2003－NL－156　(6)
2003／7／25

単語類似度の尺度比較支援ツールの作成

河部恒 †‡　柏岡秀紀 †‡　田中英輝 ‡　松本裕治 †

† 奈良先端科学技術大学院大学 情報科学研究科
‡ ATR 音声コミュニケーション研究所
Email: {kou-k,matsu}@is.aist-nara.ac.jp, {hideki.kashioka,hideki.tanaka}@atr.co.jp

単語は、自然言語における処理の単位としてもっとも基本的なものである。近年の形態素解析や構文・係り受け解析の精度の向上により、単語を与えられたテキストから自動的に同定して切り出すことが可能になっている。一方、単語の表す内容については、人間用の辞書を機械可読な形に直したり、シソーラスやオントロジーと呼ばれる知識表現を人手であるいはコーパスから半自動的に構築するなどして利用する研究がすすんでいる。以上の様な状況のもと、単語を比較する処理が様々なアプリケーションにおいて利用されている。単語の表す内容を比較する上で、各単語がどのくらい似ているのかを表す尺度として、適当な距離空間を導入して単語間の類似度が定義される。導入される類似度は単語について分かっている様々な形の情報から計算される。現在、単語類似度を測る数多くの尺度が提案されており、利用できる情報、適用するタスクにより、その最適な選択は異なる。様々に提案されている類似度の尺度から最適な尺度を把握するために、アプリケーション毎にシステムをインプリメントするのは現実的でない。我々は今回、それらを簡便に比較するためのツールを製作した。

キーワード: 単語類似度、共起、クラスタリング、シソーラス

A Tool for Comparing Measures of Word Similarity

Kou Kawabe†‡　Hideki Kashioka†‡　Hideki Tanaka‡　Yuji Matsumoto†

† Graduate School of Information Science, Nara Institute of Science and Technology
‡ ATR Spoken Language Translation Research Labs.

Words are one of the most principal unit for Natural Language Processing. For these days, improving the accuracy of morphological analysis, dependency analysis and parsing techniques enable us to identify the unit automatically from given text. The meanings of the words, on the other hand, is extracted from knowledge expression such as dictionaries for humans and thesaurus. To compare the meanings of words, a measure, word similarity, is introduced to denote the distance between these words and is calculated from several infomation about these words. A dozen of these measures are recently proposed and selecting them properly for our target task is a crucial problem. We made a tool to compare these measures for the convenience of the choice.

Keywords: Word Similarity, co-occurrence, Clustering, Thesaurus

1　はじめに

単語は、自然言語における処理の単位としてもっとも基本的なものである。人はその歴史を通じて言語の知識を辞書のように、単語を単位とした形で記述・蓄積してきたが、自然言語処理の分野でも近年の形態素解析や構文・係り受け解析の精度の向上により、与えられたテキストから単語を自動的に同定して切り出すことが高い精度で可能になっている。

一方、単語の表す内容については、人間用の辞書を機械可読な形に直したり、シソーラスやオントロジーと呼ばれる知識表現を人手であるいはコーパスから半自動的に構築するなどして利用しようと研究がすすんでいる。

1.1　単語同士の比較

このように単語を単位とする言語知識の獲得・蓄積は自然の流れであろう。さらに歩みを進めて単語間の関係を扱おうするならば、単語同士を比較するような何らかの処理が必要となってくる。

