日本語ワードネット 1.0

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*
*独立行政法人 情報通信研究機構（NICT）
**Nanyang Technological University（NTU）
jwordnet@gmail.com

はじめに

我々は、フリーかつ大規模な階層構造付き日本語概念辞書として日本語ワードネットを構築してきた。これは英語の Princeton WordNet 3.0(Fellbaum, 1998) をもとに構築したものであるが、全５万概念、１５万語義(概念と日本語のペア)により構成されている。
今回、各概念に対する定義文・例文の日本語訳を付したバージョンを 1.0 としてリリースし、http://nlpwww.nict.go.jp/wn-ja/において公開する。
本稿では、その中身とともに他の資源との連携や今後についても触れる。

日本語ワードネット 1.0 の構成

synset の構成

ワードネットでは synset と呼ばれるものを１つの概念の単位とし、それらが互いにリンクしている。synset は概念の説明である定義文、概念を言い表すのに使われる同義語(synonym)、例文、他 synset との関係を表すリンク(上位語/Hypernym、下位語/Hyponym、その他)などから成り、日本語ワードネットの synset の内容のうち当方で付与したものは、日本語の定義文、例文、同義語（以下日本語 synonym）、Open ClipArt Library の画像、日本語語彙体系とのマッピング情報である。 詳しくは図１に"seal"(02076196-n)[1]について例示する。
このうち日本語 synonym は、Princeton WordNet 3.0 の synonym を自動的に日本語に翻訳し、複数言語のワードネットと辞書を利用して曖昧性の解消を行った。 尚、曖昧性のないものはそのまま利用し[2]、また、現在人手によるチェックを継続中である。 そして、本リリース(バージョン 1.0)から新たに加わったものは日本語の定義文と例文である。これは既に部分的に語義曖昧性の解消がなされている Princeton WordNet Gloss Corpus[3]を機械翻訳し、それを下敷きに人手で構築した。
他言語のワードネットにおいても定義文、例文の翻訳が行われているものがあり(伊語、韓国語等)、これらと組み合わせることによって多言語対訳コーパスを構築することが可能となり、また、日本語母語話者が synset の意味を理解する手助けになると期待される。

他の資源とのリンク

日本語ワードネットでは情報を補強するために他の資源とも連携している。

Suggested Upper Merged Ontology (SUMO)

これは Niles と Pease(2001)によって構築され、IEEE からリリースされている上位オントロジー

1 "英語"(synset 番号)。以下同様。
2 例えば、"estivate"(00016183-v)=夏眠。
3 http://wordnet.princeton.edu/glosstag.shtml

Synset	02076196-n
Synonyms	ja 海豹, アザラシ, シール en seal9 fr phoque
Illustration	 animal/seal.png
Def (en)	“any of numerous marine mammals that come on shore to breed; chiefly of cold regions”
Def (ja)	「繁殖のために岸に上がる海洋性哺乳動物の各種；主に寒帯地域に」
Hypernyms	アシカ亜目/pinniped
Hyponyms	？/crabeater_seal ？/eared_seal 海驢/earless_seal
GoiTaikei	⟨⟨537:beast⟩⟩
SUMO	⊂ Carnivore

図 1:”seal”(02076196-n)

であるが、10万余りに及ぶPrinceton WordNet 3.0とのマッピング情報もリリースされており[4]、この情報は日本語ワードネットにおいてもそのまま利用可能である。 例えば図1の”seal”(02076196-n)について、SUMOでは”Carnivore”に包含されている。ワードネットでは「哺乳類」[5]に当たる概念の下位概念であり、上位方向に階層構造を辿っても肉食動物であるかどうかの言及がない。SUMOとのマッピング情報があることにより、その部分を補強できる。また、 SUMOには1万2千点ほどの画像も含まれ、そのうちワードネットとのリンクがある約4,600点も利用可能となっている。

Open Clip Art Library(OCAL)

これは、クリップアートのアーカイブを行っているプロジェクトであり[6]、これに収録されている画像と、我々がリリースしているマッピング情報はともにパブリックドメインである。尚、マッピングについては画像のメタデータを利用して半自動的に抽出した後、人手でチェックを行った。これは画像に付与されているメタデータのうち、タイトルと”Tags”の情報が下位語/Hyponymと上位語/Hypernymの関係にあることを利用したものである。例えば、図1に用いた画像の場合、タイトルが”Seal”で、”Tags”には”mammal”や”animal”が含まれている。 SUMOやOCALの画像とのマッピング情報を持つことにより、概念を理解する視覚的な手助けになる。

日本語語彙大系

これとのマッピングも半自動的に行ったが、曖昧性の解消に機械翻訳システムALT-J/E(Ikehara et al., 1991)を用いた。このマッピング情報にはSUMOとのそれと同様の利点があるが、こちらは麻野間(2001)によるマッピングデータと比較を行うことによってより信頼性を高めることができる。

インターフェース

日本語ワードネットは、基本的に既存のフリーな資源を用いて構築し、フリーという形でリリースしている。その効果で、当方でperlのAPIとweb interfaceを公開したところ、有志により多言語API（java, ruby, Python等）をはじめ、semantic web向けのRDF版が開発され、言語グリッド[7]のインターフェースも開発

4 http://sigmakee.cvs.sourceforge.net/sigmakee/KBs/WordNetMappings/から入手可能。

5 ”mammal”(01861778-n)

