Japanese Wordnet 2.0

Francis Bond 凡土 フランシス

Takyuki Kuribayashi 栗林 孝行

Department of Asian Studies
Palacký University

bond@ieee.org

12th Global Wordnet Conference 2023

Faculty of Arts

The Japanese Wordnet

- Started at National Institute of Information and Communications Technology (NICT)
  - **expand** model
  - cross-lingual inference of new senses (which led to OMW)
- Was hosted at sourceforge
- Now hosted at github: `https://github.com/bond-lab/wnja`
  - This release uses a new format (GWA XML 1.0)
    - ★ Allows for variant forms
    - ★ Structure can be different from PWN
  - A long slog by Takayuki Kuribayashