項共有述語項の意味関係コーパスの整備および同義・反義性判定

澤田 晋之介† 柴田 知秀†‡ 黒橋 禎夫†‡
† 京都大学 ‡ 科学技術振興機構 CREST
†‡ {sawada, shibata, kuro}@nlp.ist.i.kyoto-u.ac.jp

1 はじめに

単語やフレーズ間の同義を判定することは、言語処理において基本的かつ重要なタスクのひとつである。情報検索、情報抽出、意見集約などの言語処理システムにおいて、単語やフレーズ間の同義を判定することができれば、システム全体の精度を向上させることができる。例えば、ある銀行の HP にある対話ボットでは、「口座を開設」と問いかけると口座を開設することに関する情報が応答として返ってくるが、「口座を開く」だと関係の無い応答となってしまう。「口座を開設」と「口座を開く」が同義だとわかっていれば、「口座を開く」という問いかけに対しても正しく応答することができる。

本研究では述語と項の組 (以後、述語項と呼ぶ) のペア、その中でも特に項を共有している述語項を対象として同義判定を行う。例えば、「洗濯機を回す」と「洗濯機を使う」という述語項のペアについて、「回す」と「使う」は同義ではないが、「洗濯機を回す」と「洗濯機を使う」は同義であることからわかる通り、述語単体で扱っていたのでは正確な同義判定は達成できない。そこで、述語項を単位として扱うことでより正確な同義判定を行う。

単語の同義判定を行う手法として、コーパスから単語の分散表現を学習し、その類似度を用いるものがある。この手法の代表的なものに Mikolov らの word2vec[4] や Levy らによるその改良 [2] がある。これらの手法は、「文脈の似ている語は意味も似ている」という考えに基づいている。しかし、「畳む」と「広げる」のような反義関係にある語は似た文脈で出現しやすく、同義だと判定してしまう問題があることが知られている。この問題に対しても、「傘を畳む」と「傘を広げる」のように述語項を単位として扱うことで下記の例の「扉を開けた」と「干す」のように文脈に差ができ、うまく学習できるようになることが期待される。

(1) 傘を畳み、扉をがらんがらんと音を立てて開けた
(2) 家に帰ると、傘を広げて、干す。

また、泉ら [7] のように述語中の漢字の組み合わせなどの言語的特徴を用いて教師あり学習を行うことで、より良い同義判定を行うことができる。

本研究では泉らの構築したコーパスを用いて同義・反義判定を行った。しかし、含意関係には「議員に当選」と「議員に初当選」のような上位下位で同義に近いものと、「原因を探る」と「原因を突き止める」のような時間経過や前提といった無関係に近いものが混在していた。そこで含意関係を細分類し、上位下位関係を同義として、時間経過や前提を無関係として扱うこととした。また、分類を同義、反義、無関係の 3 値と簡易化したタスクをクラウドソーシングで実施することによって短時間でアノテーションを行えることを確認した。

2 関連研究

柴田ら [6] は述語項を単位として分布類似度を計算することで「洗濯機を回す」と「洗濯機を使う」のような文脈に依存して同義となる述語項を獲得する手法を提案している。

泉ら [7] は本研究で用いた述部意味関係コーパスを構築し、また同コーパスを用いて同義判定タスクに取り組んでいる。同義判定タスクでは柴田らの提案した述語項を単位とする分布類似度に加えて、言語的特徴を利用した素性を用いて教師あり学習を行うことで高い精度を達成している。

単語を単位として分散表現を学習する手法は Mikolov ら [4] が提案している。Mikolov らの手法の改良として、依存関係を利用して単語の分散表現を学習する手法を Levy ら [2] が提案している。これらの手法は、学習対象の単語と文脈中の単語の 1-hot ベクトルをそれぞれ入力、出力とする 2 層のニューラルネット

Copyright(C) 2018 The Association for Natural Language Processing.
All Rights Reserved.

