2C1-NFC-2-5

Wikipediaの概念ベクトルを用いた単語間関連度の推定

Estimating the Relationship between Words using Notion Vectors of Wikipedia

パトリック　チャン　　土方　嘉徳[*1]　　西田　正吾[*1]
Patrick Chan[*1]　　Yoshinori HIJIKATA　　Shogo NISHIDA

[*1]大阪大学大学院 基礎工学研究科
Graduate School of Engineering Science, Osaka University

単語間の意味的関係性（semantic relatedness）を計算することは，情報検索や自然言語処理の分野では重要な研究課題である．我々は，Wikipedia 文書のレイアウト情報と単語頻度を用いて，単語間の関連度を推定する手法を提案する．

1. はじめに

単語間の意味的関係性（semantic relatedness）は，情報検索，文書要約，単語の曖昧性解消などの分野で広く用いられている．意味的関係性は，単語間（時にフレーズ間）の論理的関係や因果的関係などを表す．本稿では，単語間の意味的関係性を単純に単語間関連度と呼ぶ．単語間関連度を計算する手法として，Wikipedia 文書を用いたものが注目されている．

中でも，Explicit Semantic Analysis (ESA) [1] は代表的な手法である．ESA は，単語を概念を要素とするベクトルで表現する．ここで概念は Wikipedia 記事が表す概念とする．それぞれのベクトル要素は，対象単語と対象概念との関連度（実際には，正規化された TFIDF）である．最後に，単語間関連度は，それぞれのベクトルのコサイン距離で計算される．

ESA は単語頻度のみを用いていることになるが，我々は単語とそれが出現する概念（Wikipedia 記事）との関連度はレイアウト情報も関連していると考える．例えば，Wikipedia 記事の一番上の章（通常，要約とみなされる）では，概念の説明に用いられる単語は注意深く選択されている可能性が高い．また，太字（通常，強調の意味で持ちいられる）の単語は他の単語よりも，対象概念との関連が強いかもしれない．

ESA は単語頻度のみに依存しているため，単語の出現頻度が低い単語間では，その関連度の計算精度が低くなると思われる．我々の方法はレイアウト情報が保管して予測するため，このような場合でも高い精度で推定できると思われる．

2. 提案手法

我々の手法は，ESA の手法において，単語の概念ベクトルの要素の値を算出するのに，TFIDF だけでなくレイアウト情報も含める．具体的には以下のレイアウト情報を用いる．

BOLD: 対象単語が太字かどうか.
ITALIC: 対象単語が斜体かどうか.
ANCHOR: 対象単語がアンカーテキストかどうか.
CAPTION: 対象単語がファイルキャプション（図など）かどうか.
LIST: 対象単語が箇条書きの中にあるかどうか.
DEPTH: 対象単語の出現する章の深さ.
HEIGHT: 対象単語の出現する章の番号.

また，レイアウト情報に加えて，ESA でも用いられていた TDIDF も用いる．

連絡先: Patrick Chan, 大阪大学大学院基礎工学研究科, 大阪府豊中市待兼山町 1-3

レイアウト情報を用いて単語間関連度を計算するには，そのレイアウトの種類が対象の Wikipedia 記事との概念との関係性を獲得する必要がある．すなわちあるレイアウトの種類（例えば BOLD）が付けられた単語が，どれほど対象の概念と関連があるかを，あらかじめ一般化しておく必要がある．

我々は，レイアウトの種類ごとにその中の単語と，対象の Wikipedia 記事の概念との関連度の正解データを獲得することにした．具体的には，3 名の評価者に 60 個の Wikipedia 記事中の 30 個の単語を提示し，その単語と対象の Wikipedia 記事の概念との関連度を 7 段階で与えてもらった．3 名の与えた関連度の平均を正解データと見なして，以下の回帰式により回帰分析（標準的な重回帰分析）を行うことにした．

$$Relevance = \beta_0 + \beta_1 * BOLD + \beta_2 * ITALIC + \beta_3 * ANCHOR + \beta_4 * CAPTION + \beta_5 * LIST + \beta_6 * HEIGHT + \beta_7 * DEPTH + \beta_8 * TFIDF$$

ある単語のインスタンスが，1 つの Wikipedia 記事中で複数出現した時に，β_n にどういう値を取るかは確定的ではない．そこで，我々は全ての単語インスタンスを対象に，いずれか一つでも対象のレイアウト種類が付与されていたら真 (1) とみなす場合（PMA）を考えた．また，Wikipedia 記事中の最も先頭に出現した単語インスタンスのみ対象にした場合（PMT）を考えた．前者はさらに，DEPTH に対して，最も浅い位置で出現したものの深さを与える場合（PMAS）と最も深い位置に出現したものの深さを考える場合（PMAD）に分けた．

