Vol.2013-ICS-170 No.12
2013/3/12

EMD を用いた英字略語の意味判断システム

田邊僚[†1]　吉村枝里子[†2]　土屋誠司[†2]　渡部広一[†2]

近年，国際化・情報化が進む中，文書中に多用されている英字表現の略語である英字略語（例えば「Compact Disk」を「CD」と表現）に対する言い換え手法を既に我々は提案し，多義を有する英字略語（200 個）に対して意味判断を行った結果 62 %の正答率を実現した．本稿ではさらに，Earth Mover’s Distance を用いた意味候補選択手法を既提案システムに加えることにより，意味判断精度 72%を実現した．

The Judgement System using EMD for Abbreviated Alphabetical Characters

RYO TANABE[†1]　ERIKO YOSHIMURA[†2]
SEIJI TSUCHIYA[†2]　HIROKAZU WATABE[†2]

Recently, In this global and informational society, as one of the method for paraphrase, the authors had proposed the way to paraphrase alphabetical characters (ex: “Compact Disk” -> “CD”) which exist in texts frequently. And the system had showed 62% of questions answered correctly with 200 abbreviated alphabetical characters which have ambiguous meanings. In this thesis, the author added such as “A method for selecting candidates of meaning with Earth Movers Distance”. As a result, the system showed 72% of questions answered correctly.

1. はじめに

現在，コンピュータは目覚ましい速度で普及が進んでおり，デスクトップ型やノート型のコンピュータ以外にもスマートフォンやタブレット端末など形を変えた状態で，老若男女問わず人々の生活に深く浸透している．同時にコンピュータは人間の代わりに思考，または実行する道具として生活に必要不可欠な存在となっている．そのため，今後はより人間の代わりに知的な情報処理を行える次世代コンピュータの実現が期待される．現在その知的な情報処理の一つとして，スマートフォンやタブレット端末の普及による電子書籍や電子新聞など様々な電子文書に触れる機会が増えている現状から，コンピュータが文書をユーザが求める形に変換（言い換え）処理することが望まれている．

上記のようなニーズに応えるため，近年，言い換え処理の研究がさかんに行われている[1]．例えば，言い換え知識を自動で獲得する手法[2]，普通名詞の言い換え手法[3]，動詞句の言い換え手法[4]がある．これらの研究は，「言い換え処理」を機械翻訳・情報検索の前処理として用いることにより，機械翻訳では翻訳しやすく，情報検索ではより的確な回答を発見しやすくすることを実現した．しかし，機械翻訳・情報検索というコンピュータ処理を目的とした言い換え処理のため，人間が理解しやすい形に語彙・構文を変換することはできない．

人間が理解しやすい形に言い換え処理を行う研究として，聾唖者に理解し易いテキスト言い換える手法[5]がある．この研究は幅広いユーザに対応することは出来ていないが，特定の人間に合わせた文書言い換え処理を可能とした．さらに多くのユーザに合わせた文書言い換え処理をするため，文体の難易度制御をしつつ日本語機能表現を言い換える手法[6]，語彙に親密度を設定することで，語彙の言い換え変換を行う手法[7]がある．どちらの研究においても難易度を設定することにより，老若男女のユーザの理解度に合わせた形に変換が可能としているが，辞書やコーパス内に語彙の意味が存在しない場合には対応できないという問題がある．特に本来の形の文字列を省略した表現である略語表現は難易度変換の障害となる．また，略語表現は専門的な用語を表すことが多く，専門分野以外の人間には表現の意味を理解することが難しいとされている．

