語の意味に基づく文の意味類似性に関する研究

小中, 史人 / KONAKA, Fumihito

(出版者 / Publisher)
法政大学大学院理工学・工学研究科
(雑誌名 / Journal or Publication Title)
法政大学大学院紀要. 理工学・工学研究科編 / 法政大学大学院紀要. 理工学・工学研究科編
(巻 / Volume)
58
(開始ページ / Start Page)
1
(終了ページ / End Page)
2
(発行年 / Year)
2017-03-31
(URL)
https://doi.org/10.15002/00014342

語の意味に基づく文の意味類似性に関する研究

STUDIES ON SEMANTIC TEXTUAL SIMILARITY WITH WORD SENSES

小中 史人
Fumito KONAKA

指導教員 三浦孝夫

法政大学大学院理工学研究科システム理工学専攻創生科学系修士課程

Semantic Similarity is one of the most important issues in Natural Language Processing. Recently we can deliver easily thanks to the growth of Social Network Services (SNS). However Vector Space Model, the dominant method in document retrieval, cannot deal with similarities between short documents produced on SNS. In this work, we propose new methods for sentence similarity. They are especially useful for Semantic Textual Similarity so that we could retrieve SNS documents and short sentences. In experiments, we show the effectivities of our approaches by some experimental results.

Key Words *: Natural Language Processing, Semantic Textual Similarity, Document Retrieval*

1. 問題の背景

インターネット上には膨大な数の文書が存在している．近年，Web ページや Google Books 等様々な形式のデータが爆発的に増加しており，中には構造的な情報を持ったデータも存在する．また，Twitter や Facebook，Instagram 等のソーシャルネットワークサービス（Social Network Service, SNS）の流行により，SNS 上で生成される文書データも増加している．SNS 上の文書データには，短い文書が多い，低頻度語を多く含む，口語，擬音語を含む，特徴的な文法，省略や強調等のノイズを含む表記（例: never/nevr, baby/babyyyyy）等の特徴を持つ．人間が膨大なデータの全てを把握することはできないため，計算機による支援が求められる．

計算機支援の 1 つとして，情報検索が挙げられる．情報検索とは，大量のデータセットからクエリに類似するデータを抽出するタスクである．文書検索での情報表現は主にベクトル空間モデル（Vector Space Model, VSM）が用いられる．このモデルは「似た意味を持つ単語は似た文脈に出現する」という仮説である Distributional Semantic Model (DSM) に基づいて単語の意味情報をベクトルで表現する．ある単語の重要度は，ある文書内におけるその単語の出現頻度（Term Frequency, TF）と，出現している文書数の逆数（Inverted Document Frequency, IDF）の積である TF-IDF 法を用いる．これにより，VSM は文書集合中の各文書とクエリ間の類似度の算出と，それによるランキングが可能という利点がある．しかし VSM は，(1) 短い文書に有効でない，(2) 語順情報の損失，(3) 同義語や類義語といった単語同士の意味関係を反映できない，という欠点を持つ．このため，VSM は SNS 上の文書に対する文書検索に適切ではない．

ユーザは SNS 上で製品やサービスに関する情報を発信しており，これらの情報を効率的に利用することで，ユーザのニーズに応えることが可能になる．しかし，SNS 上の文書に対する検索を実現するためには，VSM の欠点である語順や語義の解釈に加え，ノイズを考慮する必要がある．また，短い文書は文が少ないため，文の意味的な類似性を適切に与える必要もある．VSM の欠点を補完し，SNS 上の文書の特徴を考慮することで，SNS 上の文書も考慮した文書検索の実現を期待できる．

本研究では，計算機上で文の意味的類似性を適切に扱う上で有用な特徴量を提案する．これにより，VSM では扱うことができない短い文書やノイズを含む表現を有効に扱えることを示す．

2. 扱う問題

(1) 語の並びを考慮した文の意味的類似度手法

文の意味を考慮する上で語の並びは重要である．例えば，”A dog bites John.” と”John bites a dog.”では動作の主体と対象が異なるため，異なる意味を表現している．しかし，文書検索において主流の VSM はこれらの文を同一の文と見なす．

また，SNS 上の文書には口語表現が少なくない．口語では難解な語の使用よりも，平易な語を用いた熟語表現が好まれるため，一般に困難とされる熟語表現も考慮する必要がある．

本研究では，レーベンシュタイン距離とユークリッド距離に，辞書を用いた語の意味的類似性の概念を導入することで，単語の多義性に加え熟語表現を考慮する．これにより，文の意味的類似性を適切に捉えられることを示す [1].

(2) グラフ構造を用いた語の意味情報

辞書を用いた語同士の意味的な類似度指標は，辞書構造内部の最短経路に基づくものや，それに加えてコーパスの統計情報を考慮したものである．しかし，辞書内の階層構造には循環や輪の存在が明らかになっている．従って，既存の類似度指標は辞書情報を適切に利用できていない．

本研究では，辞書内の階層構造をグラフ構造として考え，循環や輪を考慮可能な手法を提案する．また，グラフは巨大

なため，情報を保存したまま低次元に圧縮する手法も提案する．これにより，(1) 語同士の意味類似性のより適切な解釈，(2) インデックス保存による高速な類似度計算，(3) STS における精度と速度の改善，を実現できることを示す [2].

(3) 意味の埋め込み表現を用いた文脈の再学習

計算機上で自然言語を扱う上で，語の多義性を考慮する必要がある．語ではなく語義をベクトルで表現する手法もいくつか提案されているが，これらの手法は語義ベクトルの獲得のために語ベクトルによる文脈表現を用いている．

そこで，本研究では語義ベクトルを用いた文脈表現と語義予測，並びに文脈の再学習手法を提案する．これにより，(1) 多義性を考慮した文脈表現，(2) 適切な語義の予測，(3) 適切な文の表現，が実現できることを示す [3].