3. 評価実験

提案手法の有効性を調査した．特に，ESA は単語頻度が低い場合に弱いと思われたため，単語ペアの出現頻度ごと（両方の単語が 25%タイル未満，両方の単語が 50%タイル未満，両方の単語が 75%タイル未満，全ての単語使用）に分けて，ESA と提案手法を比較した．評価には，2010 年 10 月 11 日時点での Wikipedia のダンプデータ 27GB 分を取得した．単語間関連度の正解データが必要であるが，これには WordSimilarity-353 [2] を用いた．評価指標は，Spearman の順位相関係数である．

実験結果は，図 1 のようになった．25%タイルにおいては，PMAS と PMAD が ESA を上回っていることが分かる．やはり，単語頻度が低い場合には，ESA の予測精度は低いことが分かる．PMA と PMT を比べると，PMT の精度が低い．これは，1 つの単語のレイアウト情報を用いるよりも，複数の単語のレイアウト情報を用いる方が，精度が高いことを意味する．50%タイル以上になると，提案手法よりも ESA の方が精度が高くなる．出現頻度がある程度高くなってくると，TFIDF の

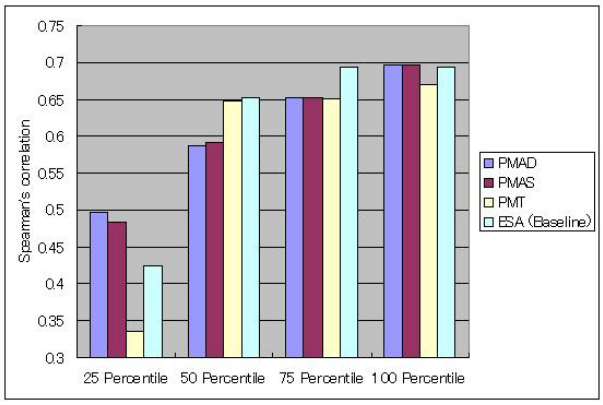

図 1: 単語間関連度に関する実験結果

表 1: 単語頻度ごとに適用手法を切り替えた場合の ESA との比較

Method	Spearman's correlation
ESA	0.696
PMAD / ESA 切替	0.708

情報だけで十分に予測できることが分かる．

出現頻度が 25%タイル未満では提案手法が良く，25%タイル以上では ESA の方が良かったので，25%タイル未満の単語ペアについては提案手法（PMAD）を適用し，25%タイル以上の単語ペアについては ESA を適用することにした．これで求めた結果は表**??**のようになる．この結果から，提案手法と ESA を切り替えて使えば，全体では少し ESA を上回ることができることが分かる．

4. レイアウト情報の調査

我々は，レイアウト情報の種類と，その中の単語と Wikipedia 記事の概念との関連度に関する関係について，詳しく調査した．表 2 に PMAD の場合の結果を載せる．関係性の統計的有意差は t 検定で行った．正規化 TFIDF が最も高い重みをもつことが分かる（統計的有意差もあり）．BOLD と ITALIC も，統計的に有意な属性である．ITALIC は BOLD よりも相関が小さくなっている．BOLD はほとんどの執筆者がその単語を強調するために使うのに対して，ITALIC は，名前に付けたり参照文献があることを表すのにつけたりと，多様な使われ方をしている．そのため ITALIC は相関が少し低くなっていると思われる．CAPTION と LIST も統計的に優位な属性である．いずれも，BOLD や ITALIC よりも相関が低くなっている．これは，BOLD や ITALIC よりも文が長くなり，ノイズとなる単語が入っているものと思われる．最後に，DEPTH と HEIGHT は統計的有意差はなかった．これは，CAPTION や LIST よりもさらに文が長くなり，ノイズとなる単語が入ったからと思われる．ただし，HEIGHT については弱い相関が認められる．

表 2: レイアウト情報の種類と概念との関連度

Attribute	Coefficient	Significance
BOLD	0.372	ρ <0.001
ITALIC	0.151	ρ <0.001
ANCHOR	0.094	ρ <0.001
CAPTION	0.048	ρ =0.003
LIST	0.063	ρ <0.001
DEPTH	0.001	ρ =0.916
HEIGHT	-0.003	ρ =0.174
TFIDF	1.60	ρ <0.001

5. まとめ

本研究では，単語間関連度を Wikipedia を用いて推定する手法を提案した．提案手法は，単語頻度を用いている ESA を改良し，Wikipedia 記事のレイアウト情報も用いた．Wikipedia のダンプデータと標準的な単語間関連度の正解データでる WordSimilarity-353 を用いて評価したところ，単語ペアの両方の単語頻度が低い場合には，提案手法は ESA を上回ることが分かった．今後の課題としては，どのレイアウト情報が単語と Wikipedia 記事の概念と関連性が高いかを分析し，単語頻度に応じて利用する手法を切り替える方法を実装する．

参考文献

[1] E. Gabrilovich and S. Markovitch, “Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis”, Proc. of IJCAI 2007.

[2] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin, “Placing search in context: The concept revisited”, ACM TOIS, Vol.20, No.1, pp.116-131, 2002.