社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－ＦＩ－ 84 (13)
2006－ＮＬ－175 (13)
2006／9／13

決定リストを利用した単語間の類似度計算法

川上 高志 †, 鈴木 寿 ††

† 中央大学大学院 理工学研究科 情報工学専攻
†† 中央大学 理工学部 情報工学科
〒112-8551 東京都文京区春日 1-13-27
E-mail: † kawakami@suzuki-lab.ise.chuo-u.ac.jp, †† suzuki@ise.chuo-u.ac.jp

近年, 形態素解析や構文解析の精度向上により, 与えられたテキストから高い精度で単語を切り出すことが可能になっている. 本研究では, 切り出した単語の意味関係を獲得することを目的とする. そこで, 任意のテキストから単語間の類似度を計算する手法を提案する. 具体的には, クラス分類問題の一つである擬似単語分類問題を正確に解く. そして, 未分類となった解の割合を求めることで, 類似度を計算する. ここでは, 擬似単語分類問題の解法に決定リストを利用する. 実験では, インターネット上に公開されている小説データから単語間の類似度を計算した. 実験結果から, 関連語や表記の揺れなど意味の似た単語間の類似度が高くなることを確認した.
キーワード: 単語間の類似度, 擬似単語分類, 決定リスト.

A calculation of Word Similarity using Decision List

Takashi KAWAKAMI † and Hisashi SUZUKI ††

† Graduate School of Science and Engineering, Chuo University
†† Faculty of Science and Engineering, Chuo University

This article proposes a method of evaluating sharply the similarity between words as one of classification problems given a sample of text, e.g., essay, story, novel, etc. The first process for evaluation is getting, on a decision list, an exact solution of the "pseudo classification" problem of words, and the latter process is calculating the similarity between words by using the ratio of the unclassified words. Given an on-web novel as a practical sample of text, the experimental result shows that the similarity between equivalent-meaning words such as some related terms or morphologically-varied expressions is comparatively high.
Keywords: Word Similarity, Pseudo Word Classification, Decision List.

1 はじめに

本研究は, 単語間の類似度計算法を提案する. 提案手法では, 単語のクラス分類問題を解くことにより, 単語間の類似度を計算する. このクラス分類問題は決定リストを利用して解く.

単語の類似性の数値化は, 自然言語処理において重要な処理である. 例えば, 単語の類似性を用いることで, 情報検索の再現率が向上することなどが知られている. また, 単語の類似性を数値化できれば, シソーラスやオントロジーといった言語知識を自動獲得することができる.

単語の類似性には, 目的に応じて様々な定義や手法が存在する [1]. 本研究では, 特に同じ意味のカテゴリに属す単語の関係を扱う. この関係を数値化したものを単語間の類似度と定義する. 例えば, 「夏」と「冬」は「季節」という同じ関係に属すため類似度は高くなる. また, 「夏」と「寺」にはこのような関係がないため類似度は低くなる.

単語間の類似度は, 人手で作成した辞書 (シソーラスなど) をベースに計算される. しかし, 辞書を用いた手法は, 大きく二つの問題がある. 一つは, 辞書に登録されている単語は固定であるということである. そのため, 辞書に登録されていない単語に関しては, 類似度を計算することができない. もう一つは, 言語知識の自動獲得を目的とする場合, 人手で作成した辞書を用いるのは都合が悪い.

これらの問題に対処するために, 任意のテキストから単語間の類似度を計算する手法を提案する.

提案手法では, 類似した単語は, 類似した文脈でテキスト中に現れると仮定する. この仮定に基づいて, 日本語の擬似単語分類問題 [2][3] を解く. 擬似単語分類問題とは, 複数の単語をシステムの側からは単一の単語にしか見えないようにしておき, どの単語であるのかを文脈から分類させるというクラス分類問題である. 仮定より, 類似した文脈からはどの単語であるのかを分類できないと考えられる. そして, このとき未分類であった単語の割合を単語間の類似度と定義できる. このように考えることで, 単語間の類似度計算を, 単語のクラス分類問題として扱うことができる.

