詳細化した同義関係をもつ同義語辞書の作成

高岡一馬　岡部裕子　川原典子　坂本美保　内田佳孝

株式会社ワークスアプリケーションズ

{takaoka_k,okabe_y,kawahara_n,sakamoto_mi,uchida_yo}@worksap.co.jp

1　はじめに

自然言語処理技術の普及にともない企業内で言語処理が利用される場面がふえている。とりわけ全文検索システムはさまざまなシステムの基幹として使用されているが、検索の網羅性を担保するには同義語辞書が不可欠である。しかし商用システムで利用できる同義語辞書はかぎられているとともに、検索対象となる企業内文書で使用される語彙が不足している問題がある。そこでわれわれは形態素解析器 Sudachi[1] の語彙辞書（以下 Sudachi 辞書）に同義関係を付与することで商用利用可能な同義語辞書を作成、公開した。[*1]

また用途により展開を制御するため同義関係を細分化したが、従来のような単層の分類ではじゅうぶんな記述ができなかったことから、同義関係に階層性を導入し詳細な記述をこころみた。

2　同義語の収集

同義語収集の手法としては従来から分布仮説にもとづいて統計的に収集する手法 [5][8] や、国語辞典や Wikipedia などの構造化された文書から抽出する手法 [6][4] が提案されている。しかし前者は同義語以外の関連語や対義語がふくまれるという欠点がある。他の情報を併用して選別する手法 [3] も提案されているが、品質をたもつためには最終的に人手による選別が必要である。後者については、本辞書作成時に試験的に利用したが、カテゴリに偏りがでる、同義語関係の一部しかえられないという理由で人手による補完が必要になった。

このような偏りを解消するため、本辞書作成にあたっては Sudachi 辞書に登録された見出し語に対し人手で同義関係を付与した。Sudachi 辞書に不足する語があれば追加をおこなった。したがって本辞書にふくまれている語はすべて Sudachi 辞書に登録されている。収載した語数を表 1 にしめす。

表 1　本辞書の収録語数

見出し語数	46,000
同義語グループ数	約 17,000
同義語対の数	約 56,000

3　同義関係の詳細化と階層性

黒田ら [7] は同義関係を異表記対、略記対、同義異語句対、誤表記対、準誤表記対、誤用対の 6 種に分類したが、実際には略称の異表記のような掛け合わせも存在しうるため、1 種の関係だけでは厳密に記述することができない。そこでわれわれは同義関係に階層を導入し複合した関係を記述することをこころみた。

本辞書では表 2 の 9 つの同義関係を定義した。その上で表 3 のように語彙素、語形種別、略語・略称、表記ゆれの 4 つの階層を定義し、同義関係を分類した。

また語形種別、略語・略称、表記ゆれの各階層で代表的な形に対して、それぞれ代表語、代表語形、代表表記を付与した。同義関係のラベルはそれぞれの階層で代表語、代表語形、代表表記を基準に付与されている。代表語形、代表表記以外の語同士の関係は代表語、代表語形、代表表記を仲介して記述される（図 1、図 2）。

辞書内では同義関係は自明な同義語彙素を除き、階層別に 3 つ組みで記述される。前述の「略語の異表記」の同義関係は（代表語, 略称・略語, 異表記）となる。

4　代表性の付与

本辞書では、検索エンジンでのサジェストや文生成などの実用上の利便性のため、語彙素内での代表的な

[*1] https://github.com/WorksApplications/SudachiDict

表 2　同義関係の定義

関係	定義	例
同義語彙素	語彙素同士の同義関係	支払い/勘定
旧称	現在の名称に対する古い名称	三菱東京 UFJ 銀行/三菱 UFJ 銀行
対訳	日本語の名称に対する他言語での名称	日本/Japan
別称	正式名称に対する別の名称（旧称、対訳以外）	社会保障・税番号制度/マイナンバー制度
誤用	誤って同義とされたもの	おもむろ/突然
略語・略称	元の名称を部分的に略したもの	流行性感冒/流感
翻字	外来語と元綴りの関係	インフルエンザ/influenza
異表記	同じ語形の異なった表記	子供/子ども
誤表記	誤って表記されたもの	シミュレーション/シュミレーション

表 3　同義関係の階層

階層	関係
語彙素	同義語彙素
語形種別	旧称 対訳 別称 誤用
略語・略称	略語・略称
表記ゆれ	翻字 異表記 誤表記

形として代表語、代表語形、代表表記を与えた。ユーザがより選好する形を代表とする必要があるが、ユーザにとって自然な表記はかならずしも高頻度に出現するとはかぎらない。用字や送り仮名についても一貫性がない。したがって代表の決定に頻度の高い出現形を採用するといった操作的な手法をとることはできない。代表性を厳密にあるいは操作的に定義することは困難である。

本辞書の作成にあたっては、代表語、代表語形、代表表記を付与する作業を言語リソース作成経験の豊富な作業者 3 名の主観によりおこない、作業者間の相互チェックをおこなっている。

5　同義関係の非対称性

本辞書では付与する同義関係を文脈から独立して同義性を判断できるものにかぎった。そのため「携帯電話/携帯/ケータイ」のような同義語グループのうち「携帯」のように多義を有する語については同義関係を有向とし、「携帯電話→携帯」「ケータイ→携帯」など多義語への同義関係のみを記述、多義語を起点とする同義関係を抑制した。

6　補助的情報

上にあげたもののほか、補助的な情報として分野情報および見出し語が体言か用言かを付与した。付与した分野の一覧を表 4 にしめす。ただし分野情報は管理上の参考情報として排他的に付与されており厳密ではない。

表 4　分野情報

人 人名 キャラ 国名 地名 地形 観光 店 店名 商品
大学名 企業名 組織名 ビジネス 政治 法律 IT 建築
交通 医療 美容 料理 ファッション 娯楽 音楽
スポーツ 教育 化学 動植物 時間 単位 色 不快

7　今後の課題

本辞書は Sudachi 辞書の収録語に対して上で述べたような同義関係を付与しているが、Sudachi 辞書内にも表記正規化として異表記、翻字、誤表記の関係が含まれており、同義語辞書との役割分担を再整理する必要がある。

また本辞書は企業内での全文検索を主目的として作成されたためビジネス文書に出現する語がおおく収録されているが、それ以外の分野はまだじゅうぶん網羅しているとはいえない。とくに SNS などで利用される打ち言葉への対応は課題である。

Sudachi 辞書は形態素解析辞書 UniDic の収録語彙を内包しており、UniDic の見出しと対応づけられた分類語彙表増補改訂版データベース [2] との関係も整理していきたい。

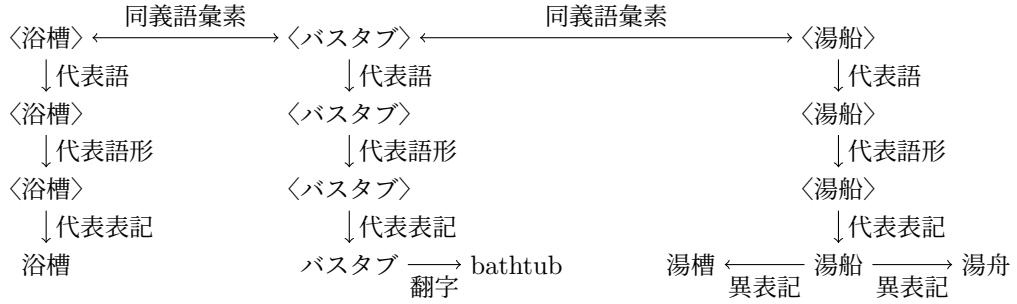

図 1　表記ゆれの例

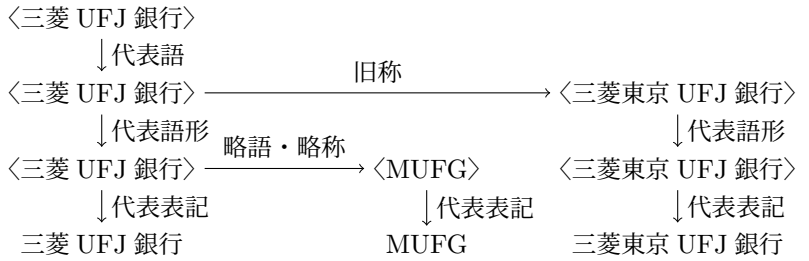

図 2　語形種別、略語・略称の例

なお本辞書は Sudachi 辞書パッケージの一部として Apache License 2.0 で公開している。

参考文献

[1] Kazuma Takaoka, Sorami Hisamoto, Noriko Kawahara, Miho Sakamoto, Yoshitaka Uchida, and Yuji Matsumoto. Sudachi: a japanese tokenizer for business. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, may 2018.

[2] 近藤明日子, 田中牧郎. 分類語彙表・unidic 見出し対応表の構築—コーパスへの網羅的・系統的な語義情報付与を目指して—. 言語処理学会第 23 回年次大会発表論文集, pp. 90–93, 2017.

[3] 澤田晋之介, 柴田知秀, 黒橋禎夫. 項共有述語項の意味関係コーパスの整備および同義・反義性判定. 言語処理学会第 24 回年次大会発表論文集, pp. 726–729, 2018.

[4] 中山浩太郎, 原隆浩, 西尾章治郎. Wikipedia マイニングによるシソーラス辞書の構築手法. 情報処理学会論文誌, Vol. 47, No. 10, pp. 2917–2928, oct 2006.

[5] 風間淳一, Stijn De Saeger, 鳥澤健太郎, 村田真樹. 係り受けの確率的クラスタリングを用いた大規模類義語リストの作成. 言語処理学会第 15 回年次大会発表論文集, pp. 84–87, 2009.

[6] 村田真樹, 金丸敏幸, 井佐原均. 複数の辞書の定義文の照合に基づく同義表現の自動獲得. 自然言語処理, Vol. 11, No. 5, pp. 135–149, oct 2004.

[7] 黒田航, 風間淳一, 村田真樹, 鳥澤健太郎. Web 文書にも対応できる日本語異表記の認定基準. 言語処理学会第 16 回年次大会発表論文集, pp. 990–993, 2010.

[8] 城光英彰, 松田源立, 山口和紀. 文脈限定 skip-gram による同義語獲得. 自然言語処理, Vol. 24, No. 2, pp. 187–204, 2017.