4D-3

文章間の類似度計算システム

生沼 貴大† 天沼 博† 松澤 和光†
†神奈川大学 大学院工学研究科 電気電子情報工学科

1 はじめに

コンピュータが文章や会話の文脈や内容を認識するには、含まれる単語の一つ一つの意味を把握し、さらにそれらの単語から常識的・日常的に連想される様々な事項を推定する必要がある。このような仕組みを機械的に実現する際には、単語間の関係としてある単語と単語がどの程度の類似性を持つか判定する機能が有用である。こうした単語間の類似性判定については既に種々の研究[1][2]があり、その一部は意味的類似度計算ツール[3]として公開されている。

しかし、実際の応用では単語間の類似性だけではなく、多くの単語が含まれる文章や記事などの間の類似性を判定したい場合も多い。例えば、あるニュースに意味がよく似た別の記事を探したり、ニュースから連想される諺を選定する[4]ような場合である。そうした単語セット同士で類似度を判定する場合には、類似度の計算が単語ペアの総当たりとなるため、処理が煩雑となってしまう問題があった。

そこで、一般的な文章および会話に用いられる基本的な単語セットに対して、提供されている類似度計算ツールを用いてあらかじめ総当たりで類似度を計算し、その計算結果を言葉の意味のデータベース（概念ベース）として用意しておく。そして、対象とする文章中にある単語セットについて対応する概念データをベクトル的に加算し、この加算したベクトル同士の余弦計算をすることによって、単語セット同士の類似度を容易に計算できるシステムを提案する。

2 提案手法

本研究の類似度計算システムは、①基本的な単語セットとあらかじめ計算した類似度を用いて概念ベクトルを生成する部分、②生成した概念ベクトル間の余弦計算により類似度を計算する部分、の２つの小システムにより構成される。また、前準備としてgainen.ID、gainen.MXの２つのデータを用意する。

2.1 gainen.ID

あらかじめ類似度を計算しておく一般的な単語セットで重複のないよう通し番号を割り振ったもの。単語セットの選定には、単語がどの程度馴染みがあるかを 1.0～7.0 の数値で表した親密度[5]を参考とし、これが 5.5 以上の名詞および 6.0 以上の動詞を中心として選出した。また、概念間の意味的類似度計算ツールや形態素解析ツールに非対応な単語は除外した。この結果、8552 単語を収録している。これは新聞記事１年分 8597 記事を対象とした場合、平均登場単語数 186 単語中 100 単語、約 54%をカバーしている。

2.2 gainen.MX

gainen.ID から総当たりの単語ペアを生成し意味的類似度計算ツール[3]を用いて類似度を計算し、結果を行列化してまとめたもの。このgainen.MX は各単語ペアの類似度を重みとみなした概念ベースであり、本システムに対応している単語を属性ごとに収録している。また、単語間の類似度の値の大部分はごく小さな値となっているが、これらは文章間の類似度計算に与える影響が極めて小さいため除外している。

2.3 概念ベクトルの生成

類似度計算を行いたい文章について次の手順により概念ベクトルの生成を行う。

(1) 対象の文章に含まれる単語とその出現頻度を記したワードリストを作成する
(2) 作成したワードリストから gainen.ID を参照して未収録単語を除外する
(3) ワードリストの単語について gainen.MX を参照して各属性との類似度と単語の出現頻度を乗算する
(4) (3)の結果を全ての属性について加算したものをその単語の重みとする
(5) 全ての単語について(3)(4)の作業を行いその結果を概念ベクトルとして生成する。

The similarity calculation system between sentences.
†Takahiro Oinuma, †Hiroshi Amanuma
†Kazumitsu Matsuzawa
†Graduate School of Engineering, Kanagawa University

