単語の定義文による辞書検索

西村 徹郎　　橋本 泰一　　徳永 健伸

東京工業大学 大学院情報理工学研究科 計算工学専攻

{nishimura,taiichi,take}@cl.cs.titech.ac.jp

1 はじめに

文章を読んでいて単語の意味が分からないとき，我々は通常国語辞典を用いて，その単語の意味を調べる．近年の電子辞書の発達により，我々は単語の表記や読みからその意味を素早く検索する事ができる．しかし，逆の場合はどうだろうか．単語の意味を知っているが単語の表記や読みを知らない場合はどうすればいいだろうか．例えば，文章を書いているときに，良い言葉が思いつかないというような場合である．

従来の電子辞書の検索機能は見出し語の読みや表記に対する検索機能がほとんどであり，内容に関する検索機能はほとんど実現されていない．単語の定義文を全文検索する機能もあるが，基本的に入力キーワードにマッチさせることがほとんどであり，柔軟な検索を行う事はできない．また，辞書の定義文は簡潔に記述されており，同じことを他の表現で言い換えることしないと，ユーザの多様なクエリにマッチさせる事は難しい．

辞書は膨大な言語資源であるが，このような状況を考えると辞書が十分に活用されているとは言えない．そこで，本稿では，ある単語の定義文からその単語を検索するシステムを提案する．例えば，このシステムは「自分が生まれた国」という入力に対して,「祖国」「母国」というような単語のリストを出力する．

同様なシステムとして MTW(Meaning To Word)[1] が存在する．このシステムはトルコ語の国語辞書を用いて実装されており，入力文と辞書の定義文の類似度を単語のマッチングを用いて計算している．しかし，渡辺 [6] によると日本語で同様の手法を用いて，ユーザが検索したい単語を正確するのは難しい．

そこで，本システムでは定義文の特徴をを利用する事で，正確な検索を実現する．さらに，複数の言語資源を組み合わせる事でユーザの多様な表現に対応できるようにした．

2 システムの概要

このシステムは，ユーザの自然文によるクエリを入力とし，その文の表す単語のリストを出力する．辞書資源として岩波国語辞典 [4] を用いている．岩波国語辞典には，各項目に見出し語と定義文が書かれており，入力とその定義文の類似度を計算し，スコアの高い順に表示する．

類似度の計算にはベクトル空間モデルを利用している．入力 q と定義文 d をそれぞれ単語に分割する．そこで，入力 q は各単語が存在するかどうかの 01 の二値ベクトル $\mathbf{q}_{\text{単語}}$ で表し，定義文 d は重み付きベクトル $\mathbf{d}_{\text{単語}} = (weight(t_1), \ldots, weight(t_n))$ で表す．重みは以下の **IDF** を用いる．N は全定義文の数を表し，$df(t)$ は単語 t を含む定義文の数を表す．

$$weight(t) = idf(t) = log(N/df(t)) \tag{1}$$

従来の検索では $\mathbf{TF} \cdot \mathbf{IDF}$ が利用されることが多いが，辞書の定義文は文章の長さが短いため **TF** の項が有効に働かない．しかし，**IDF** は文章の長さに関わらずに有効に働くので **IDF** のみ用いる．

これらのベクトルに対して内積をとることによって類似度を計算する

$$sim_{\text{単語}}(q, d) = \mathbf{q}_{\text{単語}} \cdot \mathbf{d}_{\text{単語}} \tag{2}$$

ここで，余弦や定義文に含まれる単語の数を用いて正規化を行っていないのも，定義文が短いため有効に働かないからである．

3 辞書の定義文の特徴の利用

2 節で説明した単語によるマッチングののみを用いるとユーザの検索したいものを上手く検索出来ないことがある．例えば，

> 大量のススがでる電灯

という入力に対して，

> アーク灯 (放電を利用した 電灯)
> 電光 (電灯 の光)

という語が出力される．ユーザは何らかの「電灯」を検索したいにも関わらず,「電灯」の部分のみがマッチするため,「電灯」以外のもの（電光）も検索してしまう．そのためユーザが何を検索したいのか正確に掴む

