Wikipediaを用いた段階的語釈文拡張手法による語義曖昧性の解消

大橋 勝揮　　小林 暁雄　　増山 繁
豊橋技術科学大学 情報・知能工学専攻
{ohashi@la.cs, a-kobayashi@cs}.tut.ac.jp, masuyama@tut.jp

1　はじめに

語には複数の意味（語義）を持つもの（多義語）が存在する．多義語の語義は文脈に依存して変わるため，情報検索や機械翻訳など，意味情報を用いる自然言語処理において，正しい語義を特定する語義曖昧性解消（Word Sense Disambiguation: WSD）は重要な基盤技術の一つに挙げられる．WSD の研究は古くから行われており，それらは辞書やシソーラスなどの知識ベースに基づく手法と，大量のコーパスから語義情報を統計的に学習する手法の 2 つに大別される．

前者として，近年ではフリーのオンライン百科事典である Wikipedia を用いた WSD の研究も盛んに行われている [2, 8, 7]．Wikipedia は誰でも記事の作成・編集が可能であることから，日々盛んな更新が行われており，その内容は多岐にわたる．Wikipedia には特殊な用途の記事がいくつか存在し，**曖昧さ回避ページ**もその一つである．曖昧さ回避ページは，Wikipedia 利用者が目的とする語義の記事を探しやすくするために用意された記事である．曖昧さ回避ページはその用途から，一般の辞書のように語義を端的な文章で表現しており，かつ，その文章中に，対応する語義についての詳細が書かれた記事（**語義解説記事**）へのリンクを持つことが多い．本稿では，このような Wikipedia における曖昧さ回避ページの特徴を利用し，段階的に拡張される語釈文を用いた語義推定手法を提案する．

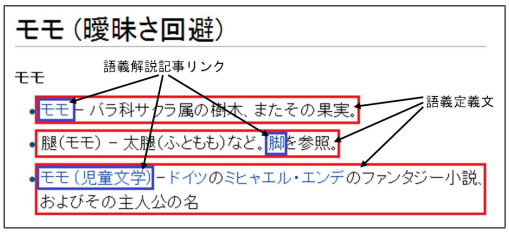

図 1: 語義定義文と語義解説記事リンク

2　関連研究

Cucerzan[2] は，語義解説記事を取得するための 1 つの手段として曖昧さ回避ページを用いている．まず，曖昧さ回避ページの他にリダイレクトやパイプ付きリンクを手がかりに，多義語から語義解説記事への対応付けを行う．多義語の語義は，語義解説記事の本文やそれに含まれるリンク，記事が持つカテゴリに加え，記事自体がどのような一覧記事からリンクされているかという情報を用いて推定を行う．しかしながら，語義によっては曖昧さ回避ページに記述はあるものの，語義解説記事は存在しないものも存在する．Cucerzan の手法では語義推定に語義解説記事を用いることから，語義解説記事が存在しない語義は推定できないという問題がある．

黒川ら [8] は，WSD に用いる知識を取得するために曖昧さ回避ページを利用している．曖昧さ回避ページに記述された語義の定義文（**語義定義文**）に含まれる名詞をキーワードとして WEB 検索を行い，その検索結果を用いて語義を推定する．しかしながら，語義によっては語義定義文が非常に短いものも存在する．その場合，検索結果が他の語義のものと類似する可能性があり，語義を正しく推定できないことがある．

3　提案手法

3.1　概要

提案手法は，WSD に用いる知識である語釈文として，曖昧さ回避ページに記述された語義定義文，その文中に含まれることがある語義解説記事リンクの有無，リンク先である語義解説記事を用いる．ここで，本稿では，”語義を説明する文章，その語義の属性や使われやすさなど，語義を特徴付ける要素”の総称を**語釈文**と呼ぶ．語釈文は語義定義文を起点とし，要素を順

に追加する形で最大 3 回まで拡張が行われる．語釈文拡張は途中の段階であっても，語義が特定できた時点で終了する．

第 1 段階では，語義定義文を語釈文として語義の推定を行う．曖昧さ回避ページは目的の語義に対する語義解説記事の探索を容易にする目的で記述されることから，それぞれの語義特有の語句を用いた，違いが明瞭になる語義定義文で記述されると考えられる．よって，第 1 段階の語釈文を用いることで精度の高い語義推定が期待できる．

第 2 段階では，語義定義文に加え，語義解説記事リンクの有無を語釈文に取り込み，語義の推定を行う．Wikipedia では著名でない題材の記事は投稿しないよう提言されている [5] ことから，語義解説記事が存在する語義は存在しない語義よりも一般に知名度が高く，使われやすいと推測できる．よって，第 2 段階では使われやすい語義に優位性が与えられた語義推定が行われる．

