内容語および機能語との共起分布を用いた述部の同義判定

中村 紘規 †　泉 朋子 ‡　柴田 知秀 †　黒橋 禎夫 †
† 京都大学大学院情報学研究科　‡NTT サイバースペース研究所
† {nakamura, shibata, kuro}@nlp.ist.i.kyoto-u.ac.jp
izumi.tomoko@lab.ntt.co.jp

1　はじめに

自然言語処理において同義性認識は非常に重要なタスクであり、情報検索や機械翻訳などのアプリケーションに有用である。本論文では、文の意味の核である述部に焦点をあて、述部の同義判定を行う。述部の意味を捉えるために人手で構築・整備されたリソースを使うことが考えられるが、カバレージの問題や分類の一貫性の問題などが存在する。

そこで大規模コーパスを用いて分布類似度の考えに基づき同義判定を行う研究が行われている [2, 4]。これらの研究では述部の意味を項に出現する名詞の分布によって捉えている。また、柴田らは動詞の多義性に対処するために、格要素と述部を組み合わせ、それに対して係り受け関係にある述語を素性としている [7]。例えば、「ブレーキを踏む」と「ブレーキをかける」の場合、どちらも「〜減速」、「〜停車」、「追突し〜」などと共起するため、「ブレーキを」が格要素であるもとでは「踏む」と「かける」は同義とみなすことができる。

このように分布類似度の研究において内容語との共起分布が用いられることが多いが、意味の類似した表現に対しては機能語との共起分布も類似していると考えらえる。上記の例においては、「〜損なう」「〜すぎる」などの機能語の分布も類似していると考えられる。そこで本研究では内容語だけでなく機能語との共起分布を用いることによって述部の同義性を判定する手法を提案する。本研究では内容語の素性としてあまり考慮されていない副詞を機能語扱いにする。

2　関連研究

動詞の同義性判定には人手により構築された辞書を用いる方法と、自動判定を行う方法がある。

人手で構築・整備された辞書には、分類語彙表や類語大辞典、LCS 辞書 [8] があげられる。分類語彙表は、語を意味によって ⟨ 感覚 ⟩ や ⟨ 作用・変化 ⟩ などの数百のクラスに分類し整理したものである。類語大辞典は、「生きる・死ぬ」や「感じる」など 100 種類の動詞・形容詞をもとに、そこから派生、連想される語をまとめて整理したものである。LCS 辞書は「本棚に本を置く ⊂ 本が棚にある」などのような動詞の包含関係を 5 階層 940 分類の細分類で分類したものである。これらの辞書を用いる方法には、カバレージの問題や分類の枠組みを変更する際の編集の手間の問題などが存在するため、コーパスなどから自動で学習できるようにすることが望ましい。

大規模なデータから自動で同義性を判定する研究として、VerbNet を拡張した Sun ら [3] の研究や、分布類似度を用いた Lin ら [2] の研究や Szpektor ら [4] の研究などがある。Sun らは Spectral Clustering を用いることで、英語における大規模な類語辞典である VerbNet をフランス語に自動的に拡張できることを示した。Lin らは、依存構造木から作成された二つのパス間の類似度を、前後に出現する名詞の分布類似度により計算した。例えば、“X wrote Y” と “X is the author of Y” の X 間、Y 間の類似度を計算することで、これらの表現間の類似度を計算している。Szpektor らは、単項の推論規則を分布類似度を用いて獲得した。例えば、“X takes a nap→X sleeps” という規則を、“X takes a nap” と”X sleeps” の X 間の分布類似度などを計算することで求めている。

3　分布類似度計算

本研究では我々が以前提案した手法 [6] に基づいて分布類似度を計算する。この手法では名詞に対して係り受け関係にある述語を素性として分布類似度を計算していたが、本研究では述語項に対して、内容語および機能語を素性として分布類似度を計算する。

まず、コーパスから、分布類似度を計算する単位 u とその素性 f を抽出し、その結果を集計することによ

減速:post　滑る:post　…　徐行:pre　脱輪:post　…
(10.1,　9.8,　…　7.9,　7.7,　…)

図 1: 「ブレーキをかける」の内容語ベクトル

り、u の頻度ベクトルを構築する。

そして分布類似度計算を関数 *weight* と関数 *measure* に分解する [1]。関数 *weight* は素性ベクトルの値を適切な値に変換するものであり、関数 *measure* は関数 *weight* で変換された値が要素であるベクトル間の類似度を計算するものである。