そこで，略語表現の難易度変換には本来の表現に言い換える手法である日本語略語を自動復元する手法[8]，カタカナ語省略形を復元する手法[9] ，略語とその原型語との対応関係のコーパスからの自動獲得手法[10]などを用いた後に難易度変換を行うことで，上記の問題に対応することができる．さらに，日本語表現・片仮名表現のみならず国際化・情報化が進む中，文書中に多用されている英字表現における略語である英字略語（例えば「Compact Disk」を「CD」と表現）には対応する研究[11]も進められている．具体的には，括弧表現に着目し「CD（コンパクトディスク）」のように英字略語の言い換え可能な語彙対（意味）を自動的に獲得する手法である．この自動取得手法では，過去に出現した英字略語の意味を英字略語とセットでデータベースに貯めることにより，英字略語の意味を判断している．しかし，この判断手法には大きく二つの問題が存在する．一

†1　同志社大学大学院　工学研究科
Graduate School of Engineering, Doshisha University
†2　同志社大学　理工学部
Faculty of Science and Engineering, Doshisha University

つは，括弧表現が英字略語の後ろに存在しない場合，過去に同じ英字略語の意味を取得していなければ判断できないという点．もう一つは，英字略語が多義を有する場合，複数の意味候補から正確な意味を選択ができない点である．これら二つの問題点から，過去に出てきていない英字略語へ対応するため，Wikipedia[12]から意味候補を取得する．そして，文書全体の内容を考慮することで，その内容に適した英字略語の意味を出力する我々の既提案システム[13]がある．既提案システムでは，144 個の英字略語に対する意味判断精度は約 71%であったが，多義を有する英字略語に限定（77 個）すると約 61%と判断精度が低いという問題点があった．そこで本稿で提案するシステムでは，我々の既提案システムを改良し，多義を有する英字略語の意味判断精度を既提案システムに比べて高い判断精度にすることを目指した．

多義を有する英字略語への判断精度向上のため，本提案システムと既提案システムとでは，新聞記事からの取得品詞，意味候補の絞込み手法，Earth Mover’s Distance[14]（以下 EMD）を用いた意味候補選択手法の三点が異なる．本稿では，前記三点の変更点の詳細とそれぞれでの意味判断精度の評価を示す．

2. 提案システムの概要

はじめに既提案システムの説明を 1～4 の順を追って説明する．

1．英字略語の意味候補を Wikipedia より取得．

2．意味候補に AF[15]を用い，意味候補に関連する自立語（属性）とその重みをそれぞれ取得．

3．英字略語を含む記事自体に重み付け．

4．英字略語の意味を選択．

はじめに 1 で英字略語の意味候補を Wikipedia より取得する．ここで，百科事典や国語辞書を用いないのは，英字略語の専門性の高さから辞典・辞書には意味が掲載されておらず，意味候補を取得できないことが多いからである．

次に 2 で意味候補に AF を用い，意味候補に関連する自立語（属性）とその重みをそれぞれ取得する．この行程により，意味候補を 2 つの概念間（既提案システムにおける意味候補と新聞記事）の関連性の計算をする際，必要となる概念ベース[16]に対応した形（自立語＋重み）で表すことができる．

また同時に，3 で英字略語を含む記事自体に重み付けを行う．記事への重み付けには記事の形態素解析後，記事自身の特徴を表すため名詞，動詞，形容詞に *tf･idf* 重み付け[17]を行う．

そして最後 4 で英字略語の意味を選択する．意味選択には，2 の結果それぞれと 3 の結果から概念ベースを用いて二つの概念間で記事関連度計算[18]を行う．記事関連度計算により，概念間の関連性が数値化されるため，その結果から数値が高い意味候補を新聞記事の意味として選択する．

以上が既提案システムの概要である．

一方で，構築した本提案システムである多義を有する英字略語の意味判断システムの構成を図 1 に示す．本提案システムでは，既提案システムと比較すると，多義を有する英字略語への判断精度向上のため，既存システムから図 1 における①～③の三点の変更を加えた．