第 3 段階では，第 2 段階の語釈文に加え，語義解説記事のインフォボックス情報と記事冒頭文を語釈文に取り込み，語義の推定を行う．インフォボックスには記事の要約情報が記述される．利用目的の一つとして情報比較を容易にすることが挙げられることから，インフォボックスはその記事が他の記事と異なる点が記述されると推測できる．一方，Wikipedia の記事冒頭文には記事の概略や分類に役立つキーワードが記述され [6]，その記事における主題の上位語が含まれることも多い [3]．特に上位語は，多義語の語義を推定するのに有効であることが知られている [1]．以上のことから，第 3 段階では語義推定精度を下げる要因になる可能性の低い語彙の増加が期待でき，再現率の向上が見込める．

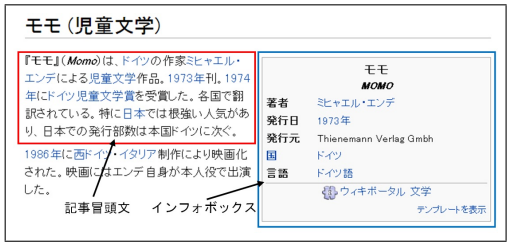

図 2: 記事冒頭文とインフォボックス

第 4 段階では，第 3 段階の語釈文に加え，語義解説記事の全文を語釈文に取り込み，語義の推定を行う．語義解説記事は語義の詳細を記述していることから，その語義に関連する語句を大量に取得できると考えられる．これにより，語義を推定するための語彙が大幅に増えることから語義推定の再現率向上が期待できる．

3.2 素性抽出

語義定義文は曖昧さ回避ページから取得する．曖昧さ回避ページは Wikipedia における箇条書きマークアップタグ（以下，タグ）である「*」「#」「;」「:」を用いた箇条書きで記述され，一行が一語義に対応することが多い．また，記事によっては，タグを連ねたり，タグの種類によりインデント幅が異なるという特徴を利用することで多段階層の箇条書きを表現している場合がある．

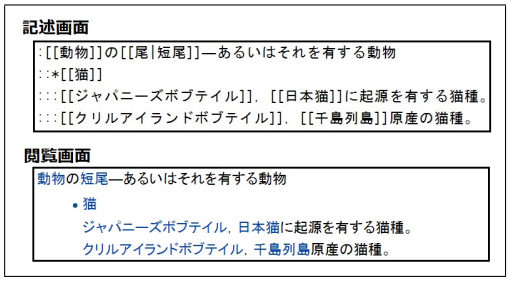

図 3: インデント幅の違いを用いた多段階層の箇条書き

多段階層の箇条書きには，

1. 語義を項目ごとに分ける
2. ある語義から派生した別の語義を記述する
3. 一つの語義定義文を複数行に分ける

の 3 種類の使われ方が存在する．1 は語義定義文としては相応しくない語句であるが，2 と 3 は語義定義文として抽出する必要があり，さらに 3 は複数行を一つの語義定義文として結合する必要がある．本手法では，多様に表現される曖昧さ回避ページから語義定義文を得るために，処理を 2 つに分けて行う．まず，曖昧さ回避ページをタグの数と種類を手がかりに根付き木に分解する．このとき，根付き木の各ノードはタグと行文字列を要素に持つ．続いて，語義定義文に相応しくないノードは除外し，複数行に分かれた語義定義文を結合する．まず，タグ「;」と「:」が交互に出現する場合，それらは 1 つの語義定義文である傾向が見られた．そこで，親ノードのタグが「;」，子ノードのタグが「:」であり，かつ，子ノードの数が 1 である場合，ノードの結合を行う．また，項目名は語義定義文に比べて短い語句で表現されることが多いことから，行文字列の形態素数が少ないノードの除外を行う．最終的に残ったノードの行文字列を，その多義語の語義の語義定義文として得る．