ワークを大規模コーパスで学習することで 1 層目が単語のベクトルを表すようになるという手法を、高速化のために近似したものである。文脈中の単語が同じ単語は同じ出力を教師データとして与えることになり、1 層目の出力、つまり単語の分散表現が近くなる。

また、フレーズに対する分散表現の学習手法を Liu ら [3] が提案している。Liu らの手法はイディオムの言語的特徴を考慮してニューラルネットワークを設計している。Liu らの注目したイディオムの言語的特徴は、イディオムと他の単語を区別するのが難しいこと、イディオム中の単語からイディオムの意味を構成できないこと、表現に揺らぎがあること、の 3 つである。イディオムを豊富に含む感情分析データセットも構築しており、そのデータセットにおいて比較手法よりも良い精度を達成している。

表現の構成性を考慮した分散表現の学習手法を Hashimoto ら [1] が提案している。Hashimoto らは学習した分散表現を構成性の判定と曖昧性解消で評価しており、同義・反義判定は行っていない。

クラウドソーシングによる言語資源の構築は最近では一般的になってきているが、その先駆的なものは Snow らの研究である [5]。Snow らは感情解析、単語類似度、含意認識、イベントの時間関係 、語義曖昧性解消の 5 つのアノテーションタスクをクラウドソーシングによって行った。非専門家によるアノテーションであっても、複数人のアノテーションをうまく統合することによって専門家によるものに比肩しうる精度のアノテーションが可能であることを示した。

3 意味関係コーパス

3.1 意味関係の分類

二つの述語項に対して、その意味関係を表 1 の 8 つに分類する。また、含意についてはその包含の向きを分類することとする。泉ら [7] は含意について細分類を考えていなかったが、結果として含意には上位下位のように同義に近いものと時間経過や前提のように無関係に近いものが混在してしまっていた。そこで含意を 3 つに細分類することで、同義・反義判定の際に使いやすい分類とした。

3.2 コーパスの整備

泉らの構築した述部意味関係コーパス [7] を元に、含意関係の細分類を行った。また、その際に誤りと思われるアノテーションの修正も行った。アノテーションの修正の際にはコーパス構築の際に用いられた言語テストを利用することでアノテーションの一貫性を保った。アノテーションは、言語学の知識のある人間が行った。アノテーション結果の分布を表 1 に示す。

表 1: 意味関係の分類と例、データ数

分類	細分類	例	データ数
同義	同義	本を買う vs. 本を購入	4804
含意	上位下位	相手を殴る vs. 相手を攻撃	703
	時間経過	原因を探る vs. 原因を突き止める	368
	前提	音楽を聞く vs. 音楽を楽しむ	480
反義	属性反義	重心が高い vs. 重心が低い	596
	視点反義	車を売る vs. 車を買う	222
	経時反義	電車に乗る vs. 電車から降りる	201
無関係	無関係	飲料を指す vs. 飲料を輸送	2620
合計			9994

3.2.1 定義および言語テスト

以下にアノテータに示した定義と言語テストを記す。アノテーションの一貫性を担保するため、含意の細分類を除いて泉ら [7] が用いたものと同じ定義、言語テストを用いた。

同義

定義: 二つの述語項が同じ出来事を表している
言語テスト: 片方の述語項を否定すると、意味が通じない

含意

定義: どちらか一方の述語項がもう一方の述語項の意味を包含している
言語テスト: 包含されている述語項を否定すると意味が通じない
含意の細分類については、以下の定義を用いた。

上位下位

定義: 表す概念が上位下位関係にあったり、情報の過不足があるもの

時間経過

定義: 時間経過を伴う前提的なもの

前提

定義: 時間経過を伴わない前提的なもの

反義

定義: 二つの述語項が同時に真であることが成立しない
言語テスト: 両方の述語項を「でも」でつなげると、意味が矛盾する
反義の細分類については、以下の定義を用いた。

属性反義

定義: 述語項の表す属性が真逆であったり、動作が真逆の方向を向いている

視点反義

Copyright(C) 2018 The Association for Natural Language Processing.
All Rights Reserved.

