日本語ワードネットにおける異表記拡張の効果

栗林孝行
豊橋技術科学大学
kuribayashi@lang.cs.tut.ac.jp

Francis Bond
情報通信研究機構/南洋理工大学
bond@ieee.org

黒田航
京都工芸繊維大学/早稲田大学
kow.kuroda@gmail.com

神崎享子
国立国語研究所
kanzaki@ninjal.ac.jp

内元清貴
情報通信研究機構
uchimoto@nict.go.jp

井佐原均
豊橋技術科学大学
isahara@tut.jp

jwordnet@gmail.com

1. はじめに

我々は日本語ワードネット(http://nlpwww.nict.go.jp/wn-ja/において公開中)の開発を続けてきたが、今年度は(語義の追加と)異表記への対応という拡張を行ったJWNを、2010年度に行ったコーパスアノテーション(パラレルコーパスへの語義タグ付与)と同一のコーパスに適用し、両者の結果を比較し、異表記対応の重要性について検証した。また、その他今年度実施した作業と今後の方針についても述べる。

2.日本語ワードネット

日本語ワードネット(以下JWN)は(独)情報通信研究機構よりリリースされている、フリーの日本語概念辞書であり、Princeton WordNet 3.0(以下英語ワードネット, http://wordnet.princeton.edu/) を基に構築されている。また日本語以外でも、フランス語、中国語等多数の言語のワードネットが同ワードネットをもとに構築されており、自然言語処理に利用されている。それら中では概念は「synset」という単位で扱われ、synsetは「synonym」という1つの概念を表すために使用される語(ときに複数の語)などから成り立っており、synsetどうしが上位概念・下位概念等の関連性をもって密にリンクしている。

また、本論文で「語義」といった場合には、語とsynsetのペアのことを指す。

3. コーパスアノテーションと異表記

我々は2010年度、JWNに不足している語義や間違った語義の発見、語義頻度情報の獲得、またより多くの例文獲得等のために人手によるコーパスアノテーションを行った。

対象としたのは京大コーパスの記事部分の冒頭1,000文、フリーのものとしてコナン・ドイル著シャーロック・ホームズシリーズより「踊る人形(The Adventure of the Dancing Men)」、「まだらの紐(The Adventure of the Speckled Band) 」、エリック・レイモンド著「伽藍とバザール(The Cathedral and the Bazaar)」である。尚、使用したJWNはバージョン1.1をバグフィックスしたもの(以下1.1改)であり、形態素への分割と語のレンマ化はChaSen[1]を用いて行った。表1にその結果を示す。

また、共同研究先である南洋理工大では同コーパスの英語・中国語版に対してそれぞれ英語ワードネット・中国語ワードネットを用いてコーパスアノテーションが行われている。[2]

	文数	語義	s	その他	対象語数
京大コーパス	1000	7669	1193	1937	10799
踊る人形	698	3746	585	915	5246
まだらの紐	702	3958	369	1299	5556
伽藍とバザール	773	6538	809	1881	9228

表1 コーパスアノテーションの結果

「文数」は 対象とした文の数
「語義」は対象となった語(「対象語」)のうち、何らかの語義が選択された数
「s」はJWNに該当する語義なしと判定された語の数
「その他」は固有名詞・機能語などアノテーション対象外の語、複合語でなければ意味をなさない語、その他エラーと判定された数

この結果を検証したところ、「s」のタグが付いたものや該当する表記がJWNになくアノテーション対象にならなかった語の中に、該当する概念がJWNに存在しないのではなく、「防空ごう」「あやうい」のように表記の違いが原因であるものが相当数含まれていた[3]。そこで、異表記対応をして拡張したJWNをこれらのコーパスに適用し、その有効性を検証することにした。

4. 異表記対応

まず、異表記対応をするために、語にカタカナ・ひらがなで読み情報を付与した。これは、読みが語の仮名表記と同一であるためと、同一synsetのsynonymかつ読みも同一であればそれらは異表記である可能性が高いためであり、またJWNを利用されている日本語学習者の方々からの要望でもある。

具体的には、

1. Jmdictの読みと意味が同じ表記、JUMANの標準表記が同じ表記を異表記のセットとしてまとめ、読み仮名を付与する
2. 1.で読みの付かなかった表記は、IPAdicを辞書として用いたMecab[4]で処理して読み仮名を付与する
3. 上記のいずれによっても読みが付かなかった場合、「読みなし」を意味するタグを付け、人手で読み仮名を

