E-052

英字略語の意味判断システム
Meaning Judgment System for Alphabet Abbreviation

田邊　僚† 吉村 枝里子‡ 土屋 誠司‡ 渡部 広一‡
Ryo Tanabe Eriko Yoshimura Seiji Tsuchiya Hirokazu Watabe

1. はじめに

近年，国際化や情報化が進むことにより，英字表現や片仮名表現が多用されるようになった．それらの表現を含む文章は専門家には理解できるが，一般的には内容を理解しづらい．そこで，文章を難解にしている表現の一つである英字略語に正式名称を付与する処理を行うことにした．

英字略語とは，例えば IC や ATM のように英字文字列を省略することにより形成されている．英字文字列を省略している為，英字略語は多義性を保有している．例えば IC の場合「集積回路」という意味と「インターチェンジ」という意味を持つ．このように多義を持つ英字略語においても，文章の内容を概念ベース[1]を用いた重み付けにより，解釈し正式名称を付与するシステムを本稿で提案する．

2. 関連技術

2.1 概念ベース

概念ベースには様々な語（概念）が，それを特徴付ける語（属性）とその重要度を表す数値（重み）の対の集合によって定義されている．ある概念 A は m 個の属性 a_i と重み w_i（>0）の対によって(1)式のように定義される．

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

2.2 関連度計算[2]

概念ベースを用い，概念 A と概念 B の関係の深さを定量的に表すのが関連度計算方式である．関連度は，0 以上 1 以下の連続的な数で表され，概念同士の関連が大きいほど関連度は高くなる．

2.3 一致度計算[2]

一致度とは，概念の属性がどれだけ一致しているかにより関連の強さを定量的に評価する手法である．

概念 A と B の属性を a_i, b_j,重みを u_i, v_j,とし，属性それぞれ L 個，M 個($L \leqq M$)とすると式(2)，(3)と表現できる．

$$A=\{(a_i,u_i)\,|\,i=1\sim L\} \quad (2)$$

$$B=\{(b_j,v_j)\,|\,j=1\sim M\} \quad (3)$$

概念 A，B の一致度 $DoM(A,B)$は式(4)で示す．

$$DoM(A,B)=\sum_{a_i=b_j}\min(u_i,v_j) \quad \min(\alpha,\beta)=\begin{cases}\alpha(\beta\geq\alpha)\\ \beta(\alpha\leq\beta)\end{cases} \quad (4)$$

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

両方の属性に共通して存在する重み分は有効であるため，一致度は一致する属性のうち小さい方の重みの和となる．

3. 英字略語の意味判断システム

英字略語の意味判断システムの構成を図 1 に示す．

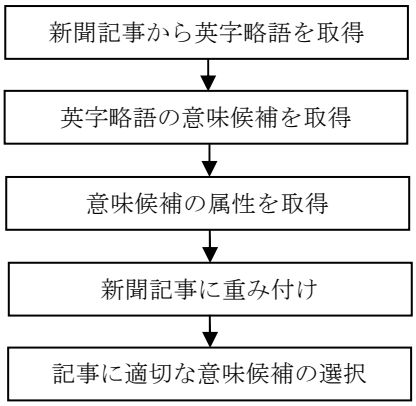

図 1　英字略語の意味判断システムの構成

3.1 英字略語の取得

新聞記事から英字略語を取得する．取得対象は，大文字･小文字一文字以上の英字文字列，英字文字列の後ろに数字がついたもの(CO2 など)とする．

3.2 意味候補の取得

英字略語は専門的な用語が多いため，国語辞典･百科事典では全ての範囲に対応することができない．そこで，英字略語を Wikipedia[3]で検索する．例として IC で検索した一部を図 2 に示す．

- 集積回路 (Integrated Circuit) - IC カード
- インタークーラー (Inter Cooler)
- インターチェンジ (Inter Change)
- イメージカラー (Image Color)
- イオンクロマトグラフィー（Ion Chromatography）
- インフォームド・コンセント (Informed Consent)
- インデックスカタログ - 星団や星雲，銀河を収載した 2 つの星表のこと(Index Catalogue)

図 2　IC 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．

3.3 属性の取得

図 2 の下線部分それぞれ X として，意味的特徴を表す単語（属性）とその重要性を表す重みの組とを Web を用

いて AF[5]で自動的に 30 個構成する．X の属性 x_i とその重み w_i の組を構成で構成されており，未定義語 X の属性 x_i とその重み w_i の組は(5)式のように構成される．

$$X = \{(x_1, w_1), (x_2, w_2) \cdots, (x_i, w_i)\} \qquad (5)$$

3.4 新聞記事に重み付け

形態素解析ソフト茶筌[4]を利用し，新聞記事内の「名詞」,「形容詞」,「動詞」などを取得する．そして，取得した語に対して $tf \cdot idf$ 重み付け[5]を行うことにより，重みを付ける．具体的には，記事 d における取得語 t の重み w_t^d を式(6)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．

$$w_t^d = \frac{tf(t, d)}{\sum_{s \in d} tf(s, d)} \times \log\left(\frac{N}{df(t)}\right) \qquad (6)$$

