述部意味関係コーパスの構築

泉 朋子† 柴田 知秀‡ 浅野 久子† 松尾 義博† 黒橋 禎夫‡
†日本電信電話株式会社　NTT メディアインテリジェンス研究所
{izumi.tomoko, asano.hisako, matsuo.yoshihiro}@lab.ntt.co.jp
‡ 京都大学大学院　情報学研究科
{shibata, kuro}@i.kyoto-u.ac.jp

1 はじめに

単語やフレーズ間の同義・反義関係を認識することは，言語処理において，最も基本的かつ重要なタスクのひとつである. QA システムや, 情報抽出, テキストマイニングなど上位の言語処理システムにおいて，これら単語やフレーズ間の意味関係が認識できれば，システム全体の精度を向上させることできる．例えば，「(信頼性に)欠ける」と「(信頼性に)乏しい」といった述部の同義関係を正しく認識出来れば，テキストマイニングの集計・分析の精度を大きく向上させることが可能である．また，「物騒だ」と「安全だ」のような反義関係を理解することが出来れば，意見マイニングにおいて相反する意見を分析したり，対話システムにおいて，システムが回答する内容がそれ以前に回答した内容と矛盾していないか確認することが出来る.

本稿では，これら述部の同義・反義認識のために必要な述部意味関係コーパスの構築について述べる．この述部意味関係コーパスは，「信頼性に-乏しい」と「信頼性に-欠ける」といった述語項のペア（約 7,300 ペア）に対し，人手で「同義」「含意」「反義」「無関係」という 4 つの意味関係が付与されている．さらに，反義関係については，「反対の観点」によって 3 つの詳細カテゴリが付与されている．1 つ目が，「多い」と「少ない」のような属性の違いを表す「属性反義」, 2 つ目が「入学」と「卒業」のように，動作の始点と終点となりうる（すなわち，時間的経過の関係になりうる）「経時反義」, 3 つ目が「売る」と「買う」のように，単体では真逆の動作であるが，格構造を交代させることで同義になりうる「視点反義」である.

また，これら作成した述部意味関係コーパスを教師データとして用いた，述部の同義性判定の実験結果を述べる.

2 既存研究

単語の意味関係を付与した大規模言語リソースとして, WordNet (Miller, 1995)がある. WordNet では，単語間の意味関係として，同義(synsets)，上位・下位，反義等の情報が付与されている．日本語 WordNet (Bond et al., 2009)にも，単語の意味関係が付与されているが，現状版(Japanese WordNet 1.1)は同義と上位下位の関係のみ付与されている.

相澤(2008)は，「A-や-B-などの-C」といったパターンを用いて，自動で同義語（i.e., A と B）と上位下位語（i.e., C は A 及び B の上位語）の評価セットを構築した．しかし，相澤(2008)が構築した評価セットは名詞句を対象にしたものであり，本稿が対象とする述部は含まれていない.

Mitchell & Lapata (2010)は，形容詞-名詞(e.g., vast-amount vs. large-quantity)，名詞-名詞(e.g., telephone-number vs. phone-call)，動詞-名詞(e.g., start-work vs. begin-carrer)という 3 種類のフレーズ間の意味の関連度を人手で付与したコーパスを作成した．しかし，関連度は 7 段階のスケールで付与されているため，その関連性が同義を表すのか，反義を表すものかの情報は付与されていない.

3 述部意味関係コーパスの構築

2 節で述べたように，既存の言語リソースを，述部の意味関係認識の評価セットとして用いることは困難である．そこで，述部の意味関係を計算機で認識するために必要な，述部意味関係コーパスを構築した.

この述部意味関係コーパスは,「信頼性に-乏しい」と「信頼性に-欠ける」のような述語項のペア（約 7,300 ペア）を対象に，人手で「同義」「含意」「反義」「無関係」という 4 つの意味関係を付与している．述部は組み合わさる項によって意味が異なるため，項とペアにすることで，述部の意味を明確にしている．また，項を入れることで,「文脈によって同義・反義となりうる述部」をデータに加えるようにした．コーパスの内訳と例を Table 1 に示す.