ことが必要である．そこで，定義文という特徴を利用する．

3.1 定義文の特徴

定義文は見出し語の様々な属性を表している．これらの属性をマッチさせる事で正確な検索を行う．属性の中で一番重要なものが上位語 [3] である．上位語とは定義文の中心となる語であり，見出し語の上位概念に当る．

上位語は表 1 に示す 3 つの方法を用いて抽出した．

表 1: 上位語の抽出方法

種類	見出し語	定義文
最後の単語	アーク灯	放電を利用した 電灯
テンプレート	愛育	可愛がって 育てる こと
係り受け	亜高木	高木 の中で小型のもの

まず，一番最後の単語が上位語であることが多いのでこれを抽出する.「放電を利用した電灯（アーク灯)」ならば「電灯」を上位語として抽出する．しかし,「可愛がって育てること（愛育)」では「こと」よりも「育てる」の方が上位語としてふさわしい．そのため,「〜こと」というテンプレートを用いてこれを抽出する．しかし,「高木の中で小型のもの（亜高木)」はテンプレートを用いてもうまく行かない．この場合，以下の手法を用いる．

1. 「〜の中で」「〜で」「〜のうち」の部分を上位語の候補として抽出する．この場合「高木」になる．
2. 上位語の候補（「高木」）が最後の単語（「もの」）の修飾語（「小型の」）を修飾していれば次に進む．
3. 上位語の候補（「高木」）が最後の単語（「もの」）の下位概念であれば，その単語（「高木」）を上位語とする．

また，上位語の修飾語の部分は重要な属性を表している．例えば,「放電を利用した電灯（アーク灯)」の上位語「電灯」を修飾する「放電を利用した」の部分はこの「電灯」どのような「電灯」であるかを表している．これを格の情報や品詞情報と同時に抽出し，正確なマッチングを行う．

入力と定義文からそれぞれ抽出した属性の集合をベクトルで表現する．例えば,「放電を利用した電灯（アーク灯)」であれば，属性の集合は以下のようになる．

{(ヲ格, 放電), (動詞, 利用する), (上位語, 電灯)}

このように,「(ヲ格, 放電)」のペアが一つの要素となる．入力のベクトル $\mathbf{q}_{\text{属性}}$ はその属性が存在するかどうかの 01 の二値ベクトルで表され，定義文のベクトル $\mathbf{d}_{\text{属性}} = (weight(t_1), \ldots, weight(t_n))$ は重み付きベクトルになる．ここで各 t_i は属性を表す．式 (1) の **IDF** の重みを用い，内積を計算する事で類似度を求める．

$$sim_{\text{属性}}(q, d) = \mathbf{q}_{\text{属性}} \cdot \mathbf{d}_{\text{属性}} \tag{3}$$

この類似度と式 (2) の単語のマッチングによる類似度を足し合わせることによって，総合的な類似度を計算する．現在のシステムでは $\alpha = 0.5$ としている．

$$sim_{\text{単語+属性}}(q, d) = \alpha \times sim_{\text{単語}}(q, d) + (1 - \alpha) \times sim_{\text{属性}}(q, d)$$

しかし，属性を抽出できても単語の完全一致では上手く検索出来ない場合がある．例えば,「奴（やっこ)」という単語に対して，

入力：武家の 使用人
定義文：武家の 奴隷

である場合,「使用人」と「奴隷」の意味はほぼ同じであるのにも関わらず，属性のスコアを足すことができない．そこで，本システムではシソーラス (日本語語彙大系 [2]) を用いて「奴隷」を「使用人」に拡張する事によって，入力と定義文の上位語の属性をマッチさせることができる．

3.2 評価

これまで説明してきた手法に対して評価を行った．テストセットとしてクロスワードのカギを利用した．クロスワードのカギのうち単語の定義を用いて質問しているもののみを 269 問利用した．各カギに対する解答は岩波国語辞典に含まれている．

