

動的なドメイン知識獲得とテキストの格解析を行なう

5 Y-1

類似文書検索手法の評価

永松 健司, 田中 英彦

{naga,tanaka}@mtl.t.u-tokyo.ac.jp

東京大学大学院 工学系研究科*

1 はじめに

自然言語処理では、二つの表現間の類似性を判定する処理が様々な場面での基礎的な指標として利用される。特に情報検索では、そのような類似性判定が中心的な役割を果たすが、昨今の爆発的に増加している電子化テキストを検索する際の不満は、検索エンジンがテキスト内の表現間に適切な類似性を判定できないためである。

[1]では、簡単な係り受け構造を持つ表現に対する類似度判定手法を提案した。この手法では、

1. 従来なら予め構成しておく必要のある知識ベースに相当する情報を、検索対象となるテキストを含むコーパスから統計的に求めることで知識ベースの構成や保守に要する手間を無くすと共に、
2. テキストの格情報を利用して意味的な処理を行なうことで、単なる統計情報による情報検索よりも的確な処理が行なえると期待できる。

本稿では、この手法の部分要素である、簡略化された格解析処理と類似文書へのマッピング処理の評価について述べる。まず、第2節で本手法の概要を説明した後、第3節で評価結果を提示し、その考察を行なう。

2 係り受け表現に対する類似度判定手法

2.1 簡略化された格解析処理

本来、格解析は大きなコストを要する処理であるため、大規模なコーパスデータに対して処理を行なうには、ある程度の簡素化を行なわねばならない。そこで、本手法で扱う係り受け表現を「係り受けの深さが一段の句構造」に限定し、かつ、名詞（形容動詞）・動詞・形容詞に限定した格フレームへとその語句表現を格解析する。

本手法での格解析は以下の手順で行なう（図1）。

1. 統計データによる形態素分割・品詞属性付与 [2]
2. 表層的情報（読点の有無、表層格情報が重複しない等）を用いて、上述の品詞を持つ形態素列を格フレームの候補となり得る複数の形態素列へ分割
3. EDR 共起辞書の動詞共起パターン情報を利用して、最も適合する組み合わせを格フレームとして出力

*"Evaluation of a Document Search Method employing Dynamic Information Acquisition and Case Analysis"

Kenji Nagamatsu, Hidehiko Tanaka

University of Tokyo, Graduate School of Engineering,

7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

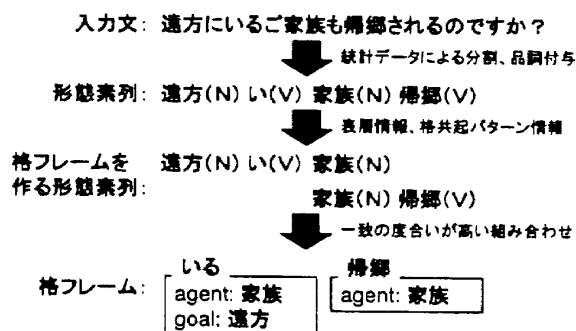


図1: 簡略化された格解析の手順

2.2 入力表現のコーパス内へのマッピングと近傍格情報を用いた類似度計算

本手法では、入力表現対それぞれの「類似事例」がコーパス内で出現している、その前後の表現の格情報の分布の相関を調べることで入力表現間の類似度を定義できるとし、以下の二段階の処理による類似度判定を行なう。

1. 入力表現をコーパス内の類似事例へとマッピング

情報・文書検索の対象となるテキストコーパスでは通常、単語の逆インデックスが作成されているため、

- (a) 格解析の結果の単語を基に、逆インデックスを利用して高速に各単語の出現位置を求める
- (b) シソーラスの利用による同義語展開はこの時点で行なう（図2の「遠い」と「遠方」等）
- (c) その上で、多くの単語の出現位置が近くに固まっている箇所を探索し、それらに得点を与えて順序付けする（図2）。