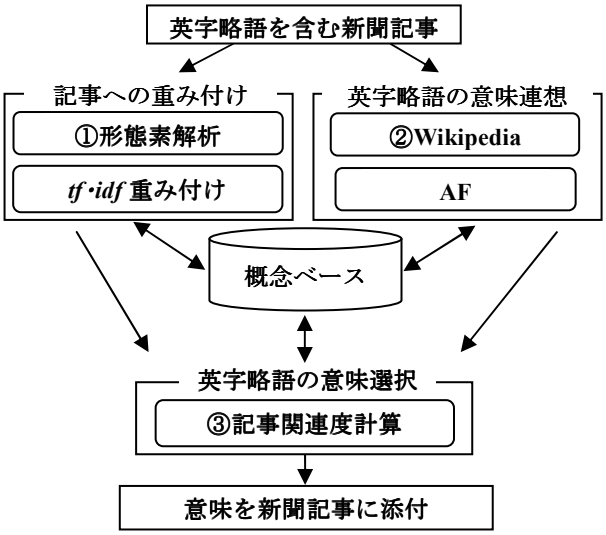


図 1　本提案システムの流れ

既提案システムでは記事への重み付けの際，記事の形態素解析結果から名詞，動詞，形容詞全てを取得し，*tf･idf* 重み付けを行っていた．しかし，記事関連度計算を行う対の相手である英字略語の意味連想結果（AF 結果：名詞）と比較すると動詞，形容詞が雑音となってしまい判断精度を下げてしまう可能性があった．そこで，一つ目の変更として本稿では記事の名詞のみに重み付けを行い取得した（①）．

二つ目は，英字略語の意味連想の際，Wikipedia で得られる英字略語の意味候補から型番や記号を削除する意味候補の絞り込み手法を取り入れた（②）．これは，既提案システムにおいて意味判断を行った結果，英字略語の意味として成りえない型番や記号などの候補が記事内の英字略語の適切な意味として誤って出力されてしまう問題が存在したからである．この問題に対応するため，前述の形で意味候補を絞り込む変更をした．

三つ目は，英字略語の意味候補選択における記事関連度計算を行う際，記事関連度計算と EMD を用いた記事関連度計算を場合分けする手法へ変更をした（③）．記事の名詞と AF 結果との間での関連度計算では，記事の名詞の個数と AF 結果の個数との間で個数差（属性数差）が少ない場合，適切な計算結果を得られるが，属性数差が多い場合は適切な計算結果が得られないという特徴がある．この特徴により，既提案システムシステムでは，属性数差が多い場

合の判断精度が低くなってしまう問題が存在した.そこで,属性数差が多い場合には,属性数に差がある場合においても柔軟に対応することが出来る EMD を用いた記事関連度計算を用いる.

3 章で全ての工程に用いられる概念ベース,4 章で①の変更を含めた記事への重み付け,5 章で②の変更を含めた英字略語の意味連想,6 章で③の変更を含めた英字略語の意味選択について述べ,7 章で提案システムの意味判断精度評価を行う.

3. 概念ベース

概概念ベースとは,複数の国語辞書から作成されており,常識的な判断を行うために構築された,語の知識の集合である.概念とは常識的な判断を行うための語で,概念はその語の意味特徴を表す単語(属性)と重みの対の集合として表現される.重みは概念の特徴をより的確に表現している属性には高い重みが,そうでない属性には低い重みが付与されている.属性には必ず概念そのものが含まれ,属性数は概念によって異なる.概念ベースには,このように定義された概念が約 9 万語収録されている.ある概念を A,その語の i 番目の属性を a_i,重みを w_i とすると,概念 A の属性とその重みは(1)式のように表すことができる.ただし,n は概念 A のもつ属性の総数とする.

$$A=\{(a_1,w_1),(a_2,w_2),\cdots,(a_n,w_n)\} \quad (1)$$

概念 A において概念自身の属性 $a_1,a_2,...,a_N$ を概念 A の一次属性と呼ぶ.一次属性を 1 つの概念と見なせば,一次属性からさらにその一次属性の意味を表す属性の集合が表現できる.a_i が持つ N 個の一次属性 $a_{i1},a_{i2},...,a_{in}$ つまり一次属性の属性を概念 A の二次属性と呼ぶ.さらに,二次属性を 1 つの概念と見なせば,二次属性からさらにその二次属性の意味を表す属性の集合が表現できる.このように,概念は N 次属性まで表現可能である.具体的に,概念「医者」の二次属性まで展開したものを表 1 に示す.

