

含意関係認識におけるベクトル空間モデルの適用

寺中 元希 斎藤博昭
慶應義塾大学大学院 理工学研究科

{teranaka, hxs}@nak.ics.keio.ac.jp

1 はじめに

含意関係認識は、機械翻訳や質問応答をはじめとする自然言語処理研究に広く共通するタスクである。自然言語処理において含意関係とは、ふたつの文 T (Text) と H (Hypothesis) において、 T が正しい場合に H も正しいと推定できる関係を指す。ここでいう含意関係は、論理学でいう意味論的含意とは異なり、人間が含意であると判断できると言える場合に成り立つものである [11]。含意関係は一般的に含意が成り立つか否かの 2 値分類を行うタスクである。以下に含意関係が成り立つ場合の例を示す。

T: 川端康成は「雪国」などの作品でノーベル文学賞を受賞した。

H: 川端康成は「雪国」の作者である。

含意関係認識タスクは、2005 年の PASCAL ワークショップ [1] を皮切りに注目され始め、日本語においても NTCIR-11 RITE-VAL [5] という評価型ワークショップが 2014 年に行われた。これらのワークショップでは、単語の被覆率などの表層的な素性を用いた機械学習による認識が主な手法となっており、文間の意味の関係に焦点を当てた研究が今後の課題となっている。

本研究では、文または単語のベクトル表現を用いることで含意関係認識における文間の意味へのアプローチを試みる。近年、深層学習によって単語のベクトル表現を学習するモデルとして skip-gram モデル [6] が注目を集めている。Skip-gram モデルによって得られたベクトル表現は、意味類似性の高い単語ほど近くに分布するという特徴がある。また最近では、skip-gram モデルの拡張であり文や文章のベクトル表現を学習するモデルである distributed memory モデル [4] が提案されている。

含意関係認識において、文のベクトル表現に焦点を当てて行われた研究は確認されず、また distributed memory モデルは非常に新しいモデルであり、これを用いた研究はまだ多くない。従って本稿では、上述のベクトル空間モデルを含意関係認識に適用し、文また

は単語のベクトル表現が含意関係認識に有効であるか検討する。また、既存のベクトル空間モデルを用いた場合との比較検討を行う。

2 関連研究

横手ら [10] は、単語間の意味類似度を WordNet と既存手法を用いて計算し、得られた意味類似度を含意関係認識に有効な類似度に変換する方法を提案している。変換後のスコアと H の単語の逆文書頻度を組み合わせたものを特徴ベクトルとし、Support Vector Machines (SVMs) によって認識を行う。意味類似度の算出手法が 1990 年代の手法であり、より有効な類似度を用いることにより認識精度の向上が期待される。

NTCIR-11 RITE-VAL において最良の結果であった Ishii ら [2] の手法は、ロジスティック回帰を用いたものであったが、特徴量のひとつに skip-gram モデルで生成した単語のベクトル表現のコサイン類似度を用いている。この特徴量 $F_{skipgram}$ は以下の式で表される。

$$F_{skipgram} = \frac{\sum_{w_2 \in NM_{H,T}} \max_{w_1 \in T} \{\cos(v(w_1), v(w_2))\}}{|\{w_2 : w_2 \in NM_{H,T}\}|} \quad (1)$$

ここで、 $NM_{H,T}$ は文 H に含まれ文 T に含まれない単語の集合であり、 $v(w)$ は単語 w を skip-gram モデルによってベクトル表現したものである。また、 $\cos(v_1, v_2)$ はベクトル v_1 と v_2 のコサイン類似度であり、 $|\{w_2 : w_2 \in NM_{H,T}\}|$ は集合 $NM_{H,T}$ に含まれる単語の総数である。

Ishii らの研究では、この特徴量の有効性の検証は詳細に行っていない。

3 ベクトル空間モデル

3.1 Bag-of-Words

Bag-of-Words モデルは最もシンプルな文書のベクトル表現方法であり、ベクトルの各次元を一つの単語に対応付け、その単語が文書中に出現する場合は 1、それ以外の場合は 0 をその次元の値とするものである。

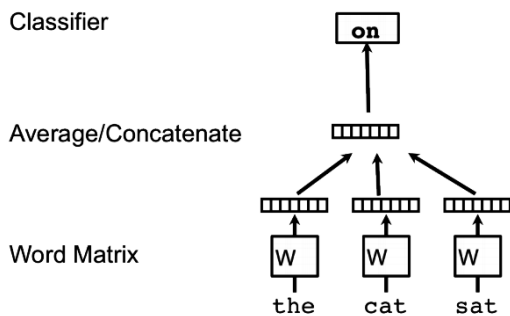


図 1: 単語のベクトル表現の学習 [4]

