自然言語処理 93－1
（1993. 1. 13）

EDR電子化辞書を用いた単語類似度計算法

崔 進　小松 英二　安原 宏
（株）日本電子化辞書研究所

用例ベース機械翻訳など、事例に基づく推論方式の自然言語処理システムにおいては、類似した文例を選出するため、単語間の類似度を計算することが必要になる。従来の単語間類似度計算方式は、分類語彙表、類語辞典、シソーラス等を用いて、単語間の距離を測定するものであった。それらはいずれも単語体系を利用した類似度計算であり、単語の持つ概念および概念体系上の上位下位関係を利用した類似度計算ではないため、多義性を持つ単語に関する処理方法の問題や、異なる言語の単語間の類似度を計算できないという問題があった。本稿では、単語の持つ概念をもとに単語間の類似度を計算する方法について述べ、単語の文法及び意味情報を記述した単語辞書、概念の上位下位関係を記述した言語に依存しない概念体系を用いた単語間の類似度計算法を提案する。単語辞書及び概念体系はEDR電子化辞書を利用した。

A CALCULATION OF SIMILARITY BETWEEN WORDS USING EDR ELECTRONIC DICTIONARY

Jin CUI,　Eiji KOMATSU,　Hiroshi YASUHARA
Japan Electionic Dictionary Research Institute, Ltd. (EDR)
Mita-Kokusai-Bldg. 4-28, Mita 1-Chome, Minato-ku, Tokyo, 108, Japan

In order to find similar sentences in natural language processing systems, (such as memory-based translation), which uses the case-based reasoning method, it is necessary to calculate the similarities between words that appear in each sentence. Research has been done in this area before, (such as calculating the distance between two words using a word thesaurus), but because the word distance has not been calculated using word sense in previous research, there are some problems that have not been solved, such as how to treat the words that have more than one meaning, and how to calculate the similarities between two words that belong to two different languages. In this paper we try to introduce a method that shows how to calculate the similarities between two words based on their word sense. Our research proves that the EDR word dictionary and concept dictionary are effective.

1 はじめに

用例ベース機械翻訳(Memory-based Translation)など、事例に基づく推論(Case-based Reasoning)方式の自然言語処理システムでは、用例文間の類似性をどう決めるかが問題のひとつとなっている[1]。一方、用例文間の類似性を決めるためには、二つの用例文にある単語間の類似度を計算する必要がある。

例えば、用例ベース機械翻訳システムにおいて、用例ベースに2つの用例文があるとする

(1) アメリカに立つ (leave for America)
(2) 証人に立つ (be a witness)

解析対象文「イギリスに立つ」が与えられたとき、用例文 (2) を排除して用例文 (1) を選出するために、単語「イギリス」と単語「アメリカ」、単語「イギリス」と単語「証人」の類似度を計算する必要がある。

これまで、分類語彙表、類語新辞典などの単語体系（単語のシソーラス）を利用した単語間の距離を測定する研究[2],[3]がある。単語体系を利用した単語距離測定法は、単語の持つ概念をもとに類似度を計算する方式ではないため、多義性を持つ単語に関する処理方法の問題や、異なる言語の単語間の類似度を計算できないという問題があった。

本稿では、単語の持つ概念及びそれらの概念の上位概念を用いて単語間の類似度を計算する方式について述べる。本方式では、単語体系を使わずに、単語の文法及び意味情報を記述した単語辞書、概念の上位下位関係を記述した言語に依存しない概念体系を用いて単語間の類似度を計算する。

以下では、2節でEDR電子化辞書の中の本文と関係する部分の紹介、3節で本文における単語類似度についての基本的な考え方、4節で単語間類似度の計算方法及び計算式、5節で本方式による類似度計算の実験について述べる。

2 EDR電子化辞書

EDR電子化辞書は、単語辞書、概念辞書、対訳辞書、共起辞書の4種類の辞書から構成されている。本研究は、その中の単語辞書及び概念辞書を利用して単語間の類似度を計算する。

2.1 単語辞書

EDR単語辞書の基本的な役割は、単語と概念の対応関係を記述し、この対応関係が成り立つときの文法的特性を与えることである。単語辞書は単語項目の集合であり、日本語単語辞書と英語単語辞書からなる。単語項目は、見出し情報、文法情報、意味情報の3つのブロックで構成されている。意味情報は単語の多義性を識別するための概念情報を与えるものであり、その単語の持つ語義を識別するものである。EDR電子化辞書において、単語の持つ概念を次のように定義する。

> 人間はある単語を見たときに、その単語によって想起されるいくつかのイメージを思い浮かべることができる。このようなイメージの中で個別の状況に依存しない高い一般性を持つものを概念と呼ぶ[4]。