2.4 概念ベクトル間の余弦計算

2.3 で生成した２つの概念ベクトルから余弦計算を行い、この計算結果が２つの概念ベクトル間の類似度となり 0～1.0 の数値で表される。

３ 評価実験

文章間の類似度を算出する例としてインターネット上の記事に対して提案手法を適用した。比較対象として、単語ペアの総当たりで余弦計算により算出した類似度を表１に示す。使用した記事は 1,2：金環日食、3：天気予報、4：コーラに発癌性物質、5：生活保護不正受給にそれぞれ関する記事である。

表１．総当たり計算との比較

手法	組み合わせ	類似度
総当たり計算	記事１－記事２	0.8039
	記事１－記事３	0.6076
	記事２－記事３	0.4629
	記事４－記事５	0.3617
	記事１－記事４	0.3477
	記事１－記事５	0.3053
提案手法	記事１－記事２	0.7203
	記事１－記事３	0.3843
	記事２－記事３	0.3191
	記事１－記事４	0.2853
	記事４－記事５	0.1809
	記事１－記事５	0.1121

記事 1 から記事 3 を含んだ組み合わせはどちらの手法で計算した場合でも高い類似度となっている。これは、記事 1 から記事 3 はどれも天文に関する事柄が記述されているため関連の無い記事 4、記事 5 を含んだ組み合わせよりも類似度が高くなり妥当といえる。

また、類似度順に並べた際の順位についてはほぼ同じ結果が得られたが、一部順位が入れ替わっている部分がある。これは、総当たり計算の場合対象となる文章に含まれる全ての単語ペアに対して余弦計算をしているため、文章の内容とは直接関係せず、どのような文章にも含まれうる単語（例えば、「する」や「ある」など）によって類似度が全体的に高くなり順位に影響を与えたと思われる。提案手法を用いた場合にはそういった問題が起こりうる単語は収録単語から除外しているため文章の内容により重点をおいた結果となっている。

次に計算量について、総当たり計算の場合は全ての単語ペアについて類似度を算出してから余弦計算するのに対して、提案手法ではあらかじめ用意したデータを参照するため余弦計算のみで類似度の算出が可能となっている。

４ おわりに

先行研究[1][2]で構築されている類似判定システムは本研究と同様の概念ベースを内蔵しているが、これらの概念データは公開されてはいない。また本研究で利用した意味的類似度計算ツール[3]は、類似計算の機能は公開されていても内部のデータは非公開である。一般にこうした概念データは辞書やコーパス等を基に構成されるため、それらの著作表現上の権利が複雑に絡んで公開は難しいかと予想される。しかし、類似度計算の機能そのものは内部データの「表現上の権利」からは独立しているため、[3]のように一般に公開が可能である。そこで本研究のように、こうした公開された機能を用いて構成した概念ベースであれば、第３者の自由な利用が可能になると考えられる。

さらに、類似性判定は概念ベース内蔵の方式に限らず、例えばシソーラス上の距離を用いるなど様々な方式が考えられる。そうした類似性判定結果を本研究により用いて概念ベース化すれば、より多様な状況に応じた利用し易い知識データとして活用が期待できるだろう。

なお、本研究では対象とする単語セットの規模を１万語弱に限ったが、複数の概念ベクトルを足し合わせて別の単語を表現することができるので、概念ベクトルの長さを変えずに対象単語の規模を拡張できる発展性がある。

参考文献

[1] 笠原,松澤,石川：国語辞書を利用した日常語の類似性判別, 情処論,Vol.38,No.7, pp.1272－1283.1997
[2] 奥村,土屋,渡部,河岡：概念間の関連度計算のための大規模概念ベースの構築, 自然言語処理,Vol.14,No.5,pp.41-64.2007
[3] 野口,清水,杉本,石川：あらゆる概念表記への対応・精度向上を目指した意味的類似度計算ツール, 情処全大第 69 回,2006
[4] 海老澤,天沼,松澤：諺を用いたニュース見出し生成法, 人工知能学会ことば工学研究会,SIG-LSE-B002,2010
[5] 天野,近藤：NTT データベースシリーズ日本語の語彙特性(第１期 CD-ROM 版),三省堂,2003