類似した文脈がクラス分類問題で未分類となるのは, クラス分類器が正確な分類をする場合のみと考えられる. それは, 誤りを含むクラス分類器は, 未分類となるべき解もいずれかのクラスに分類してしまうためである. しかし, 現状において, この要求を十分に満たすクラス分類器は存在しない. そこで, この要求を満たすクラス分類器を作成する必要がある.

提案手法では, 擬似単語分類問題と決定リスト [4][5] を組み合わせることで, 正確な分類をするクラス分類器を作成する. 決定リストは, 学習結果が if-then 形式のルールの並びであるために, 人間が容易に理解できる. そのため, 学習した決定リストを修正することができる. これは, 他のクラス分類器の学習結果がブラックボックスである点に対する大きなメリットである.

以下, 2 節で提案手法における単語間の類似度についての基本的考え方や計算方法, 3 節で提案手法における計算機実験について述べる. 4 節で考察を行い, 最後に 5 節でまとめを行う.

2 単語間の類似度計算法

2.1 提案手法の基本的考え方

テキストに含まれる表層的な情報から, どのようにして単語間の類似度を得ることができるだろうか. ここで, 次の問題を考えてみよう.

問 1. 以下の例文に現れる x を「海」もしくは「川」に置き換えて意味が通るようにせよ.

問 2. 以下の例文に現れる x を「海」もしくは「寺」に置き換えて意味が通るようにせよ.

問 3. 以下の例文に現れる x を「海」もしくは「電車」に置き換えて意味が通るようにせよ.

1. x に海水浴へ行く。
2. x の水は塩辛い。
3. 今年も x で泳ぐ。
4. x の水は冷たい。
5. 今年の夏に x へ行く。

問 1. は 1, 2 番が x を「海」に置き換えることで意味が通るようになり, 3, 4, 5 番は, 「海」でも「川」でも文の意味が通る. 同様に, 問 2. は, 4, 5 番が, 「海」でも「寺」でも文の意味が通り, 問 3 は全ての例文において, 「電車」に置き換えると文の意味が通らなくなる.

逆に, それぞれの問について, x がどちらの単語になるか, 文脈から判別する問題を考えてみよう. 問 3. は, すべての例文の x が「海」に置き換わると判別できる. しかし, 問 1. や問 2. は, どちらの単語でも文の意味が通る例文が存在する. そのため, 問 1. の 3, 4, 5 番や問 2. の 4, 5 番は, x がどちらの単語になるか判別できない. このように, 対象とする単語の組み合わせの違いで, 同じ文脈でも単語の判別結果が異なることがわかる.

このとき, 例題の単語の組み合わせが, どのような関係に属しているだろうか. 「海」と「川」は「自然」という同じ関係に属し, 「海」と「寺」は「場所」という同じ関係に属す. また, 「海」と「電車」は同じ関係に属さない. このとき, 「自然」は, 「場所」よりも具体的な概念といえる. つまり, 「海」と「川」の組み合わせは, 「海」と「寺」よりも類似した関係である.

以上のことから, 「類似した意味を持つ単語は, 類似した文脈でテキスト中に現れる」という仮説を考えることができる. そして, 類似した文脈からは, 単語の分類ができない傾向にあるといえる. Miller らは, 文脈の類似性が, 単語の意味的な類似性に寄与するという同様の仮説を, 従来研究において指摘している [6]. この仮説に基づいて, 擬似単語分類問題を解き, その結果から未分類であった単語の割合を求め, その割合を単語間の類似度 δ とする. 例えば, 問 1. は, 5 文中 3 文が未分類になるので, 「海」と「川」の類似度 δ は $\delta = 0.6$ である. 同様に, 問 2. 「海」と「寺」は $\delta = 0.4$, 問 3. 「海」と「電車」は $\delta = 0.0$ となる.