単語の出現頻度をベクトルの値とする場合もあるが、本稿では 2 値ベクトルのものを Bag-of-Words モデルとして用いる。

3.2 TF-IDF ベクトル

TF-IDF は単語に関する重みの一種であり、TF-IDF を Bag-of-Words のベクトルの値として用いたベクトル表現は、主に情報検索などの分野で用いられる。本稿では便宜上、上記のベクトル表現を TF-IDF ベクトルと呼ぶ。

ある単語 t_i の文書 d_j における TF-IDF の値 $tfidf(t_i, d_j)$ は以下の式で計算される。

$$tfidf(t_i, d_j) = \frac{n_{i,j}}{\sum_k n_{k,j}} \cdot \log \frac{|D|}{|\{d : d \ni t_i\}|} \quad (2)$$

ここで、 $n_{i,j}$ は t_i の d_j における出現回数、 $|D|$ は総文書数、 $|\{d : d \ni t_i\}|$ は t_i を含む文書の総数を表す。

3.3 Distributed Memory

Distribution memory モデルは、文書のベクトル表現を自由な次元で学習できるモデルであり、Bag-of-Words モデルの語順に依存しないという欠点を克服している。図 1 は、Mikolov ら [6] の単語のベクトル表現を学習するフレームワークで、“the cat sat on” という文において、“on” の単語のベクトル表現を学習している。Distribution memory モデルは、この機構に Paragraph id を導入したものであり、図 2 のようなフレームワークとなる。図 2 での Paragraph id は “the cat sat on” を示す id であり、現在の文脈の損失情報を表すベクトルとなる。

4 提案手法

本研究が提案するシステムの俯瞰を図 3 に示す。このシステムは入力されたテキストのペア T, H からそれらの文書ベクトル表現 $v_i(T), v_i(H)$ を生成する。 $v_i(text)$ はベクトル空間モデル v_i によって生成された

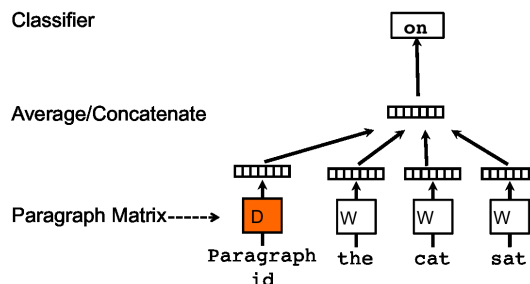


図 2: 文書のベクトル表現の学習 [4]

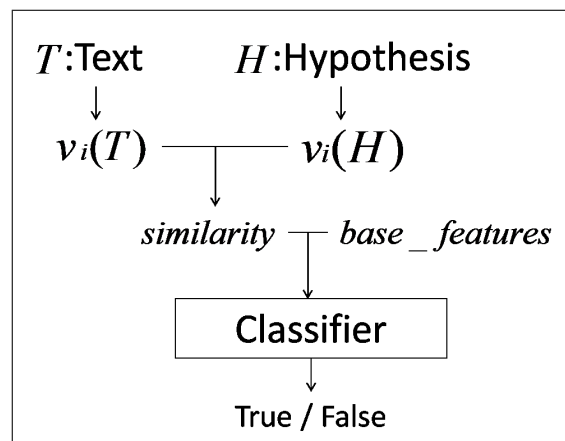


図 3: システムの俯瞰

$text$ のベクトル表現である。文のベクトル表現の詳細に関しては 4.1 節で述べる。次に、2 つのベクトル表現の類似度を計算する。この類似度と base 特徴量を組み合わせたものを分類器によって学習・分類を行う。類似度の計算法は 4.2 節で、base 特徴量に関しては 4.3 節で、分類器に関しては 4.4 節で詳細を述べる。使用したデータセットの詳細は次の 5 章で述べる。

本手法は、外部資源を必要としない。

4.1 文のベクトル表現

文のベクトル表現の生成には、3 章で述べた 3 つのベクトル空間モデルを用いた。Bag-of-Words と TF-IDF ベクトルは、語順の情報を捨てており、日本語における助詞や助動詞などは類似度計算時のノイズとなりうるため、本稿では名詞と動詞以外の単語を無視したベクトル表現を生成した。TF-IDF の計算において、式 (2) の d_j はデータセット中出现する j 番目の文を指し、 $|D|$ はデータセット中の文の総数とする。Distributed memory モデルは、フリーの Python ライブラリである gensim¹ で実装されているものを用い、データセット中の全ての文において学習を行った。パラメータは、alpha(学習率)、size(ベクトルの次元数)、