表 1　概念「医者」の一次属性と二次属性

概念	属性，重み			
医者	医師，0.334	患者，0.110	‥	治す，0.042
	医者，0.340	病人，0.270	‥	治療，0.106
	診察，0.102	看護婦，0.143	‥	医療，0.098
	病院，0.088	看病，0.055	‥	癒す，0.078
	:	:	‥	:
	保健，0.033	治療，0.049	‥	病気，0.022

4. 記事への重み付け

記事への重み付けには,形態素解析と *tf･idf* 重み付けを用いる.形態素解析とは日本語構造の制約(例えば,形容詞は名詞の前に付くことができるという法則)を利用し,単語の切り出しや品詞を同定することをいう.実際に日本語の複雑さのため完全に単語を切り出すことは難しいが,代表的な形態素解析器である茶筌[19]では様々な工夫により高い精度で単語を正しく切り出すことができる.本提案システムでは単語に切り出しに茶筅を用い,得られた「名詞」の内,概念ベースに存在する語のみを索引語として *tf･idf* 重み付けを用いることにより,記事への重み付けを行った**(変更①)**.

tf とは,索引語頻度を意味し,索引語の網羅性を示す値である.記事中に索引語がどれだけ多く出現するかを示している.何度も繰り返し使われる語は,重要であると考えられる.具体的には,記事 d における取得語 t の重み w を(2)式で定義する.s は総単語数,N は対象とする記事の総数とする.

$$w=\frac{tf(t,d)}{\sum_{s\in d}tf(s,d)}\times\log\left(\frac{N}{df(t)}\right) \quad (2)$$

例として,新聞記事「様々な細胞に変化できるＥＳ細胞は素晴らしいと賞賛されている.」に対する本提案手法における取得品詞と重み付け結果を表 2 に示す.

表 2　本提案システムにおける取得品詞と重み付け結果

索引語	品詞	*tf･idf* 値
細胞	名詞	0.283
賞賛	名詞	0.163
変化	名詞	0.123
様々	名詞	0.103

5. 英字略語の意味連想

英字略語は専門的な用語が多いため,国語辞典･百科辞典では全ての英字略語に対して意味候補を取得することができない.そこで,英字略語を Wikipedia で検索する.例として「DL」で検索した一部を図 2 に示す.

・デルタ航空の IATA **航空会社コード**
・ディーゼル機関車 - Diesel Locomotive
・差動固定装置(デフロック) - Differential Lock

図 2　DL での検索結果の一部

図 2 における下線の部分をそれぞれ意味候補として取得する.これは,図 2 における「-」の前の部分かつ,「()」と英字の部分を省いたものである.さらに本提案システムでは,削除対象ワード(図 3)を含むものを意味候補,図 2 の二行目を意味候補から削除した**(変更②)**

型番，型式，単位，記号，通貨コード，
航空会社コード，ドメイン，言語コード
コマンド，係数，登場，符号，作品，楽曲
アルバム，シリーズ，拡張子

図3　削除対象ワード

前記の通り，専門的な用語が多いことから全ての意味候補の属性を概念ベースでは取得することができない．そこで，図2の下線部分のそれぞれを*X*としてAFにより，概念化（式 1 の形）を行う．そのために，*X*の属性とその重みをWebから以下の手順で獲得する．まず，*X*をキーワードとして検索エンジンを用いて検索を行い，検索上位100件の検索結果ページの内容を取得する．次に，取得した文書群に対して，形態素解析ソフト茶筅を用いて形態素解析を行い，自立語を抽出する．最後に，概念ベースに存在する語のみを*X*の属性とし，頻度情報と特定性情報である*idf*値を掛け合わせたものを属性の重みとする．これらの行程の結果について，評価を行うと属性とその重みのセットは30個が最も関連性があるとされた．よって，AFでは属性を30個取得している．