3. 結果

まず，語順を考慮した文の意味的類似度手法では，システムに入力された文ペアの類似度を算出し，それに従ってランキングを行った．その結果，平均適合率は，提案手法は 0.8，比較手法は 0.81 を示し，ほぼ同程度の精度を達成している．比較手法では類似度の低いペアが上位にランクインしているが，提案手法ではそのようなペアを上位に含んでいない．

次に，グラフ構造を用いた語の意味情報では，単語のペアに対して類似度を算出し，ランキングを行った．その結果，人手スコアによるランキングとのスピアマン順位相関係数は，提案手法では最大 0.798，比較手法で最大 0.784 を示し，僅かではあるが，より適切に語の類似性を解釈できていると言える．一方，実行時間については，インデックスを用いる提案手法は最短 0.127 秒，比較手法は最短 2.186 秒となっており，実行時間を約 94%短縮している．また，STS では，システムに入力された文ペアの類似度を算出した結果，人手スコアによるランキングとのピアソン相関係数は，提案手法では 0.561，比較手法では最大 0.511 を示しており，比較手法の約 1.1 倍の相関を達成している．STS における実行時間は，提案手法で 2.158 秒，比較手法では 386.923 秒となっており，実行時間を約 95%短縮している．

最後に，意味の埋め込み表現を用いた文脈表現と語義の予測では，語義曖昧性解消された 2 つの単語の類似度を算出し，ランキングを行った．その結果，人手スコアによるランキングとのスピアマン順位相関係数は，提案手法が最大で 0.222，比較手法が 0.166 を示し，比較手法と比べ相関が 0.056 向上している．また，文ペアに対しても類似度算出とランキングを行ったところ，人手スコアによるランキングとのピアソン相関係数，スピアマン順位相関係数の両観点で，提案手法は比較手法よりも強い相関を達成した．図 1 に実験結果を示す．具体的にはピアソン相関係数の観点では提案手法は最大 0.455 の相関を達成し，比較手法と比べ相関が最大で 0.136 向上している．スピアマン順位相関係数の観点では提案手法は最大で 0.456 の相関を達成し，比較手法と比べ相関が最大で 0.144 向上している．

4. 結論

本研究では，SNS 上で生成される短い文書を計算機上で効率的に処理するために，文を表現する上で有効な特徴量を

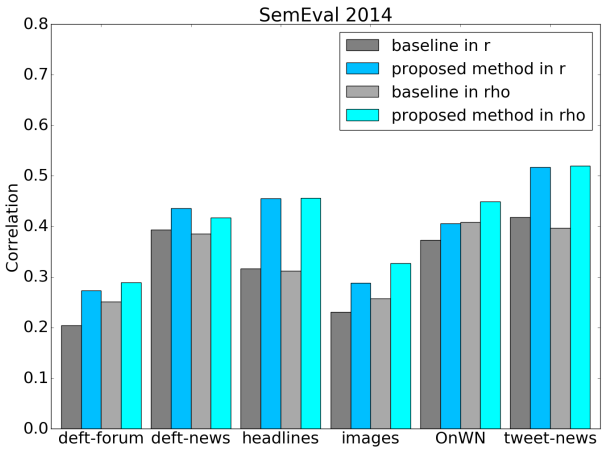

図 1: 文ペアに対する類似度付与タスクの相関係数における比較．提案手法はベースラインよりも強い相関を示している．

提案した．SNS 上で生成される文書は短いものやノイズを含むものが多く，従来の方法では対応できない．このため，文書ではなく，文を表現することで SNS 上の文書を適切に扱える．

語順を考慮した文の意味的類似度手法では，語の並びを考慮することでより適切に文の意味的類似性を捉えられることを示した．グラフ構造による語の意味情報では，グラフ情報を低次元に縮小することで，比較手法と同程度の性能を維持したまま，比較手法の約 6%の計算時間で単語間の意味的類似度を算出できることを示した．また，文の意味的類似度タスクにおいても比較手法と比べ約 1.1 倍の相関を達成した．加えて，比較手法の約 5%の時間で文の意味的類似度を算出できることを示した．意味の埋め込み表現を用いた文脈表現と語義の予測では，比較手法の約 1.3 倍の相関を達成し，多義性を考慮した文脈表現により，適切に語義を予測できることを確認した．加えて，文の意味的類似度タスクにおいても比較手法と比べ最大で約 1.5 倍の相関を達成し，適切な語義の予測により，より良い文の表現を獲得できることを示した．

残る問題点として，言語により文の構造が大きく変わることに対する手法の一般化を達成できなかったことが挙げられる．本研究では単語間の区切りが明確な英語を扱っているが，日本語のように区切りが曖昧な言語においても，適切にトークン化することで応用の可能性はあると考えられる．

参考文献

1) Fumito Konaka and Takao Miura (2015): ”Textual Similarity for Word Sequences”, in Proceedings of the 8th International Conference on Similarity Search and Applications (SISAP), pp. 244-249.
2) Fumito Konaka and Takao Miura (2016): ”Domain Graph for Sentence Similarity”, in Proceedings of the 9th International Conference on Similarity Search and Applications (SISAP), pp. 151-163.
3) 小中 史人 and 三浦 孝夫 (2017) : “意味の埋め込み表現を用いた文脈の再学習”，第 9 回データ工学と情報マネジメントに関するフォーラム (DEIM).