2.2 提案手法の詳細

提案手法では, 任意の 2 つの単語 w_1, w_2 が入力されると, 以下の手順に基づき, 単語間の類似度

$\delta(w_1, w_2)$ を計算する．以後，類似度を計算する対象となる単語を対象単語 w_i と呼ぶ．

1) 任意のテキストから対象単語 w_i を検索し，各対象単語 w_i を中心とした前後 n 単語の文脈を，適当にそれぞれ m 文集める．
2) 集めた文脈の対象単語 w_i を，システムの側からは単一の単語にしか見えないようにする．ただし，正解は確認用に保持しておく．
3) 集めた文脈の γ% を選び，決定リストを作成するための初期学習用データとする．
4) 学習用データから決定リストを作成する．
5) 決定リストを用いて未分類の単語を分類する．
6) 分類結果と確認用の正解を比較する．誤りが含まれているならば，決定リストを修正し，今回の分類結果を無効にして，手順 5 に戻る．
7) 新たに分類された単語を含む文脈を，学習用データに加え，手順 4 に戻る．新たに分類された単語がないならば，手順 8 へ進む．
8) 分類結果から未分類である単語の割合を求め，類似度 δ を算出する．

ただし，入力する 2 つの単語が同じ場合は，類似度を求める必要がないので，以上の手順を適用しない．以下で，手順 4，手順 5，手順 6，手順 8 について詳しく述べる．

2.2.1 決定リストの作成

対象単語 w_1, w_2 と学習用データから決定リストを作成する．以下の手順で決定リストを作成する．

1) 各対象単語 w_i と証拠 e_j とが同時に現れる頻度 $f(w_i, e_j)$ を，学習用データから得る．

提案手法では，決定リストで用いる文脈上の特徴として，以下のものを設定する．

- 直前の単語 w: $w-$ と表記する．
- 直後の単語 w: $w+$ と表記する．
- 前方に現れる単語 w: 前方 n 単語内に現れる単語を取り出し，それぞれ $w-n$ と表記する．
- 後方に現れる単語 w: 後方 n 単語内に現れる単語を取り出し，それぞれ $w+n$ と表記する．

例えば，対象単語を「海」，「電車」とし，以下の 2 つの例文をみる．

例文 1 「今年も海で泳ぐ。」
例文 2 「満員電車に乗る。」

例文 1 からは，「海」に対する証拠として，
“今年 $-n$”, “も $-$”, “で $+$”, “泳ぐ $+n$”, “。 $+n$”
が取り出される．例文 2 からは，「電車」に対する証拠として，
“満員 $-$”, “に $+$”, “乗る $+n$”, “。 $+n$”
が取り出される．

2) 証拠 e_j が生じた場合の信頼度 $L(w_1, w_2, e_j)$ を以下の式で求める．

$$L(w_1, w_2, e_j) = \log \frac{P(w_1|e_j)}{P(w_2|e_j)} \tag{1}$$

ここで，$P(w_i|e_j)$ を近似して以下のようにする．

$$L(w_1, w_2, e_j) = \log \frac{f(w_1, e_j) + \alpha}{f(w_2, e_j) + \alpha} \tag{2}$$

α は，式 (2) の分母が 0 の場合の不具合を回避するために設定する．提案手法では，$\alpha = 0.1$ とする．この値は最も簡単な実装であるが，十分な結果を達成できると Yarowsky によって示されている [4].

