F-034

# Wikipedia 記事の文章群を用いた多義を有する英字略語の意味判断システム
# The judgement system using sentence group of Wikipedia article for abbreviated alphabetical characters with ambiguity

後藤 大介† 後藤 和人† 土屋 誠司‡ 渡部 広一‡
Daisuke Goto Kazuto Goto Seiji Tsuchiya Hirokazu Watabe

## 1. はじめに

国際化・情報化が急速に進む中，英字略語を目にする機会は増加し続けている．英字略語とは，例えば IC(Integrated Circuit)のように，1 つ 1 つの英字文字(I,C)にそれぞれ文字列を持つ英字文字列と定義する．英字略語は多義を持つものがあるため，英字略語を含む情報はすべての人が正確に判断できるとは言い難い．そこで，英字略語が含まれる文章内の語を考慮することで，英字略語の意味を判断する手法の実現を目指す．

英字略語の意味を判断する既存手法として多義を有する英字略語の意味判断システム[1]がある．しかし，既存手法では EMD(Earth Mover's Distance) [2]を用いた記事関連度計算方式[3]を行う際に，AF（オートフィードバック）[4]を用いるため，英字略語の意味を判断する際に雑音が含まれ正しく判断できないことがある．そこで，本稿ではこの問題を Wikipedia[5]の記事冒頭文と「概要」欄の文章群を使用することで，既存手法の改善を目指す．

## 2. 関連技術

### 2.1 概念ベース

概念ベース[6]とは，電子化された国語辞書などから機械的に構築された大規模なデータベースである．ある語を概念として定義し，概念の意味特徴を表す語である属性と，その属性の重要さを表す重みの対の集合によって構成されている．ある概念 $A$ は $m$ 個の属性 $a_i$ と重み $w_i(>0)$の対により，式(1)のように表現される．ここで，属性 $a_i$ を概念 $A$ の一次属性と呼ぶ．

$$\text{概念 } A=\{(a_1,\ w_1),\ (a_2,\ w_2),\ \cdots,\ (a_m,\ w_m)\} \quad (1)$$

### 2.2 EMD を用いた記事関連度計算方式

記事関連度計算方式とは，記事間の関連の強さを定量的に表現する手法である．記事関連度は 0.0 から 1.0 の実数で表現される．記事関連度は記事間において同一，または意味の近い単語を共通して多く含むほど高い値を示す．

EMD とは，分布間の距離を表すもので，最適な輸送コストを用いて定義される．EMD を記事検索に適用する場合，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．需要地には検索要求の索引語を，供給地には検索記事の索引語を割り当てる．索引語とは記事中の名詞を指す．

### 2.3 オートフィードバック（AF）

AF とは，概念ベースに定義されていない未定義語の属性とその重要度を表す重みの組を，Web を用いて獲得する手法である．

### 2.4 *tf*・*idf* 重み付け

*tf*・*idf* 重み付け[7]とは，索引語の頻度と特定性に基づいた重み付け手法である．*tf* はある文章 $d$ に出現する索引語 $t$ の頻度を表す尺度であり，式(2)で定義される．ただし，文書 $d$ における単語の総数を $W$，索引語 $t$ の出現回数を $n$ とする．*idf* は対象とする全文章において索引語が出現する文章数を表す尺度（特定性）であり，式(3)で定義される．なお，$N$ は検索対象となる文章集合中の全文章数，$df(t)$は語 $t$ が出現する文章数である．

$$tf(t,d)=\frac{n}{W} \quad (2)$$

$$idf(t)=\log\frac{N}{df(t)}+1 \quad (3)$$

## 3. 英字略語の意味判断システム

入力記事中に英字略語がある場合，入力記事と英字略語の意味候補に対して EMD を用いた記事関連度計算を行い，その記事内における英字略語の意味として出力する．提案手法の概略を図 1 に示す．

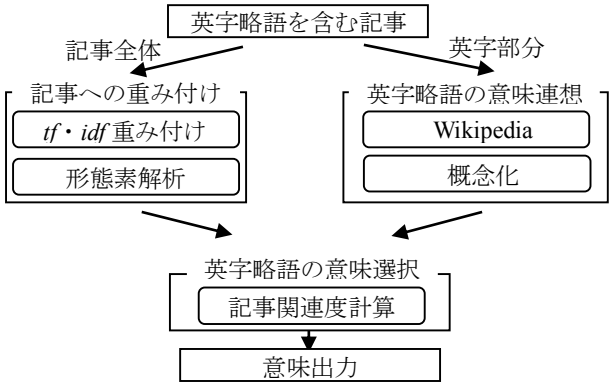


図 1 システムの概要

### 3.1 英字略語の取得

英字略語を含む記事から文字コードを用いて英字を取得する．記事中に連続して並ぶ英字は一つの英字略語として取得する．このとき，一文字の英字は「T シャツ」や「V 字回復」のように，その文字の形を意味としている場合がある．これらは本稿における英字略語の定義から外れるため取得しない．しかし，T（テスラ）や V（ボルト）のように一文字の英字でも，その前に数字がつく場合は単位となる場合が多く，これらは本稿の英字略語の定義に当てはまる．そこで数字の後につく一文字の英文字は英字略語として扱い，取得する．

### 3.2 記事への重み付け

入力記事と英字略語の意味候補に対して関連性を計算す

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

るためには，入力記事を概念ベースに対応した形で表現することが必要である．そこで，代表的な形態素解析器である茶筌[8]と *tf・idf* 重み付けを用いて記事への重み付けを行う．記事内の名詞かつ，概念ベースに存在する語を，索引語とする．この索引語に対し，*tf・idf* 重み付けを用いて，記事への重み付けを行う．その際，名詞のみに限定して取得する理由は，動詞や形容詞が英字略語の特徴を表す単語と考えられず，雑音となっている可能性があるためである．