3.5 適切な意味候補の選択

例として，IC の意味候補の一つである「集積回路」を新聞記事と関連度計算したものを図 3 に示す．

図 3 における a_1 から a_{30} は，3.3 節の手法で取得した「集積回路」の AF 結果である．一方右側の b_1 から b_k は 3.4 節の手法で取得した，新聞記事内の「名詞」,「形容詞」,「動詞」に重み付けをしたものである．

左右の属性に共通属性を考慮した関連度計算を行う．

具体的には，初めに a_1 から a_{30} と b_1 から b_k との間の全ての組み合わせにおいて，属性一致度を算出する．そして，属性一致度の和が最大となるように組み合わせを決定する．図 3 では，b_1 から b_i を a_1 から a_{30} との属性一致度の和が最大の組み合わせとして選出している．

選出した a_1 から a_{30} と b_1 から b_i との間で関連度計算を行い「集積回路」と新聞記事との関連性を定量化する．この手法をそれぞれの意味候補で行い，最も関連度計算結果が高いものを記事内における英字略語の適切な意味として，英字略語の後ろに付与する．

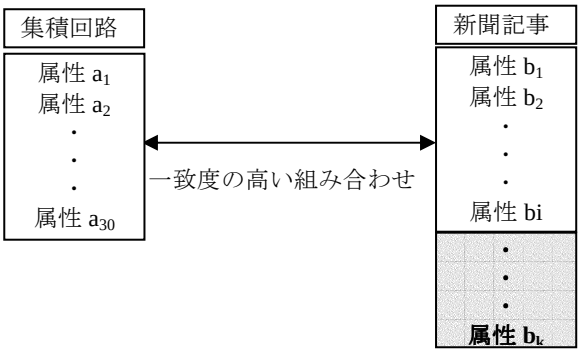

図 3 「集積回路」と新聞記事での関連度計算

4. 評価

3 節での提案手法の評価を行った．テストデータとして，新聞記事よりランダムに取得し，その内の英字文字列を含んだ 100 件の記事を対象とする．正解判定として，人間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する．

100 件の記事に対し，3.1 節で述べた英字略語の取得法を用いると英字文字列の数は 144 個となる．この 144 個に対しての正解判定結果を表 1 に示す．判定不可能とは，Wikipedia にその用語自体ない場合を示す．また，割合における小数点以下は四捨五入をしている．

表 1　判定結果(括弧内は割合(%))

文字列の種類	正解	不正解	判定不可能
一文字	0(0)	7(100)	0(0)
一文字+数字	9(64)	2(14)	3(22)
二文字	20(62)	12(38)	0(0)
二文字+数字	3(75)	0(0)	1(25)
三文字	48(82)	8(13)	2(5)
三文字+数字	6(75)	0(0)	2(15)
四文字	14(93)	1(7)	0(0)
四文字+数字	0(0)	0(0)	1(100)
五文字	2(100)	0(0)	0(0)
五文字+数字	0(0)	0(0)	0(0)
六文字	1(33)	0(0)	2(67)
全体	**103(72)**	**30(21)**	**11(8)**

5. 考察

表 1 から三文字以上の長さであると正解率は高く，逆に一文字のものは，7 個中 7 個が不正解となっており，正解率が低いことがわかる．これは，文字数が少ない英字略語ほど意味候補が多いことが原因となっていると考えられる．意味候補が多い場合，正解の意味と近い意味が同時に含まれる可能性が高くなり，判断間違いが起こる．

また 11 個存在する判定不可能の原因は新出英字略語である．新出の英字略語は Web ページが存在しないので，意味候補を取得することができない．その為，新たな手法の提案が必要である．例えば Web からその英字略語が使われている文章を見つけ，そのときに使われていた英字略語の意味を取得するなどの方法が考えられる．

6. 終わりに

本研究では，新聞記事内の重みと語，意味候補の属性とを共通属性を考慮した関連度計算を行うことにより，英字略語の適切な意味を判断する手法を提案した．この手法により，約 72%の正解率を実現することができた．

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．

参考文献

[1]奥村紀之，土屋誠司，渡部広一，河岡司，“概念間の関連度計算のための大規模概念ベースの構築”， 同志社大学理工学研究報告，Vol.14, No.5，pp.41-64，2007.

[2]荒木孝允, 奥村紀之, 渡部広一, 河岡　司，“比較対象概念の共通属性を重視する動的関連度計算方式”，同志社大学理工学研究報告，Vol.48, No.3，pp.14-24，2007.

[3]Wikipedia，http://ja.wikipedia.org

[4]ChaSen - 形態素解析器，奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座(松本研究室)，http://chasen-legacy.sourceforge.jp/，（2011/ 1/ 27）.

[5]Salton, G. and Buckley, C.: Term-weighting approaches in automatic text retrieval , Information Processing & Management Vol.41 No.4 pp.513-523 , 1988.