示発話のテキストを例とし、用言と名詞の大域的多義性解消手法についてそれぞれ説明する。

3.1 用言の多義性解消

我々の利用している格フレームは、用言とその直前の格要素を組にしているため、用言の意味ごとに詳細に分かれたものである [3]。そこで、本研究では、用言の多義性解消を、格フレームを選択する処理と捉える。この処理は、2.1 節の手順 (2c) に該当する。これに加えて本研究では、用言の多義性解消結果を保持し大域的に用いる。つまり、用言ごとにどの格フレームを選択したかを記憶しておき、同文章中に同じ用言が再び出現した場合には、記憶している格フレームと類似しているもののみ (類似度閾値:0.60) を用いることにする。この格フレーム間の類似度としては、格フレーム自動構築時のクラスタリングで用いた尺度と同じ く、格の一致度と用例集合間の類似度の積 [9] とする。以下に例を示す。

1 文目　おろし金でかぶらを おろしていきます。
2 文目　これは大きい「聖護院かぶら」です。
3 文目　このように「の」の字を描いて おろします。

この例の 1 文目では、「{ 大根 } を おろす」という格フレームが選択されるので、「おろす」に対して「{ 大根 } を おろす」という格フレームが使用されたことを記憶しておく。3 文目の「おろす」では、従来手法では「おろす」のすべての格フレームに対して先行詞同定までの処理を行っていたが、本手法では記憶している「{ 大根 } を おろす」に類似した格フレームのみを使用することになり、ヲ格の先行詞として正解の「かぶら」を補う可能性が高くなる。逆に、「{ 銀行, 郵便局, ... } から { 金 } を おろす」というような類似していない格フレームは使われず、「金」と似ている「おろし金」をヲ格の先行詞として誤る可能性が低くなる。

3.2 名詞の多義性解消

名詞の多義性解消は、日本語語彙大系において解析対象名詞に対して定義されている意味属性集合から、文脈において適切なものを選ぶ処理となる。意味属性の選択は、手順 (2c) において決定した格フレームと入力の述語項構造とのマッチング結果に基づいて行う。つまり、それぞれの格スロットごとに、入力側格要素の各意味属性と格フレームの用例群とのマッチングをとり、もっともマッチする意味属性を選択する。そして、語ごとに決定した意味属性を記憶しておき、同文章中に同じ名詞が再び出現した場合には、記憶している意味属性のみを用いる。以下に例を示す。

1 文目　まずは おすまし を頂きます。
2 文目　２０分で本格的なだしを取りました。

1 文目の「おすましを 頂く」に対して選択される格フレームは以下のようなものである。

	格	用例	入力
頂く	ガ	<主体>	-
	ヲ*	スープ	おすまし

「おすまし」に対する意味属性は<汁>,<顔つき>,<奇人>の 3 つであるが、これらと格フレームの用例である「スープ」とのマッチングをとった結果、<汁>がもっとも類似度が高いので、「おすまし」の意味属性は<汁>に決定される。これによって、「おすまし」には<奇人>という主体の意味がなくなり、2 文目以降の動詞のガ格先行詞になって誤る可能性もなくなる。

4 実験

我々の省略・照応解析システムは、料理番組の料理教示発話を構造化する研究 [11] において利用されている。そこでは、料理番組のクローズドキャプション、5 番組、813 文に対して、関係コーパス [5] の基準に従って、様々な関係の正解タグを付与している。実験は、この料理コーパスを用いて 5-fold 交差検定を行った。先行詞選好順序、分類器は料理コーパスから学習し、格フレーム辞書は料理ドメインのテキスト約 500 万文を Web から収集して作成した。評価する格は、省略

表 2: 省略解析精度 (料理)

	適合率	再現率	F 値
旧手法	696/1092 (0.637)	696/1482 (0.470)	0.541
本手法	713/1081 (0.660)	713/1482 (0.481)	0.556

表 3: 省略解析精度 (新聞)

	適合率	再現率	F 値
旧手法	515/924 (0.557)	515/1087 (0.474)	0.512
本手法	526/911 (0.577)	526/1087 (0.484)	0.527

の大多数を占めるガ格、ヲ格、ニ格とし、ゼロ代名詞検出と先行詞同定を併せて評価を行った。その結果を表 2 に示す。

表 2 より、大局的多義性解消によって精度が向上したことがわかる。しかし、この精度向上は名詞の多義性解消によるものがほとんどであり、用言の多義性解消では精度はほとんど向上しなかった。これは、用言の大域的多義性解消を用いなくても、正しい格フレームがおおむね選ばれており、精度向上にはつながらなかったと思われる。なお、大域的多義性解消が行われた用言は、調べる格フレームが 16%に減少しており、解析効率のかなりの向上がみられた。

また、新聞記事に対しても評価実験を行った。先行詞選好順序、分類器は関係コーパス [5] の 279 記事からを学習し、格フレーム辞書は新聞記事 20 年分から学習した。関係コーパス 100 記事を用いてテストを行った結果を表 3 に示す。料理の場合と同様に、大局的多義性解消による精度向上がみられる。

5 おわりに

本研究では、省略・照応解析システムに語の大域的多義性解消の枠組を導入した。多義性解消は、用言に対しては格フレームの選択、名詞に対しては意味属性の選択によって行った。さらに、その多義性解消結果を同文章の解析中は保持しておき、後の解析で利用するという大域的な手法を提案した。今後は多義性解消自身の評価をしつつ誤りを分析し、省略・照応解析システムの精度向上を目指す予定である。

参考文献

[1] Ryu Iida, Kentaro Inui, Hiroya Takamura, and Yuji Matsumoto. Incorporating contextual cues in trainable models for coreference resolution. In *Proceedings of the 10th EACL Workshop on The Computational Treatment of Anaphora*, pp. 23–30, 2003.

[2] Hideki Isozaki and Tsutomu Hirao. Japanese zero pronoun resolution based on ranking rules and machine learning. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pp. 184–191, 2003.

[3] Daisuke Kawahara and Sadao Kurohashi. Fertilization of case frame dictionary for robust Japanese case analysis. In *Proceedings of the 19th International Conference on Computational Linguistics*, pp. 425–431, 2002.

[4] Daisuke Kawahara and Sadao Kurohashi. Zero pronoun resolution based on automatically constructed case frames and structural preference of antecedents. In *Proceedings of the 1st International Joint Conference on Natural Language Processing*, 2004, (to appear).

[5] Daisuke Kawahara, Sadao Kurohashi, and Kôiti Hasida. Construction of a Japanese relevance-tagged corpus. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pp. 2008–2013, 2002.

[6] Adam Kilgarriff and Martha Palmer. Introduction to the special issue on SENSEVAL. *Computers and the Humanities*, Vol. 34, No. 1, pp. 1–13, 2000.

[7] David Yarowsky, editor. *SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguating Systems*. The Association for Computational Linguistics, 2001.

[8] NTT コミュニケーション科学研究所. 日本語語彙大系. 岩波書店, 1997.

[9] 河原大輔, 黒橋禎夫. 用言と直前の格要素の組を単位とする格フレームの自動構築. 自然言語処理, Vol. 9, No. 1, pp. 3–19, 2002.

[10] 関和広, 藤井敦, 石川徹也. 確率モデルを用いた日本語ゼロ代名詞の照応解析. 自然言語処理, Vol. 9, No. 3, pp. 63–85, 2002.

[11] 柴田知秀, 立木将人, 河原大輔, 岡本雅史, 黒橋禎夫, 西田豊明. 言語情報と映像情報の統合による教示発話の構造解析. 言語処理学会 第 10 回年次大会発表論文集, 2004.