意味関係		エントリ数	例
同義		**3,188**	信頼性-に-乏しい vs. 信頼性-に-欠ける 調子-を-落とす vs. 調子-が-悪い
含意		**1,557**	英語-が-堪能 vs. 英語-を-話す 見た目-が-幼い vs. 見た目-が-若い
反義		**716**	
	属性反義	426	命令-を-受ける vs. 命令-を-拒否
	経時反義	131	電話-を-かける vs. 電話-を-切る
	視点反義	159	サイン-を-求める vs. サイン-に-応じる
無関係		**1,817**	電話-を-繋ぐ vs. 電話-が-殺到 評価-が-低い vs. 評価-が-進む
合計		**7,278**	

Table 1: 述部意味関係コーパスの内訳と例

述部意味関係

述語項ペアに付与する意味関係として，「同義」，「含意」，「反義」，「無関係」という 4 種類の関係を付与した．「反義」に関しては，「反対の観点」によって下記の 3 つの詳細カテゴリを付与した．

(1) 属性反義
　　多い　vs.　少ない
(2) 経時反義
　　入学　vs.　卒業
(3) 視点反義
　　売る　vs.　買う

(1)は，述部が表す属性や，動作が真逆の方向に向いている反義である（属性反義）．(2)は，動作の始点と終点になりうる（すなわち，過去-未来の関係になりうる）関係を表している（経時反義）．(3)は，項が同じ場合は反対の動作を表すが，項を入れ替えることで同義になりうる関係を表している（視点反義）．

これら反義関係の分類は，我々が独自に構築したものであるが，これらの分類は自然言語処理にとって重要であると考える．例えば，視点反義は，述部単体では反義であるが，項の入れ替えにより同義になりうるため，言い換え研究において必須の知識である．また，反義知識は矛盾検出などにも使われているが(Harabagiu et al., 2006)，経時反義は，過去-未来の関係を表しているため，これらの反義関係は，必ずしも矛盾を引き起こすわけではない（例えば，「2003 年に入学し，2005 年に卒業した.」は「入学」と「卒業」という反義関係の単語を同一文に含んでいるが，矛盾はしていない）．このように，「反対の観点」の詳細な分類が，後段の処理にとって重要かつ有益な情報になるため，これらの詳細カテゴリを付与した．

述部の抽出

コーパス作成にあたり，日本語 wikipedia[1]から述部を抽出した．Wikipedia から高頻度に出現した名詞句を，上位から 5 つ飛ばしで 100 個抽出し，それぞれの名詞句に対して，相互情報量が高い上位 20 件の述部を抽出した．

アノテーション

アノテーションは，言語学の知識のある 3 人のアノテータが実施した．下記が，アノテータに提示した意味関係の定義である．また，ラベル付与の一貫性を担保すべく Chierchia and McConnel-Ginet (2000)をもとに，言語テストを作成した．

・同義

定義：2 つの述部が同じ出来事を表している
言語テスト：片方の述部を否定すると，意味が通じない
例：×「土産を買った．***でも，***（その土産を）購入した***という訳ではない.***」

・含意

定義：どちらか一方の述部がもう一方の述部の意味を包含していること
言語テスト：含意されている述部を否定することができない
例：×「土産を衝動買いした．でも，（その土産を）買った***という訳ではない.***」
○「土産を買った．でも，衝動買いした***という訳ではない.***」