我々は以前、名詞の分布類似度計算において相澤の評価セット [5] を用いて、以下の *weight* 関数と *measure* 関数の組み合わせが最も精度が高いことを示した [6]。

weight 関数:

$$weight = \begin{cases} 1(\mathsf{MI} > 0) \\ 0(otherwise) \end{cases} \quad (1)$$

ここで、$\mathsf{MI} = \log \frac{P(u,f)}{P(u)P(f)}$ であり、$P(x)$ はコーパス中での出現確率を表わす。

measure 関数:

$$measure = \frac{1}{2}(\mathsf{JACCARD} + \mathsf{SIMPSON}), \quad (2)$$

ここで、

$$\mathsf{JACCARD} = \frac{|(u_1, *) \cap (u_2, *)|}{|(u_1, *) \cup (u_2, *)|} \quad (3)$$

$$\mathsf{SIMPSON} = \frac{|(u_1, *) \cap (u_2, *)|}{min(|(u_1, *)|, |(u_2, *)|)} \quad (4)$$

である。

上記の *weight* 関数と *measure* 関数は名詞の分布類似度計算において最も精度が高かったものであるが、本論文で扱う述語項の分布類似度計算においても精度が良いと仮定し、これらの関数を用いる。

4　素性の抽出

述語項に対して、内容語の素性と機能語の素性を抽出し、それぞれ上記で述べた分布類似度計算方法に基づき述語項の類似度計算を行う。内容語の素性、機能語の素性の抽出方法を順に述べる。

4.1　内容語素性の抽出

構文解析結果から、述語項 *PA* に対して、*PA* が係っている述語、および、係られている述語を素性として抽出する。前者には *post*、後者には *pre* というフラグを述語に付与することによって、両者を区別する。例えば、「ブレーキをかけて、減速した」という文からは述語項「ブレーキをかける」に対して、素性「減速:post」を抽出し、「徐行してブレーキをかけた」という文からは素性「徐行:pre」を抽出する。

図 1 に「ブレーキをかける」のベクトルを表す。このベクトルは頻度ベクトルの値を相互情報量 (MI) にしたものである。

表 1: 機能語となる形態素の品詞・品詞細分類とその表現例

A:同文節に出現	
助動詞	書く**だろう**, 書く**ようだ**, 書く**のだ**
接頭辞	**お**書きになる, **大**満足する, **新**発売する
接尾辞	書いて**いる**, 書いて**おく**, 書いて**くれる**
付属動詞	書くかも**しれ**ない, 書いて**頂く**, 書き**続ける**
形容詞	書いて**ほしい**, 書くと**よい**, 書くしか**ない**
形式名詞	書く**こと**, 書く**もの**だ, 書く**の**は〜
副詞的名詞	書く**とき**, 書いてる**場合**, 書いた**後**
判定詞	書くこと**だ**, 書いてる場合**じゃ**ない, 書き**ます**
B:動詞に係る文節に出現	
副詞	**きっと**書く, **絶対**書く, **直ちに**書く
接続詞	**そして**書く, **しかし**書かなかった, **だから**書く
感動詞	**さあ**書こう, **じゃあ**書こう, **ねえ**書かないの
指示詞	**そこ**に書いて, **どのように**書くの, **これ**で書く
C:両方に出現	
助詞	書くの**を**〜, 書くの**かな**, 彼**は**書く

表 2: 機能語の抽出方法

	F1	F2	F3	F4
A グループ	同	同	同・係	-
B グループ	係	-	係	副詞のみ
C グループ	同	同	同・係	-

4.2　機能語素性の抽出

本研究では、機能語として表 1 に示す品詞および品詞細分類の語を選択する。動詞に関係のあると思われる機能語のみを抽出するため、抽出する対象は動詞と同じ文節中に現れるものまたは動詞に係る文節中に現れるものに限定する。また格標識「が」、「を」、「に」は機能表現として抽出しない。例えば、「子供が飛び出してきたので咄嗟にブレーキをかけたそうだ」という文からは述語項「ブレーキをかける」に対して、素性「きた_接尾辞」、「ので_助詞」、「咄嗟に_副詞」、「そうだ_助動詞」を抽出する。