1　http://chasen.naist.jp/hiki/ChaSen/

2　当該共同研究は豊橋技術大学と南洋理工大との間で結ばれ、日本学術振興会より二国間交流事業として支援を受けている

3　「防空壕」という表記であればsynsetID=02868638-nのsynonymとしてhitする

4　http://mecab.sourceforge.net/

付与する

4. 1~3 を JWN の synonym リストに適用し、同一 synset の synonym で、読み仮名が同一である表記を異表記対としてまとめる

という手順で対応を行った。

その結果、まず JWN 自体のステータスは以下のように変わった。

1.1 改	異表記対応拡張後
57178synset	57178synset
158074 語義	155317 語義
91961 語	214393 語

まず表記に異なりがあってもそれらは 1 つの語義としてまとめたため、語義の数は減少した。これは、JWN ではこれまで例えば「吸い込む」も「吸込む」もそれぞれ一語として扱っていたため語義の数が実質よりも大きくなってしまっていたという問題点が改善されたことを意味する。

しかし、語の数は 214393 まで増加しており、内訳としてはカタカナ・ひらがな表記の増加分が大きいが、それ以外にも JMdict と JUMAN から 20259 語を導入できた。

さらに、synsetID=01539063-v(Eng: draw, suck,...) では「吸い込む」「吸込む」の両者が synonym となっているのに synsetID=02765464-v(Eng: absorb, take in)では「吸い込む」の方しか synonym になっていないという類のエラーをある程度解消できた。

表 2 に拡張後の JWN を前出のコーパス約 3000 文に適用して、適用前と比較したものを示す。ただし、適用後のデータは人手によるチェックを行っていないため、適用前のデータも人手によるアノテーションを行う前の、機械によるアノテーションによるものであり、また、名詞のカウント方法に違いがあるため、表 1 と対象語数が若干異なる。

	総語数	対象語数	対象異なり語数	タグ済対象語数 (%)	対象異なり語数 (%)
京大コーパス (記事)	24615	11939	4218	9385 (78.61)	2863 (67.88)
				9766 (81.80)	3063 (72.62)
踊る人形	13483	4752	1855	3874 (81.52)	1421 (76.60)
				4332 (91.16)	1605 (86.52)
まだらの紐	13896	4848	1815	4097 (84.51)	1454 (80.11)
				4501 (92.84)	1610 (88.71)
伽藍とバザール	18067	7509	2210	5858 (78.01)	1553 (70.27)
				6618 (88.13)	1777 (80.41)
京大コーパス (社説)	27906	13300	2454	10958 (82.39)	1820 (74.16)
				11542 (86.78)	1952 (79.54)

表 2 異表記拡張前(上段)・拡張後(下段)の比較

表 2 で示したとおり、異表記拡張によってカバー率が数~10 ポイント改善しているが、副作用としてやはり語義曖昧性も増えてしまった。

その上今回の手法を取った場合、性質上どうしても読みが付いていない、正しく読み仮名が付いていないという場合がどうしても出てきてしまう。

現状では読みは手作業で修正する他なく、本原稿執筆時点では前者については修正済みであるが、後者についてはまだ修正中の段階であり、それが完了した段階で再度同一コーパスに適用してみる必要がある。

また、読みの情報は語義曖昧性解消にも一役買うと思われ、現在では読み情報の付いたコーパスはあまり多くないが、いずれ検証を行いたい。

以下に、例として「空」の読みと語義の関係を示す。

	クウ	ソラ	カラ	ウロ
01086545-a	○		○	
01497736-a	○			
08521267-n		○		
09238926-n				○
09239302-n				○
09294877-n				○
09304465-n				○
09304750-n	○			○
09436708-n	○	○		
13910116-n	○			
14455206-n	○		○	
14455552-n	○		○	

表 3「空」の読みと語義

左列は synsetID を示し、対応する読みを「○」で示した
これを見ると、形容詞(synsetID の末尾が「a」で終わっているもの)の場合は「ソラ」「ウロ」とは読まないことが分かる。

5. 表示表記

コーパスアノテーションや語句抽出などに JWN を利用する場合は異表記全てを網羅している必要があるが、文章生成に利用する際や JWN を検索した結果を表示する際などにその全てが使われると煩雑になってしまう。そこで我々は「表示表記」というものを定め、煩雑さを解消することを試みた。その基準は以下の