例えば以下のような単語の抽象化やグループ化と

いった処理が、様々なアプリケーションにおいて要求されている。

1. 単語の抽象化、汎化
   - 例 1: QA システムにおけるクエリ拡張
     “過去 10 年でもっとも被害のあった 地震 は？”
     “過去 10 年でもっとも被害のあった 災害 は？”
     「地震」→「災害」と置き換えることで検索の再現率を向上させる。
   - 例 2: 用例ベース機械翻訳におけるテンプレートマッチ
     入力文: けいはんな にはどういけばいいですか？
     用例文: 東京駅 にはどういけばいいですか？
     「けいはんな」を名詞 X として抽象化して考えることで用例文とマッチさせることができる。
   - 例 3: 言い換えにおける単語の選択
2. 単語のグループ化
   - 例 4: 単語クラスタリングやシソーラスの自動構築
     k-nearest neighbors classification[4]
     distributional clustering[9]
     デンドログラム等、
     一般的なクラスタリングアルゴリズムでは、もっとも類似性の高い単語の組からはじめてボトムアップに単語の集合をつくっていく。
3. 単語間の関係の類推
   - 例 5: 動詞の選択制限、多義性の解消

現在これら単語間の関係を比較する上で、各単語がどのくらい似ているのか・異なるかを表す尺度を客観的な基準として導入するのが一般的である。これを単語類似度 (Word Similarity) と呼ぶ。その多くは適当な距離空間[1]を導入して距離という形で表現されており、その決定に用いられる情報も様々である。

また、単語同士の比較が可能になればそこからさらに歩みを進めて、単語の集合同士の比較、すなわち文 $\Leftrightarrow$ 文の関係や文の集まりである文章間の関係なども比較することができる。現在、単語 $\Leftrightarrow$ 文章間の関係を示すには $tf \cdot idf$ や特異値分解による LSI などの手法が使われている。

しかしここでは基本となる単語 $\Leftrightarrow$ 単語の関係のみに注目し、次項でその定義を見ていくことにする。

[1] $\mathbf{R^n}$ 上のハウスドルフ空間

1.2 どうやって類似性を定義づけるか

では単語が類似しているとはどういうことであろうか。当然アプリケーション毎にどういう “類似度” が欲しいのかは異なると考えられる。

Miller ら [8] はある単語の出現を別の単語に入れ替えたとして意味が通るかという置き換え可能性 (contextual interchangeability) で説明しようとしている。また Lin[6] は汎用性と理論的な正当性を念頭にした妥当な仮定[2]をおいて類似度を定義することを試みている。また類似性をより狭義にとらえ、“同一” (概念/属性が全く同じ)、“同義” (共通な概念/属性が一つ以上存在する)、“類似” (共通な概念/属性は存在しないが共通の親ノードを持つ) のように 3 つに分類する方法もある。[17]

ここでは、類似しているという観点に以下のようなものを考えた:

1. 統語的に同じ振る舞いをする
   例: “eat dinner” と “eat apple” のように同じ動詞の目的語になる。
2. 同じ意味のカテゴリに属する
   例: “apple” と “orange” $\in$ “fruit”
3. 同じトピックでいっしょに出てくる
   例: “doctor” と “hospital”
4. 文字面が似ている
   例: “color” と “colour”
   “machine” と “machinery”

ここで重要なことは 1 から 4 のどの観点を類似度として採用するかはアプリケーション毎に考えなければならないということである。

つまりシステムの一部として単語類似度を考える場合、(i) どのような観点で、(ii) どのような距離空間を導入するかという二つの問題を考慮しなければならないが、これらの組合わせの数は多く、最適なものを見つけるためにアプリケーション毎にすべての類似度尺度をインプリメントして比較するのは現実的ではない。

しかし単語 $\Leftrightarrow$ 単語類似度の計算はその性質上アプリケーションからの独立性が高く、もしアプリ

[2] x と y の類似度は、共通性の持つ情報量に比例し、記述が異なるほど小さくなり、等価の場合は 1 である、等の 6 つの仮定をおいて類似度を導いている。

ケーション制作者がいくらかの主体的判断のみで最適な類似度尺度採用の決定ができればこれは大きな利得を生じる。

我々は今回これを実現するためのツールを製作した。以下、section 2 では本ツールで検討する類似度の尺度の定義を概観し、section 3 ではツールの構成について述べる。最後に section 4 で今後の課題を述べる。