¹<http://radimrehurek.com/gensim/index.html>

min_count(このパラメータ以下の出現頻度の単語は無視する)を変更した。

ここで、形態素解析には MeCab² を用いた。

4.2 類似度計算手法

ベクトル間の類似度の計算手法はコサイン類似度を用いた。二つのベクトル a_1 と a_2 のコサイン類似度 $\cos(a_1, a_2)$ は以下の式で表される。

$$\cos(a_1, a_2) = \frac{a_1 \cdot a_2}{|a_1||a_2|} \quad (3)$$

4.3 ベース特徴量

ベース特徴量には、2文が表面的にどれだけ類似しているかを示す被覆率を用いる。被覆率は含意関係認識において最もよく用いられる特徴量であり、この特徴量と組み合わせることで認識精度の良くなる特徴量が効果的な特徴量だと判断することができる。本稿では、文字の被覆率を採用し、Wuら[9]を参考に以下の式で表される特徴量 F_{base} を base 特徴量とした。

$$F_{base} = \exp\left(\frac{1}{N} \sum_{n=1}^N \log p_n\right) \quad (4)$$

$$p_n = \frac{\#(ngrams_{T,H})}{\#(ngrams_H)} \quad (5)$$

ここで、 $\#(ngrams_{T,H})$ は文 T と文 H で一致する n -grams の数であり、 $\#(ngrams_H)$ は H の全ての n -grams の数である。 N はパラメータとして用いる。この式は機械翻訳の分野でしばしば用いられる自動評価手法である BLEU 値 [8] を参考にしたものである。

4.4 分類器

含意関係認識は含意がある、なしの2値分類を行うタスクであるため、2値分類に強いとされる Support Vector Machines(SVMs) を分類器として用いる。SVMの実装には Classias[7] を用い、学習アルゴリズムは Truncated Gradient[3] を用いた。カーネル関数は線形カーネルを用いた。

5 実験環境

実験のデータセットには NTCIR-11 RITE-VAL で用いられたデータセットを用いた。このデータセットは訓練データセット 2692 ペア、テストデータセット 1379 ペアの計 4071 ペアの文を含んでいる。従って、文の総数は 8142 文である。

各パラメータは訓練データセットを用いた交差検定の結果を得て、ヒューリスティクスにおいて決定し

た。これにより、distributed memory のパラメータは $\alpha=0.025$, $size=500$, $min_count=2$ とし、base 特徴量では $N=1 \sim 9$ とした。

比較対象として、(1)式で示した $F_{skipgram}$ でも同様の実験を行った。skip-gram モデルの学習は Ishii らと同様に Wikipedia コーパスを用いた。Ishii らは、学習のパラメータに関して明確に述べていなかったため、他のパラメータに倣い交差検定の結果が良くなるように調整した。パラメータは distributed memory と同様のものを調整し、 $\alpha=0.025$, $size=200$, $min_count=5$ とした。

評価尺度は、NTCIR-11 RITE-VAL で採用された Macro-F1 値と正解率を用いた。Macro-F1 値 ($macroF1$) は、クラス c の集合を C , c における精度を $Prec._c$, 再現率を $Rec._c$ とすると、以下の式で表される。

$$macroF1 = \frac{1}{|C|} \sum_c \frac{2 \times Prec._c \times Rec._c}{Prec._c + Rec._c} \quad (6)$$

6 実験結果と考察

各ベクトル表現の類似度とその組み合わせを特徴量として、含意関係認識実験を行った結果を表1に記す。表1では、省力表記としてそれぞれ base:base 特徴量, bow:Bog-of-Words, tfidf:TF-IDF ベクトル, dm:distributed memory, skipgram:skip-gram を表している。

表 1: 各特徴量を用いた認識の結果

特徴量	Macro-F1	正解率
base	63.36	72.88
base+skipgram	62.91	72.88
base+bow	63.54	73.86
base+tfidf	64.48	75.27
base+dm	64.62	74.62
base+bow+tfidf	63.76	74.91
base+bow+dm	64.34	74.11
base+tfidf+dm	63.85	75.42
base+bow+tfidf+dm	64.00	74.62

表1から、skip-gram モデルを用いた手法では base 特徴量のみを用いた手法よりも Macro-F1 値が小さくなったが、文のベクトル表現の類似度を用いた手法では、Macro-F1, 正解率ともに精度が向上した。これより、文のベクトル表現の類似度は単語のベクトル表現の類似度に比べて含意関係認識に有効な特徴量で