定義: 格構造が全く同じだと真逆の意味を表すが、格を交替することで同義になる

経時反義

定義: 動作の起点と終点を表す

3.3 クラウドソーシングによる拡張

近年、クラウドソーシングによる大規模な言語資源の構築が行われている。そこで、述語項のペアを見て同義、反義、無関係の 3 値に分類するクラウドソーシングを実施することで、大規模なアノテーションを行うことを考える。PMI が高い述語と項の組み合わせはその意味が非構成的であるという仮定の下、Web コーパスで頻度の高い述語項、かつ述語と項の PMI が高いものを対象として選択した。1000 件の述語項ペアについてアノテーションを行う小規模な実施で有効性を確かめたところ、下記例のように良い結果が得られた。今後、大規模にクラウドソーシングを実施することでコーパスを拡張する予定である。

(3) 許可を取り消される vs. 許可を取る（反義）
(4) 前方に見えてくる vs. 前方にある（同義）
(5) 花粉症に悩まされる vs. 花粉症になる（同義）

4 分散表現獲得

文脈によって語義の変わる述語を扱うために述語項を単位とし、Levy らに基づき係り受け関係にある語を文脈として分散表現を学習する。学習対象は学習に使用するコーパスにおいて 5 回以上出現する述語および述語項とし、文脈については項（名詞＋格）、述語および Web コーパスにおける頻度上位 20 万以内の述語項とする。例えば、図 1 に示した例において、「目に入る」という述語項の文脈は「見て」、「通りを目指す」、「目指す」、「アパートが」の 4 つとなる。

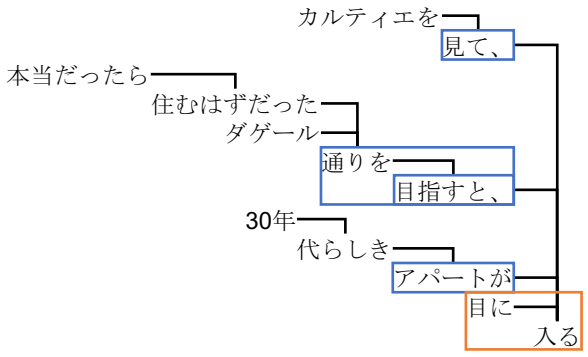

図 1: 学習に使用する文脈の例

5 同義・反義性判定

5.1 問題設定

整備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同義・反義性判定を行った。分類は同義、反義、無関係の 3 クラスとし、含意ラベルのデータについては上位下位を同義に、時間経過と前提を無関係にマージした。

5.2 ヴォイス・否定の正規化

同じ項と述語であっても下記例のようにヴォイスや否定表現により同義と反義の反転する例がある。

(6) ケーキを贈る vs. ケーキをプレゼントする（同義）
(7) ケーキを贈る vs. ケーキをプレゼントされる（反義）

このような例を機械学習で解くには、素性の組み合わせをうまく学習しなければならず、難しい。そこで、ヴォイスや否定表現を正規化したものを機械学習の入力とすることで対処する。

5.3 素性

泉ら [7] の素性をベースに有効と思われる素性を用いた。詳しくは以下の通りである。

- 分散表現のコサイン類似度
  76 億文から学習した分散表現を用いた。述語間、述語項間、述語—述語項間の類似度の最大値を素性として用いた。また、漢字の分散表現を 1 億文から学習し、述語中の先頭と末尾の漢字の類似度を素性として用いた。
- 先頭漢字、末尾漢字の組み合わせ
  「入学」の「入」と「卒業」の「卒」のような述語の先頭や末尾に位置する漢字（先頭漢字、末尾漢字と呼ぶ）が反義関係を表現する傾向がある。そこで、これらの先頭漢字、末尾漢字の組み合わせを素性として用いる。
- 述語が漢字を共有しているかどうか
  例えば「断定」と「判断」のように、位置にかかわらず同じ漢字を共有していることが同義関係を表現する傾向がある。そこで、漢字の共有情報を素性として用いた。
- 一方の述語がもう一方の述語の辞書定義文に出現するかどうか
  2 つの述語が同じ意味を表す場合に、片方の述語の辞書定義文にもう一方の述語が出現する傾向があるため、素性として用いた。

Copyright(C) 2018 The Association for Natural Language Processing.
All Rights Reserved.