2 類似度の尺度の種類

現在提案されている様々な類似度の尺度は大きく三つに分類することができる。すなわちコーパスから自動的に計算するもの、人手でつくられた知識から計算するもの、その他の方法によるものである。この節ではまずこれらを順に見ていき、最後に各尺度の評価について述べる。

2.1 コーパスから自動的に獲得するもの

大規模なテキストデータであるコーパスから獲得される類似度の尺度は共起関係に基づいて計算される。ここで共起関係とは、単語 x と 単語 y が関係 f においてコーパス中に出現することを言う。

共起関係 f の種類を以下に挙げる。

- 文法的な関係
  主語-動詞, 動詞-目的語などの格関係
  係り関係
  同一複合語内
  同一文内
  同一段落内

- 非文法的な関係
  n-gram
  size n の window 内

2.1.1 観点

共起関係に基づいた類似度で計れるのは section 1.2 の分類のうち、 1. 統語的に同じ振る舞いをする、および 3. 同じトピックでいっしょに出てくるのみである。

2.1.2 距離の定義

コーパスから自動的に獲得出来る類似度の定義を以下に挙げる。[7][1][3]

ここではベクトルモデルと確率モデルの二つに大別する。[3]

- ベクトルモデルベース[4]

$$Dice = \frac{2\,|x \cap y|}{|x| + |y|}$$

$$Jaccard = \frac{2\,|x \cap y|}{|x \cup y|}$$

$$Overlap = \frac{2\,|x \cap y|}{\min(|x|, |y|)}$$

$$Cosine = \frac{< x, y >}{|x|\,|y|}$$

- 確率モデルベース

$$point\text{-}wise\ MI(x, y) = \log \frac{p(x, y)}{p(x)p(y)}$$

$$KL\,(p, q) = D\,(p \parallel q) = \sum_i p_i \log \frac{p_i}{q_i}$$

$$IR\,(p, q) = \frac{1}{2}\left[D\,(p \parallel \frac{p+q}{2}) + D\,(q \parallel \frac{p+q}{2})\right]$$

$$\alpha\text{-}skew\,(p, q) = D\,(q \parallel \alpha \cdot p + (1-\alpha) \cdot q)$$

$$L_1(p, q) = \sum_i |\,p_i - q_i\,|$$

2.2 人手の知識から獲得するもの

人手による単語間の類似度を示す知識は、辞書的な情報として蓄えられることが多い。例えばシソーラスやオントロジーなどである。概念やカテゴリの上位下位関係等を用いて類似度が定義される。

2.2.1 観点

人手でつくられた知識をベースに計算される類似度で計れるのは、section 1.2 の分類のうち 2. 同じ意味のカテゴリに属するものである。

[3] 共起に関しての直接共起、間接共起という分類も考えられる。すなわち x と y が直接共起している状況を測るものと、x と y に対する様々な共起相手を元にベクトル表現し、その近さを測る、というものである

[4] この他に χ^2 統計量, Log-likelihood, 補完類似度など。

2.2.2 距離の定義

ここではシソーラスベースとネットワークベースの二つに大別する。

- シソーラスベース
  シソーラスとは単語を木構造に分類したもので、大きく二つの種類がある。一つはリーフにのみ単語が配置されているもの、もう一つはリーフとノードの両方に単語が配置されているものであるが、どちらの形式でも二つの単語が与えられたときに共通の親ノードを同定することができる。今これを $common(x, y)$ とすると、シソーラスベースの類似度のもっとも基本的な定義は以下で与えられる。
  - リーフのみに単語があるシソーラス (分類シソーラス):[10]

$$SIM(x, y) = \frac{2 * height(common(x, y))}{height(\mathbf{T})}$$

  - ノードとリーフに単語があるシソーラス (上位下位シソーラス): [21]

$$SIM(x, y) = \frac{2 * depth(common(x, y))}{depth(x) + depth(y)}$$

  ただし $height()$ はリーフからの木 $\mathbf{T}$ の高さ、$depth()$ はルートからの深さを表す関数[5]である。
  図 1 は上位下位シソーラスの depth を表している。

