ウィキペディアを用いた語義曖昧性解消手法と情報検索への応用

A Word Sense Disambiguation Method using Wikipedia and Its Application to Information Retrieval

道下 智之[*1] 中山 浩太郎[*2] 原 隆浩[*1] 西尾 章治郎[*1]
Tomoyuki MICHISHITA Kotaro NAKAYAMA Takahiro HARA Shojiro NISHIO

[*1]大阪大学大学院情報科学研究科
Dept. of Multimedia Engg., Graduate School of Information Science and Techn., Osaka University

[*2]東京大学知の構造化センター
Center for Knowledge Structuring, The University of Tokyo

In this paper, we propose a method for word sense disambiguation using Wikipedia to improve the performance of information retrieval. The disambiguation method is a process of maximizing the agreement between the feature vector of a document and the feature vector of a word sense in the vector space model of Wikipedia entities. Through information retrieval evaluation, we show the method is beneficial in information retrieval.

1. はじめに

近年，ウィキペディアが語義曖昧性解消のためのコーパスとして注目されている．これは，ウィキペディアでは一つのページが一つの概念に対応しており，URL を利用して語の曖昧性を解消しているためである．曖昧な語の例として「Spring」について考える．この語は，季節の春，機械のバネ，泉といった語義があるが，ウィキペディアでは「Spring (season)」,「Spring (device)」,「Spring (hydrosphere)」といった別々のページでそれぞれの語の意味が定義されている．また，ウィキペディアには曖昧な語とその語義の一覧を提示する Disambiguation Page（曖昧さ回避ページ）が存在し，曖昧性解消のコーパスとして有用な情報の一つとなっている．

情報検索（特に Web など）の分野では，新しい概念やドメインに特化した概念など，多様なクエリが発行される．そのため，これら新しい概念や専門分野を広く網羅した曖昧性解消の方法が必要であった．筆者らは，この曖昧性解消の方法として，即時性や網羅性の観点からウィキペディアを用いた語義曖昧性解消の手法が有用であると考えている．

本稿では，Disambiguation Page から語義情報を抽出し，語義曖昧性解消のための辞書を作成する手法，および作成した辞書を用いた語義曖昧性解消手法を提案する．また，提案手法を情報検索に応用し，情報検索テストコレクションによって提案した語義曖昧性解消手法を評価する．

2. 関連研究

これまで語義曖昧性解消の研究は数多く行われてきたが，最近はウィキペディアを用いた語義曖昧性解消の研究に注目が集まっている．

Mihalcea [Mihalcea 07] は，ウィキペディアの記事中のリンクを語義タグとして扱うことによって，ウィキペディアが語義曖昧性解消の学習コーパスとして利用可能であることを示している．ウィキペディアの記事中のリンクは，アンカーテキストが曖昧な語であっても，リンク先のページはコンテキストに沿った概念が参照されている．例えば，曖昧な語「Spring」がアンカーテキストになっている「A spring is an elastic object used to store mechanical energy.」という一文を考える．この「spring」 は機械部品の意味であり,「Spring (device)」にリンクが張られている．このように，アンカーテキストが曖昧な語であるとき，ハイパーリンクは語義が定義された一種のタグとして見なすことができる．この研究では，ウィキペディアを語義タグ付き学習コーパスとして利用することの有用性を示すために，学習コーパスから得られるコンテキスト情報を単純ベイズ分類器で学習させて，SENSEVAL-2 と SENSEVAL-3[*1]のテストコレクションを用いて評価している．結果として，SENSEVAL のコーパスで学習した分類器よりも，ウィキペディアから作成したコーパスで学習した分類器のほうが高い精度が得られている．

Cucerzan [Cucerzan 07] は，ウィキペディアから概念に関するコンテキストやカテゴリータグといった情報を抽出し，それらを用いることで与えられた文書に出現する固有表現の曖昧性を解消する手法を提案している．Cucerzan は，ニュース記事とウィキペディア記事に対して評価実験を行い，提案された手法が高い精度で語義曖昧性解消を行うことを示している．

本稿では，Mihalcea のようにウィキペディアを学習コーパスとして利用することはせず，Cucerzan のようにウィキペディアから抽出した情報を用いる語義曖昧性解消手法を提案する．

3. 提案手法

語義曖昧性解消を行う方法として，語の意味を定義した辞書やシソーラスを用いる方法が一般的である．本章では，Disambiguation Page から語義情報を抽出し，語義曖昧性解消のための辞書を作成する手法，および作成した辞書を用いた語義曖昧性解消手法について説明する．

3.1 語義辞書作成

Disambiguation Page の多くは，図 1 のように曖昧な語の語義をリスト構造で記述しており，リストの各要素が一つの語

連絡先: 道下 智之，大阪大学大学院情報科学研究科 マルチメディア工学専攻 西尾研究室，
〒 565-0871 大阪府吹田市山田丘 1-5
大阪大学大学院情報科学研究科 マルチメディア工学専攻マルチメディアデータ工学講座（西尾研究室），
michishita.tomoyuki@ist.osaka-u.ac.jp

∗1 http://www.senseval.org/


Spring

From Wikipedia, the free encyclopedia

Spring may refer to:

- Spring (season), a season of the year
- Spring (device), a mechanical part
- Spring (hydrosphere), a natural source of water

Art

- *Primavera* (painting) ("Spring"), a painting by Sandro Botticelli, c. 1482.
- *Spring* (painting), an oil by Lawrence Alma-Tadema
- *Spring*, a painting by Christopher Williams

図 1: 「Spring」の Disambiguation Page

表 1: 作成した辞書における「Spring」の定義

番号	概念集合	説明文
1	Spring (season)	a season of the year
2	Spring (device)	a mechanical part
3	Spring (hydrosphere)	a natural source of water
...	...	...