### 3.3 英字略語の意味連想

既存手法では，英字略語の意味候補それぞれに AF を行い，その結果を概念として取り扱っている．ただし AF は Web から属性を獲得するため，雑音が含まれることがある．そこで，提案手法では英字略語の意味候補を概念とし，属性の取得に AF を用いる代わりに，Wikipedia の記事冒頭文と「概要」欄の文章群を英字略語の意味候補の属性として利用する．

提案手法の手順として，まず Wikipedia 上の英字略語の意味候補のページから記事冒頭文と「概要」欄の文章群を抽出する．この際，英字略語の意味候補のページが Wikipedia 上に存在しない場合は，その場合は英字略語の意味候補から除外する．Wikipedia の記事冒頭文および「概要」欄の文章群には，記事の概略が書かれ，その記事において重要なキーワードが多く含まれている．そのため，多義の推定を行う際に雑音となる可能性の高い語彙は獲得されないと期待できる．Wikipedia から抽出した文章群への重み付けは，3.2 節と同様に，茶筌と *tf･idf* 重み付けを用いる．文章群内の名詞かつ，概念ベースに存在する語を索引語として取得し，*tf･idf* 重み付けを用いて文章群への重み付けを行う．この際，*idf* は Wikipedia の全ページ（2015 年 10 月時点）から算出した値とした．

### 3.4 英字略語の意味選択

EMD を用いた記事関連度計算を行い，英字略語の意味選択を行う．具体的には，二つの概念，入力記事（概念 *A*）と英字略語の意味候補（概念 *B*）に対して，EMD を用いた記事関連度計算を行った．需要地には，概念 *A* の属性を，供給地には概念 *B* の属性を割り当て，需要量と供給量はそれぞれ一次属性の重みを用いた．

入力記事と英字略語の意味候補における EMD を用いた記事関連度計算の結果，最も関連度が高い英字略語の意味候補を，入力記事内における英字略語の意味として出力する．

## 4. 評価

Yahoo!ニュース[9]から取得した多義を有する英字略語を含む 250 件の記事に対して，提案手法の評価を実施した．評価方法として，提案手法が出力した結果と，被験者 3 名が各記事に含まれる英字略語の意味を調査し回答した結果が同じであった場合に正解とした．被験者 3 名の間で回答結果にばらつきは存在しなかった．

提案手法の正答率は 74.0%であり，既存手法の正答率 67.6%と比較し 6.4%精度が向上した．これより，属性の取得方法として，雑音が含まれる可能性がある，AF より，Wikipedia の記事冒頭文と「概要」欄の文章群を利用することが有効に機能したと考えられる．

## 5. 考察

250 件の記事に含まれる英字略語の文字数の内訳，既存手法と提案手法それぞれにおいて各英字略語の文字数ごとの正解率を表 1 に示す．

**表 1　英字略語の文字数ごとの正解率**

| 文字数 | 1 文字 | 2 文字 | 3 文字 |
|---|---|---|---|
| 割合(個数) | 8.0<br>(20) | 56.4<br>(141) | 35.6<br>(89) |
| 提案手法 | 60.0<br>(12) | 70.2<br>(99) | 83.1<br>(74) |
| 既存手法 | 25.0<br>(5) | 66.0<br>(93) | 79.8<br>(71) |

表 1 より，文字数に関わらず，提案手法が既存手法より高い正解率を示した．これは，AF により取得した属性に雑音が含まれていたことに加え，Wikipedia から取得した文章群から取得した属性に英字略語の意味候補の内容を表すために有効なキーワードを含んでいたといえる．中でも 1 文字の英字略語の場合，正解率は既存手法と比較し 35.0%上昇した．これは，提案手法では 1 文字の英字略語の意味候補を単位に限定したためと考える．1 文字の英字略語の意味候補は 2 文字，3 文字の英字略語と比較して多く，類似した意味候補もより多く存在する．そのため，英字略語の意味候補を Wikipedia から取得する際，取得する意味候補を吟味することが必要である．

## 6. おわりに

本稿では，英字略語の意味候補の属性に AF を用いる代わりに Wikipedia の記事冒頭文と「概要」欄の文章群を使用することで精度向上を図った．250 件の記事を入力として評価を行った結果，提案手法は既存手法と比較し，6.4%の正解率向上を実現した．

**謝辞**

本研究の一部は，JSPS 科研費 16K00311 の助成を受けて行った．

**参考文献**

[1] 田邊僚，吉村枝里子，土屋誠司，渡部広一，“EMD を用いた英字略語の意味判断システム”，情報処理学会研究報告知能システム，2013-ICS-170，pp.1-5，2013.

[2] X.Wan, Y.Peng, “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.301-302，2006.

[3] 藤江悠五，渡部広一，河岡司，“概念ベースと Earth Mover’s Distance を用いた文書検索”，自然言語処理，Vol.16，No.3，pp.25-49，2009.

[4] 辻泰希，渡部広一，河岡司，“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”，人工知能学会全国大会論文集，2D1-01，2003.

[5] “Wikipedia”，http://ja.wikipedia.org/wiki (2016/6/25)

[6] 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”，自然言語処理，Vol.14，No.5，pp.41-64，2007.

[7] 徳永健伸（編），“情報検索と言語処理”，東京大学出版会，1999.

[8] 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）．

[9] “Yahoo!ニュース”，http:// headlines.yahoo.co.jp (2016/6/25)