2.2 概念辞書

EDR概念辞書は概念項目の集合である。概念項目は、単語辞書に記述されている概念の間に成り立つ関係を記述するものである。概念間の関係の種類によって、概念辞書は概念記述と概念体系に分けられる[5]。

概念記述は、上位下位関係以外の、ある概念と他の概念との間にどのような関係が成り立つかを記述するものである。概念記述の概念項目例を次に示す。

(C#食べる) -agent-> (C#哺乳類)

概念体系は、概念間の上位下位関係を記述するものである。概念体系の概念項目例を次に示す。

(C#犬) -kind_of-> (C#哺乳類)
(C#猫) -kind_of-> (C#哺乳類)

本研究においては、単語の振る舞い、使い方の観点から単語間の類似度を計算するため、概念記

述を使わず、概念体系を用いて単語の持つ概念間の類似度を計算する。

2.3 単語辞書と概念体系辞書の関係

単語辞書と概念体系の関係は図2.1 で示す。

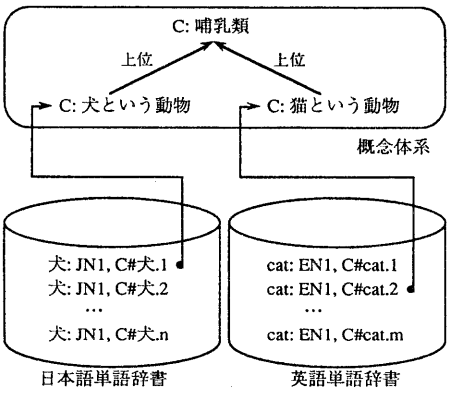

図2.1　EDR単語辞書と概念体系の関係

日本語単語辞書も英語単語辞書も共通な概念を持ち、概念体系とリンクする。概念体系は言語に依存しないので、概念レベルで単語の類似度を計算する場合、日本語単語同士、英語単語同士の類似度を計算するだけでなく、日本語単語と英語単語の類似度も計算することが可能である。

3 単語間の類似度

本研究における類似度の特徴について述べる。

3.1 単語間の類似度

本研究における単語間の類似度は、単語間の距離と違い、単語の振る舞い、使い方の観点から見た類似度である。この類似度を計算するとき、2節で述べた単語辞書及び概念体系を用い、単語の持つ概念から概念体系の上位概念または上位概念の上位概念を検索し、単語の概念及び上位概念の持ち具合により単語間の類似度を計算する。ただし、ここで注意することは、2節で述べた「犬が食べる」のような概念記述関係は本研究の類似度に取り込んでいないことである。すなわち、本研究においては、「犬」と「猫」の類似度は高いが、「食べる」と「哺乳類」の類似度は低い。

単語の振る舞い、使い方の観点から単語間の類似度を計算する場合、2つの単語の類似度が大きいほど用例文における役割が類似する。これは構文構造が決まった文に対して用いると特に有効である。

3.2 類似度の分類

本研究では、両単語の持つ概念の関係から、単語間の関係を同一、同義、類似の3つに分類し、3種類の関係に基づき単語間の類似度を計算する。

単語間の同一、同義、類似関係を説明する前に、これから使ういくつかの表記法を定義しておく。

$$Wi\ \{ i = 1, 2 \}$$

Wi は、日本語または英語の単語であり、システムの入力とする。

$$Ci = \{ c^i(1), c^i(2), \ldots c^i(n_i) \} \quad \{ i = 1, 2 \}$$

Ci は単語 Wi が持つ概念の集合であり、$c^i(j)$ $\{j=1,2 \ldots n_i\}$ は Wi の持つ概念である。

$$S^i_k = \{ s^i(k,1), s^i(k,2), \ldots s^i(k,N_{ki}) \} \quad \{ i = 1, 2 \}$$

S^i_k は Wi の k 段目の上位概念からなる集合である。$s^i(k, t)$ $\{ t = 1, 2, \ldots N_{ki} \}$ は、Wi の k 段目の上位概念であり、集合 S^i_{k-1} 内のある概念の上位概念である。

以下単語間の同一、同義、類似関係を定義する。

同一関係

両単語の持つ概念が全く同じ、すなわち、C1 と C2 が全く同じとなる関係を同一関係(identical)と呼ぶ（図3.1）。Wi と Wi は同一関係にある。同一関係にある W1 と W2 に対しては、同じ単語として扱い、類似度を計算する必要がない。

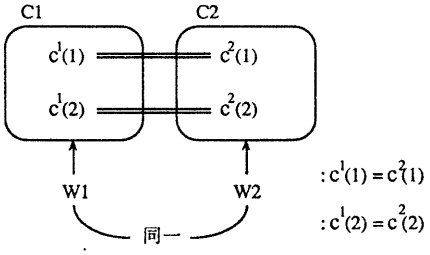

図3.1　同一関係

EDR 単語辞書を調べた結果、同一関係にある単語対には、「英国／イギリス」、「辞書／字典」、「外国／海外」などがある。

同義関係

C1 と C2 は、同一関係ではないが、共通な概念が一つ以上存在する（図3.2）関係を同義関係(synonym)と呼ぶ。同義関係においては、共通な概念数が多いほど W1 と W2 の類似度は大きい。