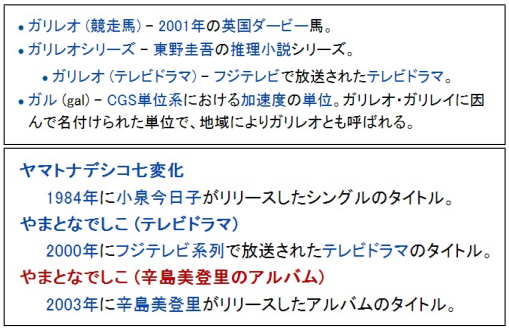

図 4: 多段箇条書きの多様性

語義解説記事リンクは前述の手法で得られた語義定義文から抽出する．調査の結果，語義定義文の多くは

- 語義解説記事リンクだけを持つもの
- 文の先頭に語義解説記事リンクを持つもの
- 文の末尾周辺に語義解説記事リンクを持つもの

の 3 つに分類され，さらに規則的な記述も確認された．語義解説記事リンクの抽出パターンを Pat. 1 に示す．ここで，**＜ LP ＞**にあたる文字列が語義解説記事リンクである．抽出された語義解説記事リンクを辿ることで語義解説記事の情報を取得することができる．

Pat. 1: 語義解説記事リンクの抽出パターン

^**＜ LP ＞**。?$	**＜ LP ＞** (は |。)
＜ LP ＞ (など)?を参照。?$	については**＜ LP ＞**
^**＜ LP ＞** [:-]	[:-→] **＜ LP ＞**。?$
[、。] **＜ LP ＞**。	**＜ LP ＞**の.+?コード
\s+**＜ LP ＞**$。	^**＜ LP ＞**\s+
＜ LP ＞の (略 | こと |. 称 |. 号 |. 名)	

3.3 スコア関数

抽出した情報を用いて，以下の式で表すスコア関数により語義の推定を行う．

$$\mathrm{Score}\,(d, a_i) = \frac{1}{1+\alpha}\left\{\mathrm{Sim}\,(\mathbf{w}_d, \mathbf{x}_{a_i}) + \frac{\alpha}{1+\beta+\gamma} \times \left(\mathrm{L}\,[a_i] + \beta\mathrm{Sim}\,(\mathbf{w}_d, \mathbf{y}_{a_i}) + \gamma\mathrm{Sim}\,(\mathbf{w}_d, \mathbf{z}_{a_i})\right)\right\}$$

ここで，a_i は多義語の i 番目の語義，$\mathbf{w}_d$ は文章 d から生成された文書ベクトル，$\mathrm{Sim}\,(\cdot,\cdot)$ は特徴ベクトル間の類似度をである．また，$\mathbf{x}_{a_i}$，$\mathbf{y}_{a_i}$，$\mathbf{z}_{a_i}$ はそれぞれ，語義 a_i に対する語義定義文，語義解説記事の冒頭文とインフォボックス情報，語義解説記事全文から生成された語義ベクトルであり，$\mathrm{L}\,[a_i]$ は語義 a_i に語義解説記事リンクが存在する場合は 1，そうでない場合は 0 をとる変数である．α,β,γ は語釈文の拡張段階を表すと共に，語義ベクトルの寄与度の調整も行う非負のパラメータである．語釈文が未拡張の時は $\alpha = \beta = \gamma = 0$，すなわち，$\mathrm{Score}\,(d, a_i) = \mathrm{Sim}\,(\mathbf{w}_d, \mathbf{x}_{a_i})$ により語義が推定される．特定の語義のスコアが他の語義のスコアよりも一定値以上大きい場合，その語義を文章 d における多義語の語義として出力する．そうでない場合，語釈文の拡張が確定し，$\alpha > 0$ に変化させる．第 2 段階では $\mathrm{Score}\,(d, a_i) = \frac{1}{1+\alpha}\{\mathrm{Sim}\,(\mathbf{w}_d, \mathbf{x}_{a_i}) + \alpha\mathrm{L}\,[a_i]\}$ により語義が推定される．それでも語義が確定出来ない場合は，β，γ を順に正の値に変化させる．語釈文拡張が第 4 段階まで到達している場合，スコアが一番高い語義に対し閾値による判定を行う．語義のスコアが閾値を超えた場合，その語義を文章 d における多義語の語義として出力し，超えなかった場合，曖昧さ回避ページには該当する語義が無いとして出力を行う．パラメータ α,β,γ，および，閾値は各拡張段階で共通しているわけではないことに注意したい．つまり，第 2 段階の α と第 3,4 段階の α は異なる値を取る可能性がある．