例として，英字略語「DL」における意味候補の一つで図2における意味候補である「ディーゼル機関車」のAF結果を表3に示す．

表3　「ディーゼル機関車」のAF結果（上位5件）

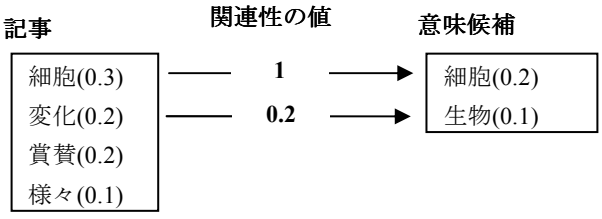
属性	重み
機関車	0.246
国鉄	0.027
鉄道	0.026
辞書	0.021
辞典	0.019

6. 英字略語の意味選択

英字略語の意味選択において，既提案手法では二つの語間にある関連性を数値として表す手法である関連度計算の中でも共通属性を考慮した関連度計算[20]を応用させた記事関連度計算を用いた．一方，本提案手法では，EMD を用いた記事関連度計算を用いた **（変更③）**．記事関連度計算で比べる二つの概念とは「記事への重み付け（4 章）」結果と「英字略語の意味連想（5 章）」の意味候補それぞれの AF 結果である．この二つの概念間の関連性を数値化し，最も関連性が高い意味候補を記事内の英字略語の意味として出力した．

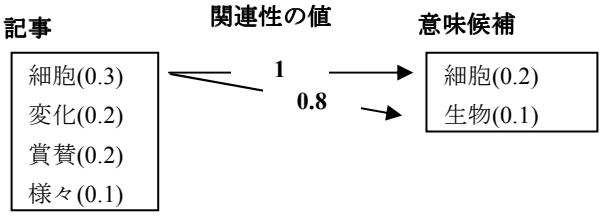
変更③に用いる EMD とは，線形計画問題の一つであるヒッチコック型輸送問題において計算される距離尺度であり，二つの離散分布において，一方の分布を他方の分布に変換する為の最小コストとして定義される．輸送問題とは，需要地の需要を満たすように供給地から需要地へ輸送を行う場合の最小輸送コストを解く問題である．EMD を記事関連度に用いる場合，記事（概念 *A*）と AF 結果（概念 *B*）の *A*,*B* 間の関連度計算に EMD を用いる．また，需要地と供給地，需要量と供給量，各需要地と供給地間の距離を定義する必要がある．そこで需要地には，概念 *A* の一次属性を，供給地には概念 *B* の一次属性を割り当て，需要量と供給量はそれぞれ一次属性の重みを用いた．その結果，需要地と供給地の距離は索引語間の関連性と見立てることができた．EMD の計算では関連性が高い語に優先して重みを輸送し，供給量がなくなるか需要量が満たされるまで輸送を行う．EMD では，最小コストを求めるため数値が低いほど関連性が高い．

記事関連度計算の計算例を図 4 に EMD を用いた記事関連度計算の計算例を図 5 に示す．



関連度＝一致度×重みの和×重み比率
1×(0.3＋0.2)×{(0.3＋0.2)÷2}
＋0.2×(0.2＋0.1)×{(0.2＋0.1)÷2}＝0.134

図 4　記事関連度計算



輸送コスト(距離×輸送量)
細胞→細胞　：(1-1)×0.2＝0
細胞→生物　：(1-0.8)×0.1 =0.02
EMD = (0＋0.02)÷0.3　＝0.067