²<http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html>

あると言える。また、base+dm の Macro-F1 値が最も高くなり、base+tfidf+dm の正解率が最も高くなった。文のベクトル表現の中でも、distributed memory モデルによるベクトルが最も効果的であると分かる。base+dm の Macro-F1 値は NTCIR RITE-VAL 参加チーム 10 チームの中でも 4 番目に良い結果となった。これより、外部の言語資源を用いない単純な 2 パターンの特徴量のみを用いた手法において、高い精度の認識ができたと言える。外部の言語資源を取り入れた特徴量を加えることで、更なる精度の向上が見込まれる。

T:衆議院で可決し、参議院でこれと異なった議決をした法律案は、衆議院で出席議員の三分の二以上の多数で再び可決したときは、法律となる。
H:衆議院で可決され参議院で否決された法律案は、衆議院で出席議員の三分の二以上の多数で再び可決されたときは、法律となる
baseの判定:N base+dmの判定:Y 正解:Y

図 4: 含意判定の例 1

T:社会主義者鎮圧法は、ビスマルクが社会主義を危惧した結果だった。
H:ビスマルクの失脚後、社会主義者鎮圧法が制定された。
baseの判定:Y base+dmの判定:N 正解:N

図 5: 含意判定の例 2

図 4, 図 5 にシステムの含意判定の例を示す。図 4 で、base のみの特徴量として用いたのは間違った判定をしているが、base+dm を用いた場合は正解の判定をしている。この場合では、T と H の文はほぼ同義であるため文の意味類似度が高く、dm の特徴量が有効に働いていることが分かる。図 5 では、T と H の文は表層的には類似しているが意味は全く異なる。そのため、意味の類似度を考慮に入れた base+dm において、正しい判定が行えた。

T:また、1847年に建国されたリベリアが、西アフリカでもっとも古い独立国となる。
H:リベリアは19世紀に独立し、共和国となった。
base+dmの判定:N 正解:Y

図 6: 含意誤判定の例

図 6 にシステムが誤判定した例を示す。この例では、T から H を推測するのに人間の一般常識が必要となる。こういった含意関係を認識することは現在の技術では困難であり、今後の課題となる。また、文間の意味が類似していてもシステムが誤判定する文のペアも確認される。より多くのデータで distributed memory モデルを学習させることにより、含意関係の認識精度の向上が見込まれる。

7 結論

本稿では、最新のベクトル空間モデルを用いた手法を含意関係認識に適用することで、認識精度が向上することを示した。しかし、ベクトル空間モデルを適用する手法として、今回はコサイン類似度を特徴量とするといった単純な方法を用いた。今後はより効果的な手法で、ベクトル空間モデルを含意関係認識に適用する研究が期待される。また、含意関係認識には 6 節で述べたような課題があるため、これらの解決策を模索していきたい。

参考文献

- [1] Ido Dagan and Oren Glickman. Probabilistic textual entailment: Generic applied modeling of language variability. In *PASCAL Workshop on Text Understanding and Mining*, 2004.
- [2] Ai Ishii, Hiroshi Miyashima, Mio Kobayashi, and Chikara Hoshino. Nul system at ntcir rite-val tasks. In *Proceedings of the 11th NTCIR Conference*, pp. 249–254, 2014.
- [3] John Langford, Lihong Li, and Tong Zhang. Sparse online learning via truncated gradient. *Machine Learning Research*, pp. 777–801, 2009.
- [4] Quoc Le and Tomas Mikolov. Distributed representation of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.
- [5] Suguru Matsuyoshi, Yusuke Miyao, Tomohide Shibata, Chuan-Jie Lin, Cheng-Wei Shih, Yotaro Watanabe, and Teruko Mitamura. Overview of the ntcir-11 recognizing inference in text and validation (rite-val) task. In *Proceedings of the 11th NTCIR Conference*, pp. 223–232, 2014.
- [6] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proceedings of NIPS*, 2013.
- [7] Naoaki Okazaki. Classias: a collection of machine-learning algorithms for classification, 2009.
- [8] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 311–318, 2002.
- [9] Shin-Hung Wu, Li-Jen Hsu, Hua-Wei Lin, Pei-Kai Liao, Liang-Pu Chen, and Tsun Ku. III&CYUT chinese textual entailment recognition system for ntcir-11 rite-val. In *Proceedings of the 11th NTCIR Conference*, pp. 233–239, 2014.
- [10] 横手健一, ボレガラダヌシカ, 石塚満. テキスト含意認識に有効な意味的類似度変換及びその獲得法. 人工知能学会論文誌, pp. 220–229, 2013.
- [11] 金子貴美, 宮尾祐介, 戸次大介. 基本文関係に分解した含意関係認識日本語評価データの構築. 言語処理学会第 19 回全国大会, pp. 928–931, 2013.