図 2 に「ブレーキをかける」のベクトルを示す。内容語素性のベクトルと同じように頻度ベクトルの値を相互情報量にしたものである。

また、機能語の抽出方法が判定結果に与える影響を調べるため、表 2 に示す F1〜F4 の 4 通りの方法で実験を行った。表中の「同」は動詞と同文節中に現れる

咄嗟に　損なう　･･･　途端　思い切って　･･･
(7.1,　6.4,　･･･　4.2,　4.1,　･･･)
図 2: 「ブレーキをかける」の機能語ベクトル

場合に抽出することを、「係」は動詞に係る文節中に現れる場合に抽出することを、「-」は抽出しないことを表す。

5　実験

Web テキスト 16 億文を形態素解析器 JUMAN・構文解析器 KNP で解析し、その結果から内容語および機能語の素性ベクトルを構築した。また内容語素性に関しては、述語項の類似度計算だけでなく、ベースラインとして述語単体の類似度計算も行なう。この場合の素性は係り受け関係にある格要素とした。

提案手法の有効性を確認するために評価セットを構築し、実験を行った。

5.1　評価セット

述部の同義性判定の評価を行うための評価セットを構築した。まず、ブログ 810 万文から、名詞とその名詞に係っている述部をペアで抽出した。名詞は日本語語彙大系の名詞意味属性である「具体名詞」に属し、かつ頻度 10 以上出現するものに限定した。作業者は、「名詞-述部（例: 本-出版する）」のペアに対して、名詞が同じで述部が異なる他のインスタンスから、同義・含意・反意・その他という 4 つの関係に属するペア (e.g., 本-出す（同義）) をそれぞれ選出した。作業者が作成したペアに対し、評価者が指針と合っているかどうかを判断し、合っていない場合は、作業者と話し合い、2 人が合意したペアを新たに選出した。本論文では、上記で抽出されたデータに対し、適切な格助詞を後から挿入したものを使用する。また、述部は動詞に限定する。

評価セットの例を表 3 に示す。同義、含意、反意、その他のペアの数はそれぞれ 338、259、181、208 であった。評価セットの中には「信号-ヲ-渡った」と「信号-ニ-ひっかかった」(反意) のように格の異なるものも存在する。

表 3: 評価セットの例

タイプ	名詞-格-述部 1	名詞-格-述部 2
同義	本-ヲ-出す	本-ヲ-出版する
	信号-ガ-光る	信号-ガ-点灯する
含意	本-ヲ-即買する	本-ヲ-購入する
	信号-ニ-間に合った	信号-ヲ-渡った
反意	本-ヲ-返却する	本-ヲ-借りる
	信号-ヲ-渡った	信号-ニ-ひっかかった
その他	本-ヲ-出す	本-ヲ-買う
	信号-ヲ-見る	信号-ガ-光る

5.2　結果と考察

同義とみなす類似度の閾値を 0.01, 0.03, 0.05, 0.1, 0.15, 0.2, ..., 0.35 と動かし、F 値が最大となる閾値を求めた。類似度尺度が複数ある場合はすべての閾値の組み合わせを試した。複数の類似度尺度を用いる場合は OR をとった。すなわち、少なくとも一つの類似度尺度がその閾値を超えている場合にシステムは同義とみなす。

実験結果を表 4 に示す。ここで、タイプ反意の中には例えば「買った」と「売った」のように国語辞典から抽出した反意関係を用いることにより、反意関係を認識できるものがある。それらは分布類似度で反意と認識する必要はないことから、それらは反意タイプとみなした。

内容語のみを用いる場合は、述語と述語項を組み合わせた場合に F 値が最も高いことがわかる。機能語のみを用いる場合は、F3 が最も F 値が高いことがわかる。内容語と機能語の両方を用いる場合は、述語と述語項と F4 の組み合わせた場合に F 値が最も高くなった[1]。

機能語のみを用いる場合、F4 の F 値が最も低かったが、これは素性に用いることができる出現が他よりも圧倒的に小さく、データスパースネスのためであると考えられる。しかし、内容語と組み合わせて用いると最も F 値が高くなった。