3) 証拠 e_j の解答が w_1, w_2 のどちらであるかを，求めた信頼度 L から判別する (表 1).

$$\begin{cases} w_1 & |L| > \beta,\ L > 0 \\ w_2 & |L| > \beta,\ L < 0 \\ \text{どちらでもない} & \text{otherwise} \end{cases}$$

ただし，信頼度 L の絶対値が，β 以下のものはリストから外す．β の値は，実験的手法もしくは最尤推定などで厳密に決めるべきである．しかし，提案手法では，2.2.3 節の方法で決定リストの信頼性を高めている．そのため，暫定的に $\beta = 0.1$ とした．

4) 信頼度 L の絶対値が高い順のリストを作成する．これが決定リストとなる．

以上より，対象単語「海」，「電車」に対して，表 2 のような決定リストが得られる．

2.2.2 決定リストの利用

実際に決定リストを用いて，擬似単語分類問題を解く．まず文中から証拠を取り出す．そして，取り出した証拠 e_j が，作成した決定リストに含まれているかどうかを調べる．もし，e_j が決定リストに含まれていれば，e_j に対する解答 w_i が分類結果となる．該当する e_j が複数存在する場合は，最も信頼度の値が大きい e_j を解答とする．

例えば，表 2 の決定リストを用いて，以下の文の x が「海」と「電車」のどちらになるか分類する．

例文 3 「今年の夏も x へ海水浴に行く。」

表 1　証拠に対する解答と信頼度の例

証拠 e_j	「海」との頻度	「電車」との頻度	解答	信頼度 $\|L\|$
乗る $+n$	11	90	電車	2.787
行く $+n$	55	50	海	0.095
…	…	…	…	…
海水浴 $+n$	70	1	海	4.154
…	…	…	…	…
満員 $-$	20	0	電車	5.303
夏 $-n$	120	12	海	2.295
…	…	…	…	…

表 2　作成できた決定リストの例

信頼度 L	証拠 e_j	解答
5.303	満員 $-$	電車
…	…	…
4.154	海水浴 $+n$	海
…	…	…
2.787	乗る $+n$	電車
2.295	夏 $-n$	海
…	…	…

例文 4「大阪の x で船に乗る。」

（例文 3, 例文 4 ともに正解は「海」とする．）

例文 3 から取り出された証拠の中で，“海水浴 $+n$” が最も大きい信頼度を持ち，“海水浴 $+n$” の解答は「海」であるので，例文 3 の x は「海」に分類される．例文 4 も同様に，“乗る $+n$” が最も大きい信頼度をもち，“乗る $+n$” の解答は「電車」であるので，例文 4 の x は「電車」に分類される．x の正解と分類結果 w_i が等しければ正しい分類であり，等しくなければ誤った分類である．

2.2.3　決定リストの修正

2.2 節の手順 2 で記録した正解と分類結果 w_i が等しくない場合，誤った分類である．このとき，解答を w_i とした証拠 e_j を決定リストから取り除く．

ここで，2.2.2 節の例文 4 を見る．例文 4 の正解は「海」である．しかし，表 2 の決定リストを用いた例では，「電車」と誤った分類をした．この分類は，“乗る $+n$” という証拠によって分類された．したがって，“乗る $+n$” を表 2 の決定リストから取り除く．

2.2.4　類似度の算出

対象単語 w_1, w_2 について擬似単語分類を行い，単語 w_1 に分類された結果の集合を C_{w_1}，単語 w_2 に分類された結果の集合を C_{w_2}，分類されなかった結果の集合を C_{other} とする．以下の式を用いて類似度 $\delta(w_1, w_2)$ を計算する．

$$\delta(w_1, w_2) = \frac{|C_{\mathrm{other}}|}{|C_{w_1}| + |C_{w_2}| + |C_{\mathrm{other}}|} \qquad (3)$$

$\delta(w_1, w_2)$ は，0 から 1 の範囲を取り，1 に近いほど 2 つの単語は類似する．

3　計算機実験

本研究では，インターネット上に存在するデータをテスト用テキストとして用いて，提案手法の有効性を確認する．今回は，テスト用テキストとしてインターネット上に公開されている著作権切れの小説データを用いることにした．これらの小説データは，夏目漱石，芥川龍之介ら 12 著者，955 タイトルである．

実験の内容としては，まず類似度を計算する対象単語を選ぶ．対象単語は，テスト用テキストに 500 回以上現れる名詞とし，実験で用いる対象単語は 609 個である．対象単語の 2 つの組み合わせ全てに対して，類似度を計算する．

対象単語を中心とした，前後 10 単語の文脈を適当に 500 文集める．そして，無作為に選んだ 50 文 (10%) を，決定リスト作成の初期学習に用いる．残りの 450 文を，決定リストによる分類と決定リストの修正に用いる．

図 1 は，実験で求めた類似度の分布を示す．横軸は，類似度を 10 区間に分割したもので，縦軸は，各区間に現れた頻度に対して常用対数をとったものである．実験で求めた類似度の上位 100 組を表 3 に示す．

4　考察

表 3 から，与えたテキストからいくつかの関係を取り出したことがわかる．「だめ」と「駄目」，「だれ」と「誰」といった漢字とその読みという関係，「4」と「8」，「夏」と「秋」といった同一のカテゴリに属す関係，「妙」と「変」，「妙」と「不思議」といった同義語の関係，「駄目」と「大丈夫」といった対義語の関係などである．いずれの関係

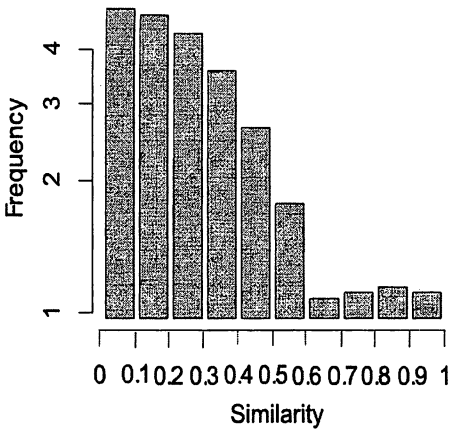

図 1　実験で求めた類似度の分布

も, 広い意味で何らかの同じ分類に属している. したがって, 提案手法は意味の近い単語の組み合わせを求めるのに有効である.

今回の実験では, 対象単語 609 個の組み合わせ全てに対して類似度を計算した. その件数は 185,136 件である. ここで, 類似度 δ が 0.5 以上のものは 112 件, 0.3 以上のものは 4,929 件であった. 図 1 からも, 求めた類似度のほとんどは 0.3 以下であることがわかる. このことから, 単語の関係を自動獲得する目的において, 全ての組み合わせに対して類似度を求めることは非常に効率が悪いといえる.

提案手法では, 決定リストに用いる信頼度を, Yarowsky[5] が用いた式 (1) で計算した. しかし, 鶴岡 [3], 李 [7] や新納 [8] は, 分類の正解率が改善される信頼度計算法を提案している. これらの計算方法を用いた場合についても検討の余地がある. また, 将来的には他の分類手法も類似度の計算に用いることができると考えられる.

提案手法は, 単語の使用例を一定数以上収集できるならば, 単語間の類似度を計算することができる. このことから, Web 検索と組み合わせることも容易である.

5　おわりに

本研究では, 任意のテキストから単語間の類似度を計算する手法を提案した. 具体的には, 擬似単語分類問題を誤りなく解き, その時に現れる未分類となった解の割合を求めることで類似度を計算する. また, 擬似単語分類問題の解法には, 決定リストを利用した. インターネット上に公開されている著作権切れの小説データを用いた実験により, 同義語, 対義語, 関連語や 表記の揺れなど意味の似た単語間の類似度が高くなることを確認した.

しかし, 提案手法は類似度の高い組み合わせが, 同義語であるのか対義語であるかといった関係の区別ができない. この区別は, 知識の自動獲得では重要であり, 今後の研究課題の一つといえる.

参考文献

[1] 河部恒, 柏岡秀紀, 田中英輝, 松本裕治. 単語類似度の尺度比較支援ツールの作成. 情報処理学会研究報告, 2003-NL-156, pp.39-44 (2003) .

