6J-02 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 *

永松健司　田中英彦†
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．特に情報検索においては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の検索に対する不満は，検索エンジンがドキュメント内の表現間に適切な類似性を判定できないことに起因する．

この点に対しては最近，XML の枠組と併せてオントロジを利用する検索手法が実用化に近付いており，検索文をオントロジのスキーマに合わせた形に解析することで，一種，関係データベース的な検索が可能となる．しかし，これも基底の単語自体に対する類似度は定義されないため，柔軟性という点では十分とは言えない．

本稿では，情報検索の場面における検索文として最も望まれる，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する．この手法は，従来なら予め構成しておかねばならない知識ベースに対応する情報を，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，対象テキストを格解析した情報を利用することで，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．

2 係り受け構造を持つ表現に対する類似度判定手法

単語間に類似度を定義する場合以上に，係り受け構造を持つ表現間の類似度を定義する場合には，知識情報の役割が重要となる．これは，単語間類似度では考慮する必要がなかった係り受け構造自体が持つ情報（意味）を適切に扱う必要が出て来ることに依る．

そこで本研究では，

- 大規模コーパスから抽出した統計情報を用いて知識ベースの代替を図る．
- その統計情報を抽出する際や，入力表現の類似事例をコーパス中で検索（2.2節で後述）する際には，格解析で得られた情報を基に処理する．

という方針を採っている．

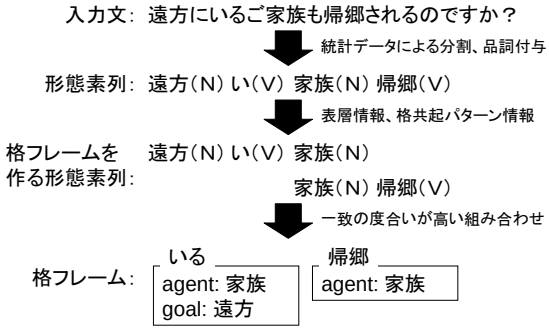

図 1: 簡略化された格解析の手順

2.1 コーパスデータを扱うための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．従って，大規模なコーパスデータに対して処理を行なうためには，ある程度の簡素化を行なわねばならない．

ここでは，情報検索の一要素という場面を考えて，本手法で扱う係り受け構造を持った語句を「係り受けの深さが一段の句構造」に限定することにする．その上で，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っている（図 1）．

1. 入力文に対して，統計データを用いた形態素分割および品詞属性の付与．
2. 上述の品詞を持つ形態素列を表層的情報（読点の有無，表層格情報が重複しない等）を用いて，格フレームの候補となり得る複数の形態素列へと分割．
3. EDR 共起辞書の動詞共起パターン情報を利用して，その一つの形態素列中で述語となり得る単語，その格要素となり得る単語を調べ，最も適合する組み合わせを格フレームとして出力．

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

単語間の類似性に関する評価結果によると，入力単語対それぞれがコーパス内で出現した箇所の前後の単語分布が類似しているということは，入力単語間の意味の類

---

*On a Similarity Measures between Language Expressions with Dependency Structures

†Kenji Nagamatsu　Hidehiko Tanaka
{naga, tanaka}@mtl.t.u-tokyo.ac.jp
Faculty of Engineering, University of Tokyo, 7-3-1 Hongo, Bunkyo-Ku, Tokyo, 113, Japan

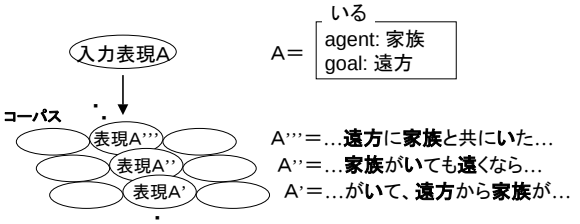

図 2: コーパス内の類似事例へのマッピング

似性と良い相関関係にある [1]．このことから，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対の類似度が定義できることが予想される．

ここで問題となるのは，入力表現からコーパス内の “類似事例” にマッピングする処理であり，当然ながら最終的に求める類似度とは別の尺度を用いて，その近さを判定する必要がある．

そこで本稿では，以下のような二段階の処理による類似度判定手法を提案する．

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) 入力表現を格解析した結果の単語を基に，その逆インデックスを利用することで高速に各単語の出現位置を求めることができる．

   (b) 更に，シソーラスの利用による同義語の展開もこの時点で行なう（図 2中，A” の「遠い」と「遠方」等）ことで，より広範な類似事例が検索できる．

   (c) その上で，多くの単語の出現位置が近く（一つの句の範囲内）に固まっている箇所を探し出した後，それらに得点を付与する（図 2）．

   以上の処理は既存の検索エンジンで行なわれている処理と同等のものであり，すなわち，本手法が既存の検索手法の延長として実現できることを示している．

2. **類似事例の近傍の格情報間で相関値を計算**

   (a) 上述のマッピングにより得点付けられたコーパス内の類似事例それぞれについて，得点の高い方から，その出現位置の前後のある近傍内の文に対して格解析を行なう．

   (b) 入力表現の格情報との比較を行ない，各類似事例の得点を調整する．

   (c) 入力表現対それぞれに対する類似事例の得点が上位の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計算し，それを入力表現対の類似度と定義する．

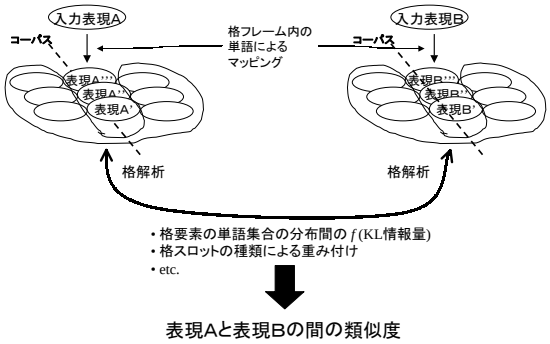

図 3: 類似事例の近傍の格情報間で相関値を計算

このステップ (c) における相関を示す値には，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ことを考えている．

2.3 本手法の利点

このような方法を採ることによって，以下のような利点が期待できる．

- 実際の日本語の使用状況に即した類似度が期待される．連続して現われる表現はお互いに関連した意味を表すことが多いという事実からも，近傍格情報の相関を用いることで，本来は知識ベースを必要とする意味の関連性が扱える．
- それと関連して，純粋なテキストコーパスから統計情報を抽出することが可能なため，オントロジ等の知識ベースを予め構築しておく必要がない．
- 格フレーム内のスロットに重みを付け，スロット内単語間の類似度と重みから格フレーム同志の類似度を定義するという手法も，コーパス内事例へのマッピングという処理により，部分的ではあるが取り込むことができる．
- 既存の検索エンジンの機能の延長として実現することができる．

3 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，二つの表現間に類似度を定義する．

本手法によって求めた最終的な類似度の妥当性に対しては，今後，文書検索処理として実装した後，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．単語対に対する類似性規準の心理実験による評価． 情報処理学会第 54 回全国大会，第 2 巻, pp. 89–90, Mar. 1997. 3C-7.