以下の 3 つの手法を比較した．

ベースライン 単語のマッチングのみを用いる手法

属性 属性を利用する手法

シソーラス さらにシソーラスを利用した手法

評価基準として正解率を用いた．正解率は上位 n 位まで検索した結果に正答が含まれる割合を表している．結果を表 2 に示す．

表 2: 本システムの手法と従来の手法との比較

n	**ベースライン**	**属性**	**属性+シソーラス**
10	92 (34.2%)	99 (36.8%)	**105 (39.0%)**
50	128 (47.6%)	**136 (50.6%)**	**136 (50.6%)**

属性を用いることによって正解率が向上し，シソーラスを用いることでさらに向上している．しかし，上位 50 位までに検索出来ないものが 133 問もあり，十分に検索できているとは言えない．そこで，これらのエラーの原因を調べ 6 つに分類した．各エラーの件数を表 3 に示す．

表 3: エラーの種類

エラーの種類	エラーの数
表記の違い	20
否定表現の違い	10
上位語の違い	20
その他の属性の違い	14
推論	17
情報の不足	34

「表記の違い」とは漢字と平仮名の違い，送り仮名の違い等で完全一致しなかったため，検索結果から漏れてしまったものである．

「否定表現の違い」の例として「空き地」の定義文を以下に示す．

入力：未使用 の土地
辞書：何にも 使っていない 土地

「未使用」のように「否定の接頭辞」を用いて否定を表現する時と,「ない」を使って否定を表す時がある．これらの情報は，ルールを用いて情報を正規化すれば検索出来る．

「上位語の違い」の例として「乱視」の定義文を以下に示す．

入力：物が変にゆがんで 見える こと。
辞書：物の形をはっきり見ることのできない 眼。

入力と辞書で指しているものが異なっている.「乱視」の意味は，入力では眼が見えない状態の事を指し，辞書では見えない眼のことを指している．

「その他の属性の違い」の例として「金貨」の定義文を以下に示す．

入力：金の お金
辞書：金を 主成分とする貨幣

本システムでは表層格を用いて抽出しているので,「金の」と「金を主成分とする」では単語のスコアを得る事は出来るが，属性情報のスコアを得る事ができない．これは意味解析まで行わないと正確に検索出来ない．

「推論」の例として「船長」の定義文を以下に示す．

入力：船の一番 偉い 人。
辞書：船の乗組員の 長。

「長」という言葉と「偉い」という言葉の関連性を得るためには「長は偉い」という推論が必要であり，そのような知識ベースが必要である．

「情報の不足」の例として「もやし」の定義文を以下に示す．

入力：ひょろっと細長くて 白い野菜。
辞書：大豆・麦などの粒を水にひたし暗所で 発芽させたもの。

「ひょろっと細長くて」という情報は辞書の「もやし」の定義文には書かれていない．辞書に情報が不足しているという問題を解決する一つの方法として言語資源をもっと増やすという手がある．シンプルであるが，辞書によって定義文が異なる書き方をしていたり，ある辞書には記述されていない情報が他の辞書にある場合もあるだろう．これは 4 節で説明する．

4 複数の言語資源の統合

3 節では入力と定義文で記述されている情報が異なっているものが多く存在していた事が分かった．一つの辞書資源だけでは，正解率の向上に限界がある．そこで，複数の言語資源を利用する事で様々な表現・情報を追加する．新しい言語資源として広辞苑 [5] と Wikipedia[1] を使用した．

4.1 言語資源の比較

各言語資源の比較を表 4 に示す．

表 4: 各言語資源の比較

言語資源	単語数	平均文数
岩波国語辞典	5 万	2.10
広辞苑	23 万	3.13
Wikipedia	22 万	7.08

広辞苑は国語辞典と百科事典を兼ねた辞典である．単語数は約 23 万語あり各定義文の長さも岩波国語辞典より多い.「アーケード」を岩波国語辞典と広辞苑で引いてみると，以下のように記述されている．

岩波：かまぼこ形の天井をもった通路。
広辞苑：建築物で、柱列上に同型のアーチが連続したもの。街路に沿って、通路をなす。

[1] http://ja.wikipedia.org/

岩波では上位語が「通路」，広辞苑では上位語が建築物であるが,「通路をなす」と書かれてあるのでほぼ同じことを表現している．このように，同じ国語辞書という特徴を共有しているため，同じものを表しているが，表現が異なる単語が多い．