機能語を用いることで精度が向上した例を表 5 に示す。「幕-ヲ-上げる」と「幕-ヲ-切る」は内容語素性では同義と判定できなかったが、機能語素性では同義と判定することができた。内容語素性を用いた場合、共通の素性としては「演奏」や「登場」くらいしかないため類似度が低いが、機能語素性の場合は「こうして」、「正に」、「そして」、「共に」などが共通の素性としてあり類似度が高くなっている。また、「棚-ニ-並べる」と「棚-ヲ-片付ける」は内容語素性では同義と判定してしまったが、機能語素性では同義でないと判定することができた。内容語素性では、共通の素性として「拭く」、「整理」、「スッキリする」などがあり類似度が高くなっているが、機能語素性では「棚-ニ-並べる」に特徴的な素性が「重ねて」、「ズラリ」、「雑然」などであるのに対し「棚- ヲ-片付ける」に特徴的な素性は「ごそごそ」、「ちょっと」、「あげる」などであり

[1] それ以外の組み合わせの精度は紙面の都合上、割愛する。

表 4: 実験結果 (正例: 同義/含意, 負例: 反意/その他)

		Precision	Recall	F	閾値
内容語	述語	0.602 (580/963)	0.972 (580/597)	0.744	述語:0.01
	述語項	0.661 (436/660)	0.730 (436/597)	0.694	述語項:0.03
	述語+述語項	0.618 (573/927)	0.960 (573/597)	0.752	述語:0.05, 述語項:0.05
機能語	F1	0.623 (561/901)	0.940 (561/597)	0.749	F1:0.01
	F2	0.622 (565/908)	0.946 (565/597)	0.751	F2:0.01
	F3	0.618 (580/939)	0.972 (580/597)	0.755	F3:0.01
	F4	0.672 (403/600)	0.675 (403/597)	0.673	F4:0.01
内容語 +機能語	述語+述語項 +F3	0.654 (562/859)	0.941 (562/597)	0.772	述語:0.1, 述語項:0.15, F3:0.2
	述語+述語項 +F4	0.671 (548/817)	0.918 (548/597)	**0.775**	述語:0.05, 述語項:0.15, F4:0.1

表 5: 機能語を用いることで精度が向上した例

「幕-ヲ-上げる」と「幕-ヲ-切る」(同義)

	類似度	共通の素性
内容語	0.030	演奏:post, 披露:post, 登場:pre
機能語	0.336	こうして_指示詞, 正に_副詞, そして_接続詞, ...

「棚-ニ-並べる」と「棚-ヲ-片付ける」 (その他)

	類似度	共通の素性
内容語	0.301	拭く:pre, 整理:post, スッキリする:post, ...
機能語	0.060	とりあえず_副詞, つ_接尾辞

類似度が低くなった。

6 おわりに

本論文では、内容語および機能語との共起分布に基づき、述部の同義判定を行う手法を述べた。今後の課題としては内容語素性と機能語素性の統合方法の検討や、同義判定結果のアプリケーションでの利用などがあげられる。

参考文献

[1] James R. Curran and Marc Moens. Improvements in automatic thesaurus extraction. In *Proceedings of the Workshop of the ACL Special Interest Group on the Lexicon (SIGLEX)*, pp. 59–66, 2002.

[2] Dekang Lin and Patrick Pantel. Concept discovery from text. In *Proceedings of Conference on Computational Linguistics(COLING 2002)*, pp. 577–583, 2002.

[3] Lin Sun, Anna Korhonen, Thierry Poibeau, and Cedric Messiant. Investigating the cross-linguistic potential of verbnet -style classification. In *Proceedings of the 23rd International Conference on Computational Linguistics(COLING2010)*, pp. 1056–1064, 2010.

[4] Idan Szpektor and Ido Dagan. Learning entailment rules for unary templates. In *Proceedings of the 22nd International Conference on Computational Linguistics(COLING2008)*, pp. 849–856, 2008.

[5] 相澤彰子. 大規模テキストコーパスを用いた語の類似度計算に関する 考察. 情報処理学会論文誌, Vol. 49, No. 3, pp. 1426–1436, 2008.

[6] 柴田知秀, 黒橋禎夫. 超大規模ウェブコーパスを用いた分布類似度計算. 言語処理学会 第 15 回年次大会, pp. 705–708, 2009.3.

[7] 柴田知秀, 黒橋禎夫. 文脈に依存した述語の同義関係獲得. 情報処理学会 第 199 回自然言語処理研究会, 2010.11.19.

[8] 竹内孔一, 乾健太郎, 竹内奈央, 藤田篤. 意味の包含関係に基づく動詞項構造の細分類. 言語処理学会第 14 回年次大会, pp. 1037–1040, 2008.