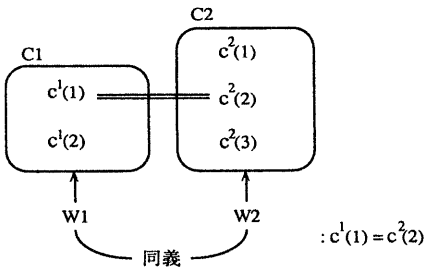

図3.2　同義関係

EDR 単語辞書を調べた結果、同義関係にあるの単語対には、「男／男子」、「犬／スパイ」、「国語／日本語」などがある。

類似関係

C1 と C2 の間に共通な概念が存在しないが、ある上位概念の階層 k まで S^1_k と S^2_k の間に共通な概念が存在する（図3.3）。この関係を類似関係(similar)と呼ぶ。類似関係においては、共通な上位概念数が多いほど W1 と W2 の類似度は大きい。

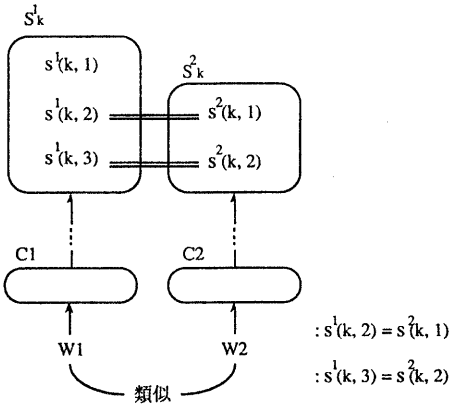

図3.3　類似関係

類似関係にある２つの単語の類似度は、共通な上位概念数のほか、両単語の上位概念の数にも関係する。すなわち、５つの上位概念の中で１つが他と同じである場合より、２つの上位概念の中で１つが他と同じである場合の方が類似度は大きい。

EDR 単語辞書、概念体系を調べた結果、類似関係にある単語対には、「部長／社長」、「部長／私」、「アメリカ／イギリス」などがある。

4 単語間類似度の計算法

本研究では、任意の２つの単語が入力されると、以下の手順に基づき、単語間の類似度を計算する。

1) 単語辞書を検索し、両単語の語義からなる概念集合に基づき、同義関係での類似度（類似度 α）を計算する
2) 概念体系を検索し、両単語の概念集合の上位概念集合に基づき、類似関係での類似度（類似度 β）を計算する。
3) 類似度 α と類似度 β をもとに、単語間の類似度（類似度 δ）を計算する。

ただし、単語とその単語自身あるいは同一関係にある単語対に対しては、類似度を計算する必要がないので、以上の 1) 2) 3) を適用しない。

類似度 α と類似度 β をもとに単語間の類似度 δ を決めるので、必要に応じて同義関係（類似度 α）

の重みまたは類似関係（類似度β）の重みを大きくしたり小さくしたりして、自由に単語間類似度の特性を変化させることが可能である。

図4.1は、システムの構成を示すブロック図である。

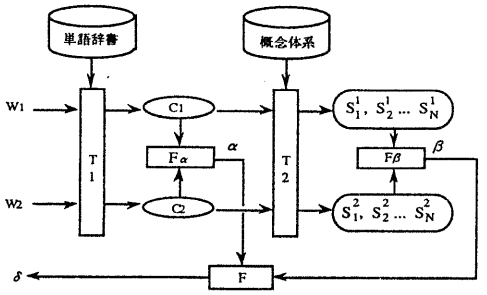

図4.1　システム構成図

W1とW2が入力されると、まず、単語辞書検索部T1を用いて単語辞書を検索することによりC1及びC2を生成し、概念集合類似度計算部F_αを用いて類似度αの値を算出する。次に、概念体系検索部T2を用いて概念体系を検索することにより各階層での上位概念集合S^1_kと及びS^2_kを生成し、上位概念集合類似度計算部F_βを用いて類似度βの値を算出する。最後に、単語類似度計算部Fを用いてαとβをもとにW1とW2の類似度δの値を算出して出力する。

以下類似度α、類似度β及び類似度δの値を算出する方法について説明していく。

類似度α

類似度αは、同義関係においての類似度であり、C1とC2の間に存在する共通な概念の数により類似度αの値を算出する。

集合Xに存在する要素の数を｜X｜で表わすと

$$\alpha = |C1 \cap C2|$$

類似度β

類似度βは、類似関係における類似度である。

概念集合C1、C2を与えると、まず、N段目までの概念体系を検索し、各段階での上位概念集合S^1_k及びS^2_k｛k＝1, 2 ... N｝を生成する。Nは概念体系の階層数以下とする任意の整数を指定できる。次にそれぞれのS^1_kとS^2_kに基づき、階層kでの類似度β_kを計算する。最後に、β_k｛k＝1, 2 ... N｝から類似度βの値を算出する。

