6T-6

否定文認識による潜在関係検索の精度向上手法

落合伸彦 † Nguyen Tuan Duc ††
Danushka Bollegala †† 石塚満††

† 東京大学工学部電子情報工学科 †† 東京大学大学院情報理工学系研究科

1. はじめに

潜在関係検索では、入力エンティティペアと類似関係を持つエンティティペアを検索する。例えばクエリ {(Google, Youtube), (Microsoft, ?)} に対して、“Google bought YouTube” や “Microsoft acquired the company Powerset for $100M.” という文を根拠にしてマイクロソフトが買収した企業の一つの “Powerset” を “?” の答えとして出力することが出来る。既存の潜在関係検索エンジン [1, 2] は上記の文から、エンティティ“Microsoft” と “Powerset” との関係を “X acquired * Y” などのような語彙パターンで表現している。しかし、上記の手法では、“Microsoft withdraws the proposal to acquire Yahoo.” という否定的な文に対しても、“Microsoft” と “Yahoo” 間の関係で “X * to acquire Y” という語彙パターンが抽出され、 (Google, YouTube) との関係類似度が高くなってしまい、“Yahoo” を “?” の答えとして出力してしまう。そこで本研究では、Duc らの潜在関係検索エンジン [1, 2] を基にして、機械学習を用いた否定文判断方法、否定文マーカーを用いた正確な検索結果の出力手法を提案する。また、提案手法を用いて、既存の潜在関係検索エンジンの精度を向上できることを確認する。以降、第 2 節では、エンティティ間類似性を利用した潜在関係検索を紹介し、第 3 節で否定文を判断する方法と否定文マーカーを用いた正確な検索結果の出力手法を提案する。第 4 節で評価と実験結果を示す。最後に、第 5 節でまとめと今後の課題について説明する。

2. エンティティペア間の関係類似性を利用した潜在関係検索

2.1 エンティティと関係の Indexing

潜在関係検索を行うために，Web などのテキストコーパスからエンティティを発見し，エンティティ間の関係を抽出する必要がある。そこで，本研究は既存研究 [1] で用いられている，エンティティとエンティティ間の関係を Indexing する手法を利用する。例えば、“It is now official: Microsoft acquires San Fransisco based company Powerset for $ 100M.” という文からは、3 つのエンティティペア (Microsoft, SanFransisco), (Microsoft, Powerset), (Sanfransisco, Powerset) が抽出される。次に，エンティティ間の関係を認識するために，エンティティペアの 2 つのエンティティが出現した位置の文脈から，関係を特徴づける語彙パターンを抽出する。表 1 に語彙パターンの抽出例を示す。

表 1 エンティティペア (Microsoft, Powerset) の語彙パターン抽出例

文 (NE tagged)	It is now official: Microsoft acquires Powerset for $ 100M
変数置き換え	It is now official: X acquires San Francisco based company Y for $ 100M
サブシーケンス	now official: X acquires San Francisco based company Y for $ 100M
Stemming	now offfici: X acquir San Francisco base companiY for $ 100M
関係を特徴づける語彙パターン	X acquir * Y, X * San Francisco * Y, offici: X * Y, X * compani Y for $ 100M, acquir San Francisco base compani Y, ...

3. 否定文認識手法

3.1 否定文の判定

まず、否定的な単語 (e.g., “not”, “never”, “n’t”, “no” の否定を表す単語) を含む文について判断する。上記の形式上のものにさらに、意味上の否定 (e.g., “failed”, “resisted”, . . .) を加える。これはより語彙上と意味上の否定を判定するかとができる。

3.2 機械学習

テキストコーパスから得られた文章に対して、手動で判断した m個の肯定文と n 個の否定文を学習データとして分類器を学習させる。単に “not” があるなどの形式的な否定文だけではなく、“Microsoft withdraws the proposal to acquire Yahoo.”などの意味上の否定を判断するために、上記ので判断したものを学習データとして用いた。

初め、文章を bag-of-words モデルで表現する。次に、先の方法だと機械学習に用いるデータの次元が大きくなってしまい、否定を表す要素の割合が小さくなると考えられるので、文章中で否定文の要素と関係のない動詞、形容詞、副詞以外を除いて学習させる。これにより、関係抽出対象の文を分類器に入れた際に肯定文の場合は “1”、否定文の場合は “-1” と出力する。分類器の中身には SVM(Support Vector Machine) を用いる。これにより任意の文章を分類器に入れた際に肯定、否定を判断する。