表 2: 実験結果 (F1)

	同義	反義	無関係	micro avg.	macro avg.
Majority Baseline	0.711	0.000	0.000	0.551	0.237
コサイン類似度素性のみ	0.774	0.054	0.694	0.692	0.507
提案手法	**0.839**	**0.641**	0.755	**0.791**	**0.745**
– ヴォイス・否定の正規化	0.838	0.590	**0.758**	0.789	0.729

- WordNet の同一 synset にあるかどうか
  WordNet は人手で整備された辞書として信頼の置ける情報であるため、素性として用いた。
- 分類語彙表の類・部門・中項目・分類項目
  分類語彙表は人手で整備されたシソーラスであり、同じ項目に属する場合には無関係でない傾向があるため、素性として用いた。
- Web コーパスにおける「たり」構文での出現頻度
  反義の単語同士は「売ったり買ったり」のように「たり」構文に出現しやすい。そこで、Web コーパスにおける「たり」構文への出現頻度を素性として用いた。

6 実験・考察

整備したコーパスを訓練データとし、項共有述語項の同義・反義性判定を行った。モデルのトレーニングには scikit-learn の線形 SVM を用いて、5 分割交差検定の平均値で評価した。多クラス分類の手法には one-vs-rest 法を用いた。結果を表 2 に示す。コサイン類似度素性に他の素性を加えることで大きな精度向上を達成している。また、ヴォイス・否定の正規化を行うことで特に反義クラスについて精度向上を達成している。

下記が、ヴォイス・否定の正規化をすることによって正しく判定できた例である。

(8) ケーキを贈る vs. ケーキをプレゼントされる（反義）

(9) レーダーが捕らえる vs. レーダーに写らない（反義）

泉らはコーパス構築時に同義+含意 vs. 反義+無関係の 2 クラス分類の実験をしており、同義+含意の F1 スコアが 91.5%という精度を報告している。しかし、泉らの実験はコーパス構築途中で行われていること、我々は含意を細分類した上で 3 クラス分類にしていることから、コーパス、問題設定ともに違うので直接比較できる値ではない。

7 結論

本稿では、日本語の項共有述語項の意味関係コーパスの整備、および同義・反義性判定について述べた。このコーパスには、項共有述語項ペアに対し、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、含意、反義関係についてはさらに 3 つずつの細分類が付与されている。

同義・反義性判定実験では、このコーパスを訓練データとして用い、またヴォイス・否定表現を正規化して扱うことで良い精度で同義・反義関係を判定することができた。

本稿で整備したコーパスは無償で公開予定である。クラウドソーシングの小規模な実施で良い結果が得られたため、今後同様のタスクを大規模に実施してコーパスを拡張する予定である。

参考文献

[1] K. Hashimoto and Y. Tsuruoka. Adaptive joint learning of compositional and non-compositional phrase embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1, pp. 205–215.

[2] O. Levy and Y. Goldberg. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 302–308. Association for Computational Linguistics.

[3] P. Liu, K. Qian, X. Qiu, and X. Huang. Idiom-aware compositional distributed semantics. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1215–1224. Association for Computational Linguistics.

[4] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger eds., *Advances in Neural Information Processing Systems 26*, pp. 3111–3119. Curran Associates, Inc.

[5] R. Snow, Brendan, D. Jurafsky, A. Ng. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 254–263.

[6] 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.

[7] 泉朋子, 柴田知秀, 浅野久子, 松尾義博, 黒橋禎夫. 述部意味関係コーパスの構築. 言語処理学会 第 20 回年次大会, pp. 690–693.

Copyright(C) 2018 The Association for Natural Language Processing.
All Rights Reserved.