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

素性の種類		説明
同義 識別 素性	**辞書定義文**	-述部ペアに対し，相手の定義文内に現れているか否かの 2 値素性 -述部ペアの辞書定義文内の語彙の重なり（内容語の重なり実数値）
	用言属性	-述部ペアが共通して保持している用言属性 -述部ペアが保持している用言属性の重み付重なり度（より詳細な属性に重みづけ）
	分布類似度	-述部ペアに対し，述部の内容語を単位とした分布類似度 -述部ペアに対し，「項-述部」を単位とした分布類似度
	モダリティ・否定	-述部ペアが共通して保持している機能表現 -述部ペアの機能表現の重なり度
反義 識別 素性	**複合語・「たり」構文**	-複合語の Web の出現頻度と Ngram スコア(google ngram を使ったことを脚注に) -述部ペアを「たり」で接続した「たり」構文の Ngram スコア
	接頭・接尾辞	-接頭・接尾辞の組み合わせ（表層文字列） -接頭・接尾辞の組み合わせ文字列の Web 出現頻度と Ngram スコア
品詞		-述部の品詞

Table 2: 同義性判定に用いる素性一覧

・反義

定義：2 つの述部が真であることが成立しない

言語テスト：両方の述部を「でも」でつなげると，意味が矛盾する

例：#「土産が多い．**でも，**（その土産が）少ない．」

反義の詳細カテゴリに関しては，下記の定義を用いた．

・属性反義

定義：述部が表す属性が真逆であったり，動作が真逆の方向に向いている．

・経時反義

定義：動作の起点と終点を表す．動作の起点・終点の関係を表す反義（動作の繰り返しも含む）はある程度**必然性**があるものにする．

・視点反義

定義：格構造が全く同じだと真逆の意味を表すが，格を交代することで**同義**になる．

4　述部の同義性判定

作成したコーパスを訓練データとして，教師あり学習を用いた述部の同義性判定を行った[2]．これは，本稿で作成した述部意味関係コーパスを訓練データとして利用できるものかを考察するとともに，今後，同義の述部コーパスの自動獲得が可能かどうかを検討するために行った．

同義性を判定するための特徴として，「同義らしさ（同義識別素性）」と「反義らしさ（反義識別素性）」という 2 種類の言語的特徴を用いた．これらの特徴を，Table 2 にまとめる．

4.1　同義らしさを表す言語的特徴（泉 et al., 2013）

・辞書定義文の重なり

2 つの述部が同じ意味を表す場合，片方の述部がもう一方の述部の辞書定義文に出現する傾向がある．例えば，「出来上がる」の定義文に「すっかりできる．**完成する**．」という同義の述部が出現する．この辞書定義文における定義の相互補完性，及び定義文同士の語の重なりを第一の特徴として用いる．

・用言属性

2 つの述部が同義の場合，それら述部の抽象的な意味クラス（用言属性）も共通している．そこで，日本語語彙大系（池原他, 1999）の用言属性を用いて，述部の用言属性の重なりを特徴として用いる．

・分布類似度（柴田・黒橋, 2010）

69 億文から計算した，述語項及び述語間の分布類似度を素性として用いる．

・モダリティ・否定

述部の否定表現やモダリティ表現の重なりを素性として用いる．

4.2 反義らしさを表す言語的特徴

4.1 であげた同義らしさを表す素性は，同時に反義関係の述部にも当てはまる傾向がある．たとえば，分布類似度などは，同義関係も反義関係も高いスコアを出す傾向がある．そこで，同義関係と反義関係を正しく識別するために，「反義らしさ」を表す言語的特徴を加える．

[2] 述部の同義性判定では，「含意」関係も「同義」として扱った．これは，「同じ事を表している表現を認識する」というアプリケーションでの有用性を指向した場合，「同義」も「含意」も同様の性質を保持しているためである．

	Precision	Recall	F-Score
Baseline-WordNet	0.977	0.349	0.514
提案手法	0.899	0.932	**0.915**
同義識別素性のみ	0.730	0.917	0.812
反義識別素性のみ	0.721	0.972	0.828

Table 3: 実験結果

・複合語と「たり」構文

反義の単語同士は複合語を作りやすいという傾向がある(e.g., 売り買い). また,「売ったり買ったり」のように，対象を表す「たり」構文に出現しやすい．そこで，複合語と「たり」構文への出現のしやすさを「反義らしさ」を表す特徴として用いる．