3.3 否定文の場合の正確な検索結果の出力方法

本手法では分類器により否定文として判定された任意の文章に対して、n-grams を生成する際に “X * acquired Y neg_sent” のように “neg_sent” という否定文マーカーを付ける。これにより、エンティティ・ペア関係の表現を考える際に共起頻度が変わり正確な結果が出力される。表 2 に否定文の場合の語彙パターンの抽出例を示す。

4. 評価

4.1 データセット

評価データとして、既存研究 [1, 3] に用いられている 12,000 ウェブページのテキストコーパスを使った。また、評価のために 8 個の関係種類のクエリセットを利用する [2]。これらの関係種類

Negative Sentence Detection for Improving Precision of Latent Relational Search

† Nobuhiko Ochiai (Dept. of Information and Communication Engineering, Faculty of Engineering, The Univ. of Tokyo)
†† Nguyen Tuan Duc, Danushka Bollegala, and Mitsuru Ishizuka (Grad. School of Information Science and Technology, The Univ. of Tokyo)

表 2 エンティティペア (Microsoft, Yahoo) の否定の場合の語彙パターン抽出例

文 (NE tagged)	Microsoft withdraws proposal to acquire Yahoo
分類器導入	Microsoft withdraws proposal to acquire Yahoo(否定)
変数置き換え	X withdraws proposal to acquire X (否定)
Stemming	X withdraw propos to acquir Y
関係を特徴づける語彙パターン	X * acquir Y neg_sent, X * to acquir Y neg_sent, ...

表 3 単純な否定文認識の結果

否定文と判定	正しく否定文と検知	誤って検知	検知されない
5.5%(53/956 個)	90.5%(48/53 個)	9.5%(5/53 個)	1.3%(12/956 個)

表 4 単純な否定文判断器による潜在関係検索の性能の向上

実装前	1 番目: { 'melbourne', } (Score = 0.6686) ... 12 番目: { 'canberra', } (Score = 0.2434)
実装後	1 番目: { 'canberra ' } (Score = 0.3852) 2 番目: { 'sydney ' } (Score = 0.0855)

は関係抽出システムの評価 [4]、関係類似度計算アルゴリズムの評価 [5, 6] や既存の潜在関係検索エンジンの評価 [1, 3, 2, 7] によく用いられる関係種類である。形式上の単純な否定文を判定する時には、テキストコーパスの最初の 10 個の文書から得られた 956 個の文章を解析し、精度向上の確認にクエリ {(Japan, Tokyo), (Australia, ?)} を用いた。形式上の単純な否定を表す語に加えて、意味上で否定を表す語も判定する時には、精度向上の確認にクエリ {(Google, YouTube), (Microsoft, ?)} を用いた。この時に形式上の単純な否定文判定で良い結果が得られていたので、精度向上の確認のみを行った。機会学習による否定文判定の時にはテキストコーパスの最初の 58 個の文書から得られた 2707 個の文章を学習させた。様々なパターンの否定文と肯定文を学習させるために形式上の単純な否定文判定の時より多くの文章を解析した。

4.2 否定文判断器の性能

形式上の単純な否定文判定の時は、12,000 ウェブページのテキストコーパスのうち初めの 10 個の文書から得られた 956 個の文章に対して、5.5%(53/956 個) は否定文と判定された。そのうち、90.5%(48/53 個) は正しく検知され、9.5%(5/53 個) は 誤って検知された ("not only…but also", "not so long ago" 等の否定語を含む非否定文)。検知されなかった否定文は 1.3%(12/956 個) あった。表 3 それは、"nothing like…", "nor", "failed" などの否定文である。機械学習で文章をそのまま学習させた分類器は、任意の文章を全て肯定文、つまり "1" と出力した。文章中で否定文の要素と関係ない動詞、形容詞、副詞以外を除いて学習させた分類器も、任意の文章を全て肯定文、つまり、"1" と出力した。