[2] 鶴岡 康雅, 近山 隆. 単語多義性における漸進的なコーパス自動タグ付け. 第 58 回情処全大, Vol.2, pp.57-58 (1999).

[3] 鶴岡 慶雅, 近山 隆. ベイズ統計の手法を利用した決定リストのルール信頼度推定法. 自然言語処理, Vol. 9, No. 3, pp3-20 (2002).

[4] David Yarowsky. Decision Lists For Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French. in *Proceedings of the 32nd Anual Meeting of the ACL*, Las Cruces, NM, pp.88-96, (1994).

[5] David Yarowsky. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. in *Proceedings of the 33rd Annual Meeting of the ACL*, Cambridge, MA, pp.189-196 (1995).

[6] George Miller and Charles Walter. Contextual correlates of semantic similarity, Language and Cognitive Processes, Vol. 6, No. 1, pp. 1.28 (1991).

[7] 李 航, 竹内 純一. 証拠の強さと信頼度を考慮した日本語同形異音語の読み分け. 情報処理学会研究報告, 1997-NL-119, pp.53-59 (1997).

[8] 新納 浩幸. 複合語からの証拠に重みをつけた決定リストによる同音異義語判別. 情報処理学会論文誌, Vol.39, No.12, pp.149-156 (1998).

表 3　単語間の類似度を計算する実験の結果 (上位 100 組)

w_1	w_2	δ	w_1	w_2	δ	w_1	w_2	δ
4	8	0.998	3	9	0.733	三	二	0.559
4	5	0.992	1	5	0.727	きょう	今夜	0.557
1	4	0.962	3	6	0.725	きょう	今日	0.557
6	7	0.958	0	2	0.723	今	今日	0.556
2	4	0.953	1	6	0.702	兄	父	0.556
8	9	0.952	5	9	0.701	ここ	そこ	0.555
5	6	0.947	だめ	駄目	0.701	公開	発行	0.555
2	6	0.939	だれ	誰	0.672	不思議	妙	0.547
2	8	0.938	1	7	0.666	お前	俺	0.544
2	3	0.927	作成	公開	0.655	先生	奥さん	0.543
3	4	0.925	母	父	0.645	どこ	誰	0.541
3	5	0.921	おまえ	お前	0.644	三	五	0.539
7	8	0.917	0	8	0.642	みんな	皆	0.537
1	2	0.900	0	3	0.625	僕	自分	0.537
5	7	0.899	夏	秋	0.624	二	五	0.535
2	7	0.898	あれ	これ	0.622	兄	母	0.532
2	5	0.888	千	百	0.612	健三	細君	0.530
3	8	0.884	彼	彼女	0.602	これ	今	0.530
1	3	0.884	ほんとう	本当	0.601	わし	俺	0.530
6	8	0.878	ちがい	違い	0.599	あいつ	これ	0.529
6	9	0.873	変	妙	0.595	ここ	これ	0.528
3	7	0.864	これ	それ	0.594	これ	そこ	0.527
4	7	0.851	いま	今	0.584	三	四	0.527
4	6	0.840	はず	筈	0.583	何	誰	0.527
1	8	0.839	おれ	俺	0.581	妻	母	0.526
5	8	0.834	年	月	0.580	いや	だめ	0.525
1	9	0.819	二	四	0.572	五	四	0.524
7	9	0.800	わけ	訳	0.567	冬	秋	0.524
4	9	0.774	これ	今日	0.563	なに	誰	0.514
0	5	0.772	あたし	お母さん	0.562	彼	彼等	0.514
0	4	0.762	彼ら	彼等	0.562	僕	君	0.513
2	9	0.747	大丈夫	駄目	0.560	冬	夏	0.513
お母さん	お父さん	0.736	だめ	大丈夫	0.559	彼	自分	0.513
						女	男	0.511