義を示している．これを利用し，各要素ごとにテキスト情報とリンク先の概念を抽出し，それぞれを語義の「説明文」,「概念集合（語義を示す概念と語義に関連する概念の集合)」として抽出する．このとき，以下のルールを満たすように抽出する．

- HTML の見出しが「See also」となっている内容は除外する．この内容は関連項目であり，語義とはなりえないからである．
- ネストされているリストは，その上位にあるリストの要素に含まれるものとする．
- Disambiguation Page の最初の一文にリンクが含まれる場合，これはリストの要素と同様に語義とする．これは，最も一般的な語義はリスト構造とは別に最初の一文で定義されることが多いためである．

この手法から抽出された「概念集合」,「説明文」に語義の識別子となる「番号」を付与して，曖昧な語の語義が定義されている辞書を作成した．表 1 は，作成した辞書に定義されている「Spring」である．

3.2 語義曖昧性解消

本手法では，曖昧な語の語義および文書を特徴ベクトルとして表現し，文書の特徴ベクトルと最も類似度の高い特徴ベクトルをもつ語義がその文書で使われていると識別する．特徴ベクトルは，各次元をウィキペディアの各概念とし，その要素は 0 か 1 の二値で表現する．つまり，特徴ベクトルの次元数は，ウィキペディアの概念数と同じである．

語義の特徴ベクトルは，辞書で定義されている「概念集合」に含まれる概念とその概念を記述した記事に含まれるリンク先の概念を 1，これらの概念に含まれないものを 0 として表現する．

文書の特徴ベクトルは，文書のテキストから抽出したウィキペディアの概念とその概念に関連する概念を 1，これらの概念に含まれないものを 0 として表現する．テキストからの概念抽出は，テキスト中からウィキペディアのタイトルと合致した名詞，または名詞が連続する句を抽出することで行う．ただし，抽出したタイトルのうち曖昧な語であるものは除去する．

表 2: 提案手法による情報検索精度の比較

	適合率	再現率	F 値	MAP	R 適合率
提案手法あり	0.177	0.137	0.154	0.121	0.159
提案手法なし	0.127	0.207	0.158	0.085	0.124

また，関連する概念の取得には，ウィキペディアシソーラス [Nakayama et al 07] を用いた．

本手法では，語義の特徴ベクトルと文書の特徴ベクトルとのコサイン類似度が最大になったものに曖昧性解消を行う．ただし，この類似度が最大になる語義が複数ある場合（類似度がすべて 0 など）は，それらの語義すべてに曖昧性解消を行う．つまり，語義を複数持つことを許容する．

4. 実験・結果

提案手法の評価として，計算機科学に関する文献から作成された情報検索テストコレクション CISI[*2]を用いて実験を行った．文章集合の曖昧な語に対してはあらかじめ提案手法で語義曖昧性解消を行っておき，手動で曖昧性解消を行ったクエリ 19 個を用いて，情報検索の性能評価を行った．

実験結果を表 2 に示す．提案手法を用いることによって，適合率が 0.05（相対比 39%）増加した．このことより，提案手法は検索結果の適合文書の割合を増やすのに効果的であることがわかる．適合率と再現率を考慮に入れた評価指標である F 値では，ほとんど差がなかった．しかし，MAP や R 適合率がそれぞれ 0.036（相対比 42%），0.035（相対比 28%）増加していることから，検索結果の上位に適合文書が増えていることがわかる．この結果は，提案手法が情報検索の性能を向上するのに有用であることを示している．

なお，今回使用したテストコレクション CISI は，計算機科学に関する文献から作成されており，クエリにも多数の計算機科学の専門用語が使われていた．ウィキペディアでは，クエリに出現した専門用語が定義されており，それを用いた提案手法もそれら専門用語を網羅することができていた．

5. まとめ

本稿では，ウィキペディアを用いた語義曖昧性解消手法を提案し，情報検索に応用することで，その有用性を示した．今後の課題として，Web 検索のようにさらに多様なクエリを持つテストコレクションで本手法を評価することを検討している．

参考文献

[Mihalcea 07] R. Mihalcea: Using Wikipedia for Automatic Word Sense Disambiguation, *In Proc. of NAACL-HLT*, pp. 196-203 (2007)

[Cucerzan 07] S. Cucerzan: Large-Scale Named Entity Disambiguation Based on Wikipedia Data, *In Proc. of EMNLP-CoNLL*, pp. 708-716 (2007)

[Nakayama et al 07] K. Nakayama and T. Hara and S. Nishio: Wikipedia Mining for an Association Web Thesaurus Construction, *In Proc. of WISE 2007*, pp. 322-334 (2007)

∗2 ftp://ftp.cs.cornell.edu/pub/smart/