4.3 否定文判断器を利用する潜在関係検索性能の向上

形式上の単純な否定文判定では、表 4 のようにクエリ {(Tokyo, Japan), (?, Australia)} に対して、既存の潜在関係検索エンジンの出力結果は 1 番目が ['melbourne',] (Score = 0.6686) 、12 番目が ['canberra',] (Score = 0.2434) であった。それに対して、実装後の出力結果は 1 番目が ['canberra '] (Score = 0.3852)、2 番目が ['sydney '] (Score = 0.0855) であった。

さらに、意味上の否定を表す語も含めた判定では、クエリ {(Google, Youtube), (Microsoft, ?)} に対して、提案手法では表 5 に示すように "neg_sent" が付加し、 誤りの出力結果である "Yahoo" のスコアが大きく下がっている。

表 5 クエリ {(Google, Youtube), (Microsoft, ?) } の出力結果

既存の出力結果	{('yahoo'), 'X and * Y .','X acquir Y ' ..., 0.1291,)} {('powerset'), 'X * acquir Y', 'X to buy Y' ..., 0.1143,} {('parlano'), 'X * acquir Y' , 'X to buy Y'..., 0.09524,} {('hotmail'), 'X to acquir Y , 'X buy * Y'..., 0.05192,} {('softimage'), 'X to buy Y for,... 0.05011,}
提案手法の出力結果	{('powerset'), 'X to buy Y neg_sent' ..., 0.1239,)}, {('parlano'), 'X nor Y' neg_sent, ',..., 0.09419,}, {('softimage'), 'X 's Y for neg_sent,..., 0.09415,}, {('hotmail'),'X * acquir Y neg_sent' ..., 0.07219, }, {('yahoo'), 'X * to acquir Y neg_sent', ..., 0.06312,}

機械学習による否定文判定では上手く肯定と否定の区別が出来なかったため、精度向上の確認が出来なかった。

5. 終わりに

本稿では、否定文認識の精度向上のために機械学習を用いた否定文か否かの判定方法、否定文マーカーを用いた正確な検索結果の出力方法を説明した。形式上の判定方法を用いると潜在関係検索の精度は向上したが、さらなる精度向上と検知されない文章をなくすために機械学習を用いた分類器による判定を行った。テキストコーパス全体に占める否定文の割合が小さい上に bag-of-words モデルでの表現では、分類器に学習させるデータポイントの次元が大きくなり、否定を特徴づける要素の効果が小さくなると考えた。そのため、文章中で否定の要素と関係のない動詞、形容詞、副詞以外を除いたもので学習させた分類器を用いて行ってみたが、良い結果は得られなかった。これは否定文の素性を捉えられていないために割合の少ない否定文を検知出来なかったと考えられる。今後は、より詳細に否定文を判定するために Stanford parser を用いて構文木を構成し [8]、構文木から得られた素性で学習を行い、否定文認識を行う予定である。

参 考 文 献

1) Duc, N. T., Bollegala, D. and Ishizuka, M.: Using Relational Similarity between Word Pairs for Latent Relational Search on the Web, *Proc. of WI'10*, pp. 196 – 199 (2010).
2) Duc, N. T., Bollegala, D. and Ishizuka, M.: Cross-Language Latent Relational Search: Mapping Knowledge across Languages, *Proc. of AAAI'11*, pp. 1237–1242 (2011).
3) グェン トアンドゥク, ボレガラダヌシカ, 石塚満: エンティティペア間類似性を利用した潜在関係検索, 情報処理学会論文誌, Vol. 52, No. 4, pp. 1790–1802 (2011).
4) Banko, M. and Etzioni, O.: The Tradeoffs Between Open and Traditional Relation Extraction, *Proc. of ACL'08*, pp. 28–36 (2008).
5) Bollegala, D. et al.: Measuring the Similarity between Implicit Semantic Relations from the Web, *Proc. of WWW'09*, ACM, pp. 651–660 (2009).
6) Bollegala, D. et al.: Relational Duality: Unsupervised Extraction of Semantic Relations between Entities on the Web, *Proc. of WWW'10*, ACM, pp. 151–160 (2010).
7) Kato, M. et al.: Query by Analogical Example: Relational Search using Web Search Engine Indices, *Proc. of CIKM'09*, pp. 27–36 (2009).
8) Apostolova, E., Tomuro, N. and Demner-Fushman, D.: Automatic Extraction of Lexico-Syntactic Patterns for Detection of Negation and Speculation Scopes, *Proc. of ACL'11*, pp. 283–287 (2011).