4 実験

提案手法の語義推定性能を評価するために，Yahoo! 知恵袋[1]の質問文を用いた語義推定実験を行った．多義語を含む質問文 271 件（うち，曖昧さ回避ページに正解語義が存在するものは 239 件）に対し正解の語義を付与し，5 分割交差検定による平均正解率を算出した．学習ではパラメータ α,β,γ，および，語義を決定する閾値を算出する．比較対象として，α,β,γ が固定された，拡張が行われない語釈文を用いた正解率を算出した．また，語義定義文，記事冒頭文とインフォボックス，語義解説記事全文の 3 種類の文章のうち，いずれか 1 つだけを用いて語義を推定したときの正解率も算出した．さらに，黒川らの手法を参考に，WEB 検索を用いて語釈文を拡張する手法を実装し，これも比較対象として正解率の算出を行った．曖昧さ回避ページに正しい語義が存在しない質問においては，語義が無いと出力された場合を正解として計上している．正解率は，曖昧さ回避ページに正しい語義が存在する質問文だけを対象にしたもの（正解率 1）と，正しい語義が存在しない質問文も対象にしたもの（正解率 2）の 2 つを求めた．これらの正解率を表 1 に示す．

また，提案手法において，語義がどの拡張段階で決定したのか，その数を表 2 に示す．このとき，各パラ

[1] http://chiebukuro.yahoo.co.jp/

表 1: 語義推定正解率

	正解率 1	正解率 2
提案手法	**0.51**	**0.46**
固定した語釈文を用いる手法	0.48	0.42
語義定義文を用いる手法	0.33	0.31
冒頭文とインフォボックスを用いる手法	0.29	0.29
語義解説記事全文を用いる手法	0.45	0.42
WEB 検索を用いる手法	0.44	0.39

メータ，および，閾値は交差検定において最も高い正解率を示した組み合わせを用いた．

表 2: 語義が決定した際の拡張段階

拡張段階	推定成功	推定失敗	正解率
第 1 段階	15	3	83%
第 2 段階	3	2	60%
第 3 段階	36	17	68%
第 4 段階（語義あり）	78	111	41%
第 4 段階（語義無し）	2	4	33%
合計	134	137	49%

表 1 から，固定した語釈文を用いて推定したときと比べ提案手法の正解率が高いことから，語義推定において語釈文を段階的に拡張したことの有効性が確認できる．さらに，WEB 検索を用いて語釈文を拡張する手法よりも精度が高いことから，語義推定において提案した語釈文拡張は，より有効な語句を取得できたことがわかる．表 2 より，第 1 段階が最も正解率が高いことから，語義定義文による語義推定は高精度に行えることが確認できる．以上より，提案した段階的語釈文拡張は WSD において有効に働くことが分かる．

5 おわりに

語義曖昧性解消の問題に対し，段階的に拡張される語釈文を用いて語義の推定を行う手法を提案した．実験より，提案手法は固定した語釈文や WEB 検索を用いて拡張した語釈文を用いて語義を推定する手法よりも，高い精度で推定できることを示した．今後の課題として，語義解説記事を語釈文として用いる際，冒頭文やインフォボックス以外の素性選択を考えたい．例えば，連語は多義語の語義を推定するのに有効であることが知られている [4]．語義解説記事には多義語自身が含まれることも多いため，連語を抽出し，語釈文拡張に組み込めるのではないかと考える．また，曖昧さ回避ページには箇条書きで記述された語義の他に，語源となる語義が冒頭文に記述される場合があるため，それらの抽出手法を考案したい．

参考文献

[1] S. Banerjee and T. Pedersen. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the 18th international joint conference on Artificialintelligence*, IJCAI'03, pp. 805–810, San Francisco, CA, USA, 2003. Morgan Kaufmann Publishers Inc.

[2] S. Cucerzan. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods inNatural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 708–716, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[3] J. Kazama and K. Torisawa. Exploiting Wikipedia as external knowledge for named entity recognition. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 698–707.

[4] Y. K. Lee and H. T. Ng. An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In *Proceedings of the ACL-02 conference on Empirical methods in naturallanguage processing - Volume 10*, EMNLP '02, pp. 41–48, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[5] Wikipedia. Help:infobox. ウィキメディア財団. `http://ja.wikipedia.org/wiki/Wikipedia:`記事を執筆する, 参照 Dec 23, 2013.

[6] Wikipedia. Wikipedia:素晴らしい記事を書くには. ウィキメディア財団. `http://ja.wikipedia.org/wiki/Wikipedia:`素晴らしい記事を書くには, 参照 Dec 23, 2013.

[7] 胡, 谷田. Wikipedia のカテゴリ情報を用いた twitter ユーザの関心分野の抽出. 信学技報 言語理解とコミュニケーション研究会, 第 113 巻, pp. 17–21, 2013.

[8] 黒川, 新里, 黒橋. 段階的文脈拡張による多義性解消. 言語処理学会第 17 回年次大会, pp. 544–547, Mar. 2011.