以下の式を用いて階層kでの類似度β_kの値を算出する。N_{ki}｛i＝1, 2｝はWiのk段目の上位概念集合内の異なる概念数であり、CS_kはS^1_kとS^2_kの間に共通な上位概念の数である。

$$\beta k = (1 + K_{\beta 1} * CS_k)(1 + K_{\beta 2} * (\frac{CSk}{Nk1} + \frac{CSk}{Nk2})) - 1$$

CS_kの値が大きいほど類似度β_kの値は大きい、$CS_k = 0$のとき、β_kは0になる。CS_k/N_{ki}は、共通な上位概念数対上位概念集合内の概念数の比率である。N_{k1}またはN_{k2}の値が大きいほどβ_kは小さくなる。上記の式で示したように、5つの上位概念の中で1つが他と同じ（$CS_k/N_{ki} = 1/5$）である場合より、2つの上位概念の中で1つが他と同じ（$CS_k/N_{ki} = 1/2$）である場合の方が類似度β_kは大きくなる。

$K_{\beta 1}$を用いてCS_kの重みを調整し、$K_{\beta 2}$を用いて$CS_k/N_{k1} + CS_k/N_{k2}$の重みを調整する、$K_{\beta 2} = 0$のとき、$N_{k1}$と$N_{k2}$は$\beta_k$に作用しなくなる。

類似度βを計算する式を以下に示す。

$$\beta = K_1 * \beta_1 + K_2 * \beta_2 + \ldots + K_N * \beta_N$$

K_k｛k＝1, 2 ... N｝を用いて各階層での類似度値β_kの重みを調整する。

類似度δ

類似度δは、単語間の類似度である。以下の式を用いて類似度δの値を算出する。

$$\delta = 1 - e^{-(K_\alpha * \alpha + K_\beta * \beta)}$$

δは、0から1の範囲内を変動し、1に近いほど両単語は類似する。

K_αを用いてαの重みを調整し、K_αの値を大

きくするほど類似度αの値は大きくなる。K_βを用いてβの重みを調整し、K_βの値が大きくするほど類似度βの値は大きくなる。K_αとK_βの値を調整することによりδの変動範囲を調整することもできる。

δの値はシステムの出力になる。

5 実験・評価

本方式の有効性を検証するために、EDR単語辞書及び概念体系を利用し、単語間の類似度を計算する実験を行なった。

重み値$K_{\beta 1}$、$K_{\beta 2}$、K_β、K_α、K_1、K_2 ... K_N及び上位概念をたどる回数Nは、本類似度計算法だけではなくて、単語辞書及び概念体系の構造にも関係する。今回は実験しながら決めていく方式をとった。サンプルとして、N=2のときの、重み値の1つの組を以下に示す。

表5.1　重み値のサンプル

K_α	K_β	$K_{\beta 1}$	$K_{\beta 2}$	K_1	K_2
0.45	0.028	2.75	8.25	1	0.05

以上の重み値を用い、単語間の類似度を計算した結果を表5.2に示す。

表5.2　単語間の類似度を計算する実験の結果

W1	W2	δ
辞書	辞書	1
外国	海外	1
男	男子	0.9909
イギリス	アメリカ	0.8978
イギリス	証人	0
犬	猫	0.8380
犬	哺乳類	0.4604
食べる	哺乳類	0
犬	cat	0.7423
計算機	computer	0.9648
apple	orange	0.7446
sun	moon	0.8193
red	blue	0.5058

6 おわりに

単語辞書から得られる単語の概念集合の間の類似度を計算し、さらに概念体系を用い、当該概念集合の上位概念集合の間の類似度を計算し、それを重み付け、補正する手段を用いて単語間の類似度を計算する方式について述べた。本方式を用いると、任意の言語の任意の2つの単語に対して、類似度を計算することができる。

参考文献

[1] 松原仁：推論技術の観点からみた事例に基づく推論、人工知能学会誌、Val.7、No.4、pp.567-575 (1992)

[2] 黒橋禎夫、長尾真：格フレームにおける意味マーカと例文の有効性について、情報処理学会自然言語処理研究会報告 91-11 (1992)

[3] Sumita, E. and Iida, H.：Experiments and Prospects of Example-Based Machine Translation; Proc. of the 29th Annual Meeting of the Association for Computational Linguistics (1991)

[4] 日本電子化辞書研究所：TR-19　日本語単語辞書（第2版再改訂）

[5] 日本電子化辞書研究所：TR-20　概念辞書（第3版）