- オントロジー、ネットワークベース
  このタイプの知識表現とシソーラスとの大きな違いは、木構造ではなく閉路を持つ一般のネットワーク構造をしていたり、アークに付与される関係が複数あったり、ノードに重みがついている、等である。一般的には単語ノードが関係を表すアークで結ばれており、各アークには距離が定義されている。
  具体的には Semantic network や EDR 概念辞書 [22]、辞書定義文のネットワーク [11] などが存在する。例えば EDR 概念辞書を使った類似度の計算方法には 崔らの方法 [17] がある。

$$SIM(x, y) = 1 - e^{-(w_\alpha * \alpha + w_\beta * \beta)}$$

  ここで α は同義関係における類似度、β は類似関係における類似度、w は重みである。

[5]明らかに $depth(x) = height(\mathbf{T}) - height(x)$ が成り立つ

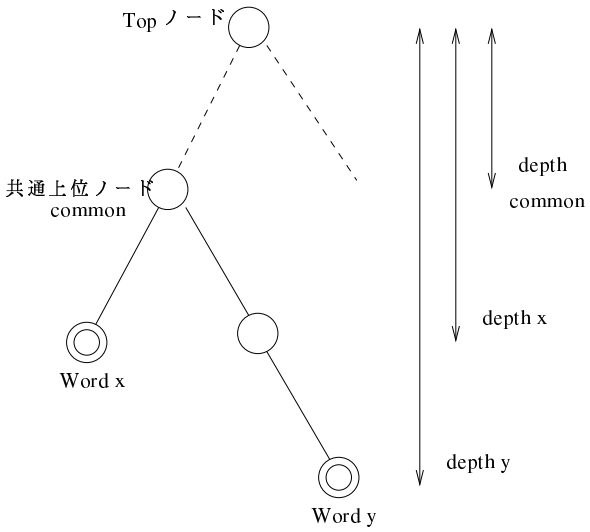

図 1: 上位下位シソーラスの場合

2.2.3 特徴

人手で書かれた知識から獲得する類似度のもっとも大きな問題点は、単語セットが固定であるということである。大規模なものでも 10^4 から 10^5 語程度であり、未知語に関しては類似度を計算することができない。

また、各シソーラスのフォーマットの違いが挙げられる。木構造と言っても分類語彙表 [15] や角川新類語辞典 [18] のようにリーフまでの木の高さが全てそろっているものと、NTT 語彙体系 [20] のようにバラバラなものがあり、単に深さの基準だけでは偏りが生じる可能性がある。

2.3 その他

- 上の方法の混合
  池原らはベクトルの基底に NTT 語彙体系の意味属性をとる方法を提案している [14]。
  笠原らの方法 [23] はシソーラスを利用して次元を圧縮した後特異値分解する。
  また、稲子ら [13] は国語辞典とテキストコーパスの組み合わせている。

- ヒューリスティクス
  文字単位の DP マッチング

2.4 各尺度の評価

これまで報告されている各尺度の比較を見てみる。

Lee[3] は、動詞と目的語の (n,v) ペアを決定するタスクに確率モデルベースの尺度[6]を使って比較実験を行っている。それによると $\alpha SKEW$ がもっとも性能が良く、ついで $Jaccard, IR, L_1$ の順であった。

また、Lin は [5] 係り関係の共起 (w1,w2,f) をベクトルモデルの尺度で比較することでシソーラスを構築し、既存のシソーラスとの比較を行っている。それによると、cosine, dice, Jaccard の間にほとんど差はなく、MI がわずかに性能が良かった。

3 システムの構成

以上のように様々な選択肢がある中で、今回我々はタスク毎にどの類似度尺度が最適であるかを簡単に比較することを目的としてツールの作成を行った。

考慮した点:

- 拡張性
  今後新しい類似度を定義した場合でも容易にシステムに組み込めるように object oriented にする。

- 汎用性
  どんなコーパスでも解析できるように一般的な日本語の前処理である形態素解析 ChaSen [16] + 係り受け解析 CaboCha[19] を仮定する。またシソーラスにおいては分類、上位下位どちらも扱える中間形式となる class を定義する。

システムの入出力は以下の通りである。

- 入力
  ChaSen+ CaboCha の解析済みコーパス
  およびシソーラスデータ (角川、分類語彙表、NTT 語彙体系)

- 出力
  $SIM(w_i, w_j)$ が要素 (i,j) であるような類似度マトリックス

- パラメータ
  頻度足切り (n)
  各モデルごとのパラメータ (window size など)、各計算サイクル毎に変更するもの

システムの概略図を以下に挙げる。

section 2 で示した 3 つの分類のうち、現在コーパスから計算できるベクトルモデルおよび確率モデルの類似度尺度の部分のインプリメントが終了している。なお実装は C++ で行っている。

[6] L_2, tau, conf, L_1, IR, Jaccard, α SKEW の 7 つ

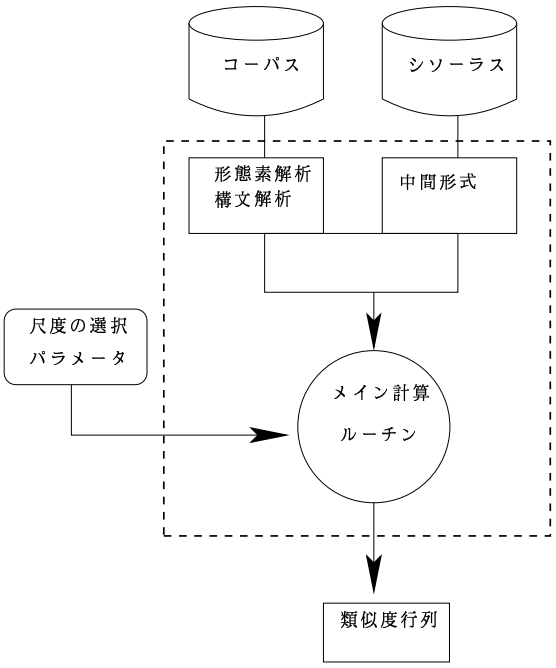

図 2: システムの構成

4 まとめ

現在提案されている様々な類似度尺度を簡便に比較するためのツールの作成を行った。これを用いることで類似度尺度を簡便に比較できるようになった。

今回は各尺度の評価は行わなかった。それは類似度単体での客観的な評価方法がないからである。それは section 1 で述べた通り、アプリケーションの精度向上に類似度尺度がどれだけ寄与したかという視点で行わなければならない。

4.1 Future Work

今後以下のような作業を予定している。

- 高速化

- GUI インターフェースの製作

- コーパスベース以外の尺度計算の実装
  角川類語新辞典、分類語意表、NTT 語彙体系等、現在利用できるシソーラスやオントロジーなどを一般的な形で取り込み、コーパスベースの類似度と比較出来るようにする。

- モジュール化
  将来的には計算部分を独立したモジュールと

し、アプリケーションとなるシステムにそのまま組み込めるようにする。

4.2 多義性について

人手の知識からの尺度において、今回は意味が一意に定まると仮定しているが現実には多義性をもち、同じ単語が木の複数の場所に現れる。その場合、最小値をとる方法、平均値をとる方法、中央値をとる方法、重みづけなどの方法などを考慮する必要がある。

4.3 Euclid vs Riemann

ここで述べた距離空間は全てユークリッド空間を仮定しているが、その妥当性の保証はどこにもない。むしろある種の次元は他の次元に比べより重要な役割を担っているかも知れず、もっと一般的なリーマン空間を導入するメリットは大いにあると考えられる。その場合計量の決定が課題となる。あるいは、確率分布を点と見なして距離を定義する情報幾何 [12] を援用することも考えられる。

5 謝辞

本研究は通信・放送機構の研究委託「大規模コーパスベース音声対話翻訳技術の研究開発」により実施したものである。

参考文献

[1] Robert Dale, Hermann Moisl, and Harold Somers, editors. *Handbook of Natural Language Processing, Chap.19.* Marcel Dekker, 2000.

[2] Donald Hindle. Noun classification from predicate-argument structures. In *28th Annual Meeting of the Association for Computational Linguistics * Proceedings of the Conference*, pp. 268–275, 1990.

[3] Lillian Lee. Measures of distributional similarity. In *37th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (ACL) *University of Maryland, USA*, pp. 25–32, 1999.

[4] Lillian Lee and Fernando Pereira. Distributional similarity models; clustering vs. nearest neighbors. In *37th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (ACL) *University of Maryland, USA*, pp. 33–40, 1999.

[5] Dekang Lin. Automatic retrieval and clustering of similar words. In *COLING-ACL '98, Proceedings of the Conference, Vol.2 *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pp. 768–774, 1998.

[6] Dekang Lin. An information-theoretic definition of similarity. In *Machine Learning *Proceedings of the Fifteenth International Conference (ICML '98)*, pp. 296–304, 1998.

[7] Manning and Schuetze. *Foundations of Statistical Natural Language Processing, Chap.8.* The MIT Press, 1999.

[8] George Miller and Charles Walter. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, pp. 6:1–28, 1991.

[9] Fernando Pereira, Naftali Tishby, and Lillian Lee. Distributional clustering of english words. In *31st Annual Meeting of the Association for Computational Linguistics * Proceedings of the Conference*, pp. 183–190, 1993.

[10] Eiichiro Sumita and Hitoshi Iida. Experiments and prospects of example-based machine translation. In *29th Annual Meeting of the Association for Computational Linguistics * Proceedings of the Conference*, pp. 185–192, 1991.

[11] Jean Veronis and Nancy M. Ide. Word sense disambiguation with very large neural networks extracted from machine readable dictionaries. In *COLING-90: 13th International Conference on Computational Linguistics, Helsinki *Vol.2*, pp. 389–394, 1990.

[12] 甘利俊一, 長岡浩司. 情報幾何の方法. 岩波講座 応用数学 対象 12. 岩波書店, 1993.

[13] 稲子希望, 笠原要. 国語辞典とテキストコーパスを用いた単語の類似性判別. 情報処理学会論文誌 Vol.43,No.10, pp. 3239–3242, 2002.

[14] 池原悟, 村上仁一, 木本泰博. 単語意味属性を使用したベクトル空間法. 自然言語処理, Vol.10 No.2, pp. 111–128, 2003.

[15] 国立国語研究所. 分類語彙表. 国立国語研究所資料集 6. 秀英出版, 1993.

[16] 松本, 北内, 平野, 松田, 高岡, 浅原. 形態素解析システム 茶筌 version 2.2.9 使用説明書. 奈良先端科学技術大学院大学, 2002.

[17] 崔進, 小松英二, 安原宏. EDR 電子化辞書を用いた単語類似度計算方法. 情報処理学会自然言語処理研究報告 NL-93-01, 1993.

[18] 大野晋, 浜西正人. 角川類語新辞典. 角川書店, 1981.

[19] 工藤拓, 松本裕治. チャンキングの段階適用による日本語係り受け解析. 情報処理学会論文誌, Vol. 43, No. 6, pp. 1834–1842, 2002.

[20] 池原, 宮崎, 白井, 横尾, 中岩, 小倉, 大山, 林（編）. 日本語語彙体系. 岩波書店, 1997.

[21] 長尾真. 自然言語処理. 岩波講座 ソフトウェア科学 15. 岩波書店, 1996.

[22] 日本電子化辞書研究所. EDR 概念辞書. 日本電子化辞書研究所, 1995.

[23] 笠原要, 稲子希望, 加藤恒昭. 単語の属性空間の表現方式. 人工知能学会誌, Vol.17, No.5, pp. 539–547, 2002.