Wikipedia はオンラインのフリーの百科事典である．単語は約 22 万語あり定義文の長さも岩波より多い．Wikipedia は専門家が書いているのではなく，一般のユーザが自由に編集する事が出来る．そのため，実際のユーザのクエリに近いのではないかと考えた．岩波国語辞典と Wikipedia で「アーチ」を引いてみると以下のように記述されている．

> 岩波：上方が半円形をなす構造物。
> Wikipedia：アーチは、下部が上に凸な曲線形状をした梁、もしくは上に凸な曲線形状そのもの。

岩波国語辞典では,「アーチ」は「構造物」を指しているのに対し，Wikipedia では「梁」や「形状」を指しており，別のものを意味している．このように岩波国語辞典と Wikipedia では異なる情報が記述されていることが多い．

4.2 評価

これらの言語資源を追加してシステムを再構築した．テストセットは 3 節と同様にクロスワードのカギ 269 問を用いた．このテストセットは岩波国語辞典に正答が存在するものについて収集したので，広辞苑や Wikipedia には存在しない物がある．そのため，2 通りの実験を行った．1 つはテストデータを全て利用するもので言語資源の追加による全体的な効果を調べる．もう一つは 2 つ言語資源に共通するテストデータのみを用いて，各言語資源の効果を調べた．

表 5: 複数の言語資源によるシステムの評価

n	岩波	岩波+広辞苑
10	105 (39.0%)	121 (45.0%)
50	136 (50.6%)	**157 (58.4%)**
n	**岩波+Wikipedia**	**岩波+広辞苑+Wikipedia**
10	105 (39.0%)	**128(47.6%)**
50	146 (54.3%)	154 (57.2%)

表 6: 岩波国語辞典と広辞苑の比較

n	岩波	広辞苑	岩波+広辞苑
10	81 (33.8%)	83 (34.6%)	**103 (42.9%)**
50	122 (50.8%)	118 (49.2%)	**141 (58.8%)**

表 7: 岩波国語辞典と Wikipedia の比較

n	岩波	Wikipedia	岩波+Wikipedia
10	29 (33.3%)	38 (43.7%)	**46 (52.9%)**
50	45 (51.7%)	48 (55.2%)	**61 (70.1%)**

表 5 に全テストデータを用いた実験結果を示す．この結果によると，各言語資源を追加する事によって，正解率は向上しているが，Wikipedia の効果はあまりないことが分かる．上位 50 位までの結果だと，Wikipedia を追加しない方が結果が良い．

表 6 に岩波国語辞典と広辞苑の比較を，表 7 に岩波国語辞典と Wikipedia の比較を示す．この結果によると，広辞苑も Wikipedia も正答が存在するテストデータに対しては有効に働いているようである．これは，Wikipedia が岩波国語辞典とは異なる情報が記述されているためだと考えられる．

5 まとめ

本研究では，単語の意味から単語を検索するシステムを構築した．本システムの構築において，辞書に記述されている定義文という検索対象の特徴を利用した．単語の定義文は様々な属性を表しており，上位語と上位語を修飾する文節が特に重要な属性を表しているという考えに基づいて，属性情報を用いた検索手法を提案した．さらに，言語資源を増やす事によって多様なクエリに対応出来るようにし，正解率を向上させる事ができた．

参考文献

[1] Ilknur Durgar El-Kahlout and Kemal Oflazer. Use of wordnet for retrieving words from their meanings. In *Proceedings of the Second International WordNet Conference*, pp. 118–123, January 2004.

[2] 池原悟, 宮崎正広, 白井悟, 横尾昭男, 中岩浩巳, 小倉 健太郎 大山芳史, 林良彦（編）. 日本語語彙大系. 岩波書店, 第 1 版, 1997.

[3] 正津康弘. 国語辞典とシソーラスの統合に関する研究. 修士論文, 2003.

[4] 西尾実, 岩淵悦太郎, 水谷静夫（編）. 岩波国語辞典. 岩波書店, 第 5 版, 1994.

[5] 新村出（編）. 広辞苑. 岩波書店, 第 5 版, 1998.

[6] 渡辺渉. 自然文による単語検索システムの構築. 学士論文, 2004.