・接頭・接尾辞の組み合わせ

「入学」の「入」と「卒業」の「卒」のような部分文字列（接頭・接尾辞と呼ぶ）が反義関係を表現する傾向がある．そこで，これらの接頭・接尾辞の組み合わせを素性として用いる．

5　実験・考察

3 節で作成したコーパスを訓練データとし，述部の同義性判定実験を行った．モデルのトレーニングには，LIBSVM(Chang & Lin, 2011)を用いて，5 分割交差検定の平均値で評価した．比較手法として，日本語 WordNet を用いて，入力の述部ペアが synsets にあれば，「同義」という評価を行った．Table 3 に実験の結果を示す．

Table 3 が示すように，提案手法は，0.92 という高い F 値を達成することが出来た．下記が，WordNet の synsets にエントリが存在せず，本提案手法でのみ正しく識別できた同義の述部である．

(3) デビュー-に-至る vs. デビュー-を-迎える
(4) 状況-を-踏まえる vs. 状況-を-見る

また，「同義らしさ」を表す特徴に「反義らしさ」を表す特徴を加えることで，全体の F 値が大幅に向上した．

エラー分析の結果，提案手法では「気に病む」と「気が弱い」のような，慣用的表現の同義性を判定できないことが分かった．これは，慣用表現の意味にはメタファーといった複雑な意味解釈が必要であり，提案手法の素性だけではそれらを表現することが出来なかったのが原因である．慣用表現の同義性判定に関しては，今後の課題としたい．

6　結論

本稿では，日本語の述部意味関係コーパスの構築について述べた．このコーパスには，述語項のペアに対し，同義・含意・反義・無関係という 4 つの意味関係が人手で付与されており，反義関係に関しては，さらに反対の観点によって 3 つの詳細カテゴリが付与されている．

述部の同義性判定実験では，このコーパスを訓練データとして用いることで高い精度で同義の述部を認識することができ，コーパスの有用性が確認できた．

本稿で構築した述部意味関係コーパスは無償で公開している[3]．今後は，これらのコーパスを用いて，述部意味解析研究の加速を目指すとともに，今後もコーパスの規模を拡大していく予定である．

References

相澤 彰子(2008). 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,* 1426-1436.

Bond F, Isahara H, Fujita S, Uchimoto K, Kuribayashi T and Kanzaki K (2009). Enhancing the Japanese WordNet, *The 7th Workshop on Asian Language Resources, in conjunction with ACL-IJCNLP 2009,* Singapore.

Chierchia, G., and McConnell-Ginet, S. (2000). *Meaning and grammar: An introduction to semantics (2nd ed.).* Cambridge, MA: The MIT press.

Chang, C-C., and Lin, C-J. (2011). LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology (TIST), 2(3), No.27.*

Harabagiu, S., Hickl, A., and Lacatusu, F. (2006). Negation, contrast, and contradiction in text processing. *Proceedings of the 23rd National Conference on Artificial Intelligence (AAAI-06).*

池原悟，宮崎正弘，白井諭，横尾昭男，中岩浩巳，小倉健太郎，大山芳史，林良彦（1999）. *日本語語彙大系 CD-ROM 版.* 岩波書店.

泉朋子・ 柴田知秀・齋藤邦子・松尾義博・黒橋禎夫(2013). 複数の言語的特徴を用いた日本語述部の同義判定. *自然言語処理, Vol. 20, 4,* 539-561.

Mitchell J., and Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive Science, Vol. 34, 8,* 1388-1429.

Miller, G. A. (1995). WordNet: A lexical database for English, *Communications of the ACM, Vol. 38, 11,* 39-41.

柴田知秀，黒橋禎夫（2010）. 文脈に依存した述語の同義関係獲得. *情報処理学会研究報告(IPSJ SIG Technical Report),* 1-6.

[3] http://nlp.ist.i.kyoto-u.ac.jp/