図 5　EMD を用いた記事関連度計算

7. 評価

この章では，①品詞の取得制限，②意味候補削除，③EMD の変更を既提案システムに加えることにより，2012 年 11 月から 12 月末までの二ヶ月間における Yahoo!ニュースヘッドライン[21]記事から取得してきた多義を有する英字略語 200 件に対して，意味判断結果の精度向上が実現したかを検証する．正解判定として，人間(3 人)がその記事での英字略語の意味を調べ，出力と調べた結果が同じならば正解とした．今回使用した 200 件において，人間 3 人の間で意味選択にばらつきは存在しなかった．

図 6 より，新聞記事より取得する品詞を名詞に限定（①の変更）することで 3.5%意味判断精度が向上していることが分かる．この結果から，既提案システムでは「動詞」,「形容詞」が雑音として働いていたと考えられる．また，

Wikipedia から取得する意味候補の内, 型番やコードなどの意味候補を削除（②の変更）することでさらに 5%意味判断精度が向上した．しかし，EMD（変更点③）を加えることにより判断精度が 1%低下していることがわかる．この精度低下が起こる原因は属性差がない記事関連度計算の場合，共通属性を考慮した記事関連度計算に比べて EMD の判断精度が劣ることにあると考えた. そこで属性差ごとの，共通属性を考慮した記事関連度計算と EMD を用いた記事関連度計算の意味判断結果を図 7 に示す.

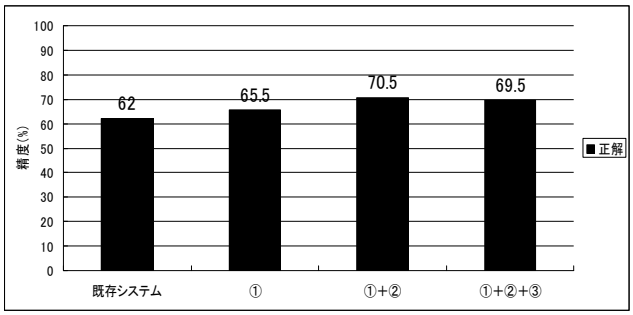


図 6　三点の変更による意味判断結果

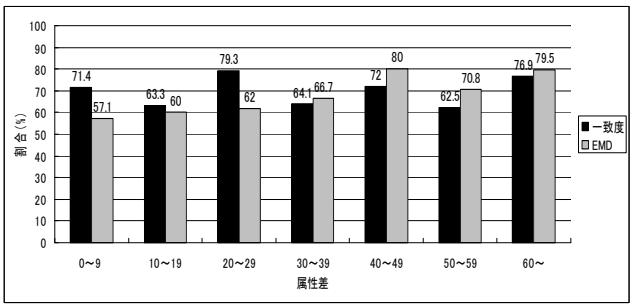


図 7　属性数ごとの意味判断結果

図 7 より，属性差が 30 までの場合，共通属性を考慮した記事関連度計算の方が意味判断の精度がよいことがわかる. 一方で 30 以上の属性差がある場合 EMD を用いた記事関連度計算の方が意味判断の精度が良い，このことより，属性差が 30 以内の場合には共通属性を考慮した記事関連度計算を使用し，30 以上の場合には EMD を用いた記事関連度計算を使用することでより良い結果が得られるといえる. 上記のように場合分けした判断結果を図 8 に示す.