2. 類似事例の近傍の格情報間で相関値を計算

- (a) こうして順序付けられたコーパス内の類似事例の上位のものから、その出現位置の前後のある近傍内の文を格解析

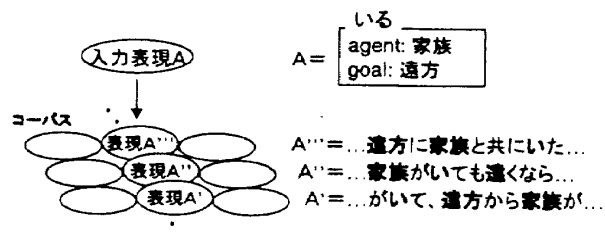


図2: コーパス内の類似事例へのマッピング

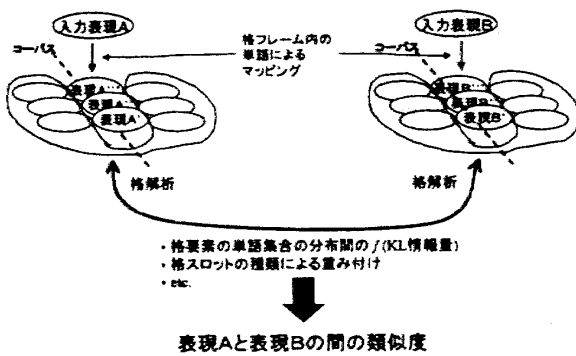


図3: 類似事例の近傍の格情報間で相関値を計算

- (b) 入力表現対それぞれに対する類似事例の得点が高い方からある一定数を選び、それらの間で格要素の単語集合の相関を示す値を求め、それを入力表現対の類似度と定義（図3）

この最後のステップ(b)における相関値には、KL情報量を基にしたものや、格スロットの種類によって重みを付けた格要素単語の一致度など、いくつかの方法を併せて評価する。

3 本手法の構成要素の評価・考察

ここでは、上で説明した手法の構成要素である、簡略化された格解析処理とコーパス内類似事例へのマッピング処理について評価を行なった。

3.1 簡略化された格解析処理の評価

評価の手順は、EDR コーパスから取り出した1000文に対して、本手法による格解析を行なった結果の出力データと実際の正解データとを直接比較し、形態素の文字列と品詞属性、および格情報が一致しているかどうかを調べるといったものである。

第1ステップの形態素分割と品詞属性付与については、以下の評価結果が得られている[2]。

既知データ	全形態素	98.2%
	限定した品詞のみ	99.8%
未知データ	全形態素	93.2%
	限定した品詞のみ	97.9%

つまり、形態素分割・品詞属性付与に関しては、ほとんど問題のないことが示されている。これ以降のデータは、形態素分割・品詞属性付与を正しく行なえる文の中から、評価用の文を取り出している。

次に格情報の一致を調べた結果を示す。

格フレームが完全に一致したもの	76.5%
格要素が過剰に付加されたもの	8.3%
格要素が不足したもの	7.8%
格情報が間違っていたもの	18.6%

ここでの誤検出の原因としては、連体修飾語句の誤り、助詞「の」の多義性に由来するものなどが多かった。

これらは日本語テキストに多く現れる表現であるため、表層的表現からの判断だけでなく、係り側と受け側の語句間の共起情報などのある程度、意味を反映する情報を利用する必要がある。

3.2 類似事例へのマッピング処理の評価

次に、類似事例へのマッピング処理（2.2節のステップ1）の評価を行なった。ここでは、EDR コーパスから、格フレーム一つからなる構造を持つ200文を取り出し、本手法でコーパス内類似事例に対応させた結果が正解（その入力文自体）である割合を評価した。

この実験で用いたスコアリングは、格フレーム内の形態素文字列、格情報の一致個数により順序付けを行なっている。また、EDR コーパスから統計的に抽出してある共起語辞書により展開される同義語の個数は平均2.1個であった。

ランクトップの解が正解である割合	91.5%
上位5位までに正解が含まれる割合	99.5%

最上位の解が正解でない場合の誤検出の原因は、同義語展開や格解析処理の誤りに由来するものが多かった。また、その場合でも、大半の出力結果が入力文とはほぼ同内容であった。これはEDR コーパスが新聞記事を多く含むものであり、同様の内容を持つ文章がいくつか存在しているためである。

今回の実験は入力文自体が単純であり、なおかつ探索する正解も自分自身という容易なタスクだったために、比較的高い精度になったようである。

4 おわりに

本稿では、情報検索において必要とされる係り受け構造を持った表現間で類似度を計算する手法を提案し、その構成要素について評価を行なった。本手法では、大規模なテキストコーパスを利用し、格解析処理により抽出された格情報の分布の一致度を用いて類似度を定義する。

本稿では、部分要素の評価のみ行ない、最終的な文書間の類似度の妥当性に対してはまだ評価していない。これは、今後、あるコーパス中の表現を検索語句とし、別のコーパス中の表現を検索した結果が、どの程度、許容されるかを人間の判断により評価する予定である。

参考文献

- [1] 永松, 田中. 係り受け構造を持つ表現に対する類似度判定手法の提案. 情報処理学会第55回全国大会, 第2巻, pp. 87-98, Sept. 1997. 6J-2.
- [2] 永松, 田中. 文字 n -gram データからの k -nn 法に基づく統計的形態素推定. 情報処理学会第55回全国大会, 第2巻, pp. 344-345, Sept. 1997. 4AE-1.