6 http://www.openclipart.org/

7 http://langrid.nict.go.jp/jp/

されたと考えられる。尚、これらは「はじめに」に記した日本語ワードネットのHPからもリンクさせて頂いている。

日本語ワードネットとWikipedia

Wikipedia は多くの言語処理関係者に期待をもたれているデータであり、一部にはそれがWordNet に取って変わる言語資源になると期待する人も少なくない。だが、以下に述べる理由から、現状はそれにはほど遠い。
Wikipediaは確かに圧倒的な被覆率を誇るデータである。一例を挙げれば、Sumida, et al. (2008) は日本語 Wikipedia から約240万対の上位語・下位語対を自動獲得した。そのデータの内訳は、上位語の異なり語数が約9.4万，下位語語の異なり語数が約11万であった。下位語の異なり数は圧倒的である。その一方、自動獲得したデータは、黒田ら (2009) でも報告したように、人手で整備の前の状態ではかなりノイズが多かった。問題は二つある。第一に、日本語特有の問題として、上位・下位語共に異表記の含有率が多い。定量的な評価はしていないが、控え目に見積もっても、データの1，2割は異表記で占められている。第二に、自動獲得された上位語の多くは複合名詞句であり、獲得されたままの形で意味処理にかけるのは難しい。先頃「上位語階層データv1.0」がALAGIN フォーラムから公開されたが、これは隅田らの方法で獲得した元のデータの上位語約9.4万語のうち、前処理によって除外された上位語集合を階層化したものである (仕上がりは約6.9万)。この人手クリーニングの動機の一つは、日本語WordNet と対応のある上位語オントロジーを構築することだった。階層化の前の素の上位語は日本語WordNet には10%程度の対応しかなかったが、階層化で日本語WNとの対応率が95%に向上した (Kuroda, et al. 2010)。
これが意味していることは、Wikipedia がWordNetに取って代わるような言語資源になるためには、上位語の階層化処理の自動化が不可欠だということである。

今後の課題

現状のバージョンの課題として、今回詳細に触れなかったが、送り仮名や異表記が不統一という問題がある。そのため、例えば「引き抜く」について、「引抜く」「引きぬく」「ひき抜く」「ひきぬく」等もカバーできるように、統一的に修正を行っていく予定である。
もう一つ、英語のワードネットであるPrinceton WordNet をもとに構築してあるために、日本語概念のカバー率は十分ではない。日本語では一般的だが英語にはない概念 (「ご飯"cooked rice"」、「ランドセル」等) が存在する。予備調査では、一般的な日本語の概念が欠けているだけではなく、誤訳や英語から直接訳せない語や正しいsynsetに割り振られていない日本語が発見された。そこでカバー率と精度向上をめざして、現在、コーパスアノテーションを行っているところである。これを行うことにより、対象としたコーパスをワードネットの例文とすることができ、また、結果は語義曖昧性解消にも用いることも可能であるという利点もある。
また少数ではあるが、日本語では別個の概念であるにも関わらず、ワードネットでは１つの概念となっているものも存在する。たとえば、"player"(10439851-n)では日本語synonymとして「プレーヤー」「選手」等を当てているが、「選手」はHyponym"card player"(09894654-n)と矛盾してしまう。そのため、「『プレーヤー』は全てのHyponymについて適用されるが、『選手』は"card

player”には適用されない」等、個々の日本語 synonym について何らかのタグを付けることが望ましいであろう。synset や階層構造の妥当性については、段階的に検討を行っていきたいと考えている。

参考文献

麻野間 直樹. 2001. 「WordNet と日本語語彙大系のオントロジ対応」. In NAACL Wokshop on WordNet & Other Lexical Resources, pages 89 94. Pittsburgh, USA.

F. Bond, H. Isahara, K. Kanzaki, and K. Uchimoto. 2008. Boot-strapping a WordNet using multiple existing WordNets. In 6th International conference on Language Resources and Evaluation (LREC 2008). Marrakech.

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.

J. W. Breen. 2004. JMDict: a Japanese-multilingual dictionary. In Coling 2004 Workshop on Multilingual Linguistic Resources, pp. 71 78. Geneva.

T. Charoenporn, V. Sornlerlamvanich, C. Mokarat, and H. Isahara. 2008. Semi-automatic compilation of Asian WordNet. 言語処理学会第 14 回年次大会, pp. 1041 1044.

EDR. 1990. Concept dictionary. Technical report, Japan Electronic Dictionary Research Institute, Ltd.

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

S. Ikehara, M. Miyazaki, S. Shirai, A. Yokoo, H. Nakaiwa, K. Ogura, Y. Ooyama, and Y. Hayashi. 1997. Goi-Taikei: A Japanese Lexicon. Iwanami Shoten, Tokyo. 5 volumes/CDROM.

S. Ikehara, S. Shirai, A. Yokoo, and H. Nakaiwa. 1991. Toward an MT system without preediting: Effects of new methods in ALT-J/E. In Third Machine Translation Summit: MT Summit III, pp. 101 106. Washington DC.

K. Kuroda, F. Bond and K. Torisawa. 2010. ”Why Wikipedia needs to make Friends with WordNet” in The 5th International Conference of the Global WordNet Association (GWC-2010), Mumbai.

E. Nichols, F. Bond, T. Tanaka, S. Fujita, and D. Flickinger. 2006. Robust ontology acquisition from multiple sources. In Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge, pp. 10 17. Sydney.

A. Sumida, N. Yoshinaga, and K. Torisawa. 2008. Boosting precision and recall of hyponymy relation acquisition from hierarchical layouts in Wikipedia. In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC-2008).

K. Uchimoto, Y. Zhang, K. Sudo, M. Murata, S. Sekine, and H. Isahara. 2004. Multilingual aligned parallel treebank corpus reflecting contextual information and its applications. In Gilles Se rasset, editor, COLING 2004 Multilingual Linguistic Resources, pp. 57 64. COLING, Geneva, Switzerland.

P. Vossen, ed. 1998. Euro WordNet. Kluwer.