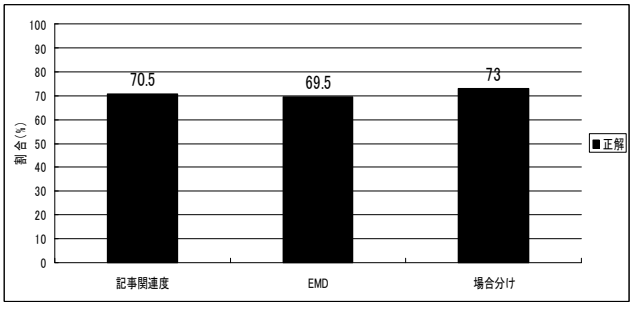


図 8　属性数ごとの意味判断結果

謝辞　本研究の一部は，科学研究費補助金(若手研究(B)21700241)の補助を受けて行った．謹んで感謝の意を表する．

参考文献

1) 乾健太郎, 藤田篤(2004).“言い換え技術に関する研究動向.” 自然言語処理,11 (5), pp.151-198.
2) Yamamoto,K.(2002).“Acquisition of Lexical Paraphrases from Texts.” Proceedings of Computerm2 Workshop of COLING2002, pp.22-28.
3) 藤田篤, 乾健太郎(2003).“語彙・構文的言い換えにおける変換誤りの分析.” 情報処理学会論文誌,Vol. 44 ,No. 11,pp.2826-2838.
4) 降幡建太郎, 藤田篤, 乾健太郎, 松本裕治, 竹内孔一 (2004). “語彙概念構造を用いた機能動詞結合の言い換え.”, 言語処理学会第 10 回年次大会発表論文集, PP.504-507.
5) Inui, K., Fujita, A., Takahashi, T., Iida, R., and Iwakura, T.(2003). “Text Simplification for Reading Assistance: A Project Note.” Proceedings of The Second International Workshop on Paraphrasing: Paraphrase Acquisition and Applications, Workshop of ACLOS, pp.9-16.
6) 松吉俊, 佐藤理史(2008).“文体と難易度を制御可能な日本語機能表現の言い換え.” 自然言語処理,Vol. 15,No. 2 ,pp.75-99.
7) 洞井智彦, 吉村枝里子, 土屋誠司, 渡部広一(2011). “語句変換による難解文から平易文への言い換え手法.” 情報処理学会研究報告. ICS-163(1), pp.1-6.
8) 石井直樹, 平石智宣, 延澤志保, 斎藤博昭, 中西正和 (2000). “日本語略語の自動復元. ” 情報処理学会研究報告,SIG-SLP, pp.23-30.
9) 野呂康洋, 桝井文人, 河合敦夫 (2003). “WEB 文書中の間接共起情報を利用したカタカナ語省略形の推定. ” 2003 年度電気関係学会東海支部連合大会講演論文集, p.257.
10) 酒井浩之, 増山繁, (2005). “略語とその原型語との対応関係のコーパスからの自動獲得手法の改良. ” 自然言語処理,Vol. 12,No. 5,pp207-231.
11) 岡崎直観, 石塚満(2007). “日本語記事からの略語抽出. ”人工知能学会全国大会論文集,2G4-4.
12) http://ja.wikipedia.org/wiki
13) 田邊僚, 吉村枝里子, 土屋誠司, 渡部広一(2011). “英字略語の意味判断システム. ” FIT2011,E-030,pp.271-272.
14) X.Wan, Y.Peng(2006).“The Earth Mover’s Distace as a Semantic Measure for Document Similality” Proceeding of the 14th ACM international conference on Information and knowledge management ,pp.301-302.
15) 辻泰希, 渡部広一, 河岡司(2003).“www を用いた概念ベースにない新概念およびその属性獲得手法” 第 18 回人工知能学会全国大会論文集,2D1-01.
16) 小島一秀, 渡部広一, 河岡司(2004).“連想システムのための概念ベース構築法－語間の論理的関係を用いた属性拡張.” 自然言語処理,Vol.11,No.3,pp.21-38.
17) 徳永健伸 (編),“情報検索と言語処理”, 東京大学出版会, 1999.
18) 荒木孝允, 渡部広一, 河岡司(2006).“共通・類似属性を考慮した概念間関連度計算方式” 情報処理学会第 68 全国大会講演論文集,4N-2.
19) 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原正幸(2002). “日本語形態素解析システム『茶筌』 version2.2.9 使用説明書”.
20) 渡部広一, 奥村紀之, 河岡司(2006). “概念の意味属性と共起情報を用いた関連度計算方式.” 自然言語処理,Vol.13,No.1,pp.53-74.
21) http://headlines.yahoo.co.jp