ようになっているが、上から順に優先度は高く設定している。

1. コーパスの出現頻度が高い表記
2. JUMAN[5]の代表表記と一致する表記
3. ひらがな表記より漢字表記を優先
4. 旧字体より新字体を含む表記を優先
5a. より字数の少ない表記を優先(カタカナ表記のもの以外)
5b. より字数の多い表記を優先(カタカナ表記のもののみ)

現状ではまだ有用な頻度情報が得られるほどサンプル数が充分ではないため、1. の頻度情報は今のところ使用していない。しかし特に 3. 以降の基準を適用すると、人間の目で見て表示表記が不自然になってしまう場合もあるため、頻度情報も重要であると考えられる。

6. まとめと今後

我々は JWN の異表記対応拡張を行い、コーパスに適用した際のカバー率の数~10 ポイントの向上をみた。

またコーパスアノテーションによってワードネットに存在しない概念が判明したため、出現頻度の高かったものを中心に追加を行っている。「名人」(将棋のタイトルの)等日本語独自のものは我々が単独で JWN に追加し、英語・中国語にも存在する「昨年」(Eng: last year; Mcn: 昨年)等は南洋理工大学と情報を共有しながら追加を進めており、それらも異表記対応をした上で JWN に適用する予定である。

その上で、昨年度作成した形容詞類の語尾データを用いた拡張も行う予定である(こちらも異表記対応を要する)。

上記の事情により、バージョン 1.2 からはリリースの際のフォーマットも変更となる。

また、現在、ワードネットと日本語語彙大系[6]の対応情報の修正と、新たなパラレルコーパス約 3,000 文に対するコーパスアノテーションを進めている。

アノテーション対象はやはりパラレルコーパスで、内訳は表 2 の比較でも用いた京大コーパスの社説部分からの 1,000 文とシンガポール政府観光局による観光案内コーパス[7]のうち約 2,000 文である。

それが終わった段階で我々は約 6,000 文に対してアノテーションを実施したことになるが、JWN 検索時に出現頻度の高いものから順に表示したり、出現頻度のあまりにも低い語義を除外したり、表示表記をより自然なものにするために語義の頻度や読みの頻度を利用するが、それを実現するにはある程度の量のサンプルが必要であるため、今後もコーパスアノテーションを続けて行きたい。

特に後者について、JWN 検索時に出現頻度の高いものから順に表示したり、出現頻度のあまりにも低い語義を除外したり、表示表記をより自然なものにするために語義の頻度や読みの頻度を利用するが、それを実現するにはある程度の量のサンプルが必要であるため、今後もコーパスアノテーションを続けて行きたい。

参考文献

5 http://nlp.ist.i.kyoto-u.ac.jp/index.php?cmd=read&page=JUMAN&alias[]=%E6%97%A5%E6%9C%AC%E8%AA%9E%E5%BD%A2%E6%85%8B%E7%B4%A0%E8%A7%A3%E6%9E%90%E3%82%B7%E3%82%B9%E3%83%86%E3%83%A0JUMAN

6 http://www.kecl.ntt.co.jp/icl/lirg/resources/GoiTaikei/

7 http://www.yoursingapore.com/content/traveller/ja/experience.html

岡部浩司, 河原大輔, 黒橋禎夫(2006) 格フレームを用いたかな表記語の曖昧性解消 言語処理学会第 12 回年次大会

岡部浩司, 河原大輔, 黒橋禎夫(2007) 代表表記による自然言語リソースの整備 言語処理学会第 13 回年次大会

栗林孝行, Francis Bond, 黒田航, 内元清貴, 井佐原均, 神崎享子, 鳥澤健太郎(2010) 日本語ワードネット 1.0 言語処理学会第 16 回年次大会

黒田航, 栗林孝行, Francis Bond, 神崎享子, 井佐原均 (2011) 日本語ワードネットの異表記対応と並行コーパスへの語義タグづけ 言語処理学会第 17 回年次大会

C. Fellbaum, ed.. (1998) WordNet: An Electronic Lexical Database. MIT Press.

F. Bond, H. Isahara, S. Fujita, K. Uchimoto, T. Kuribayashi and K. Kanzaki (2009) Enhancing the Japanese WordNet in The 7th Workshop on Asian Language Resources, in conjunction with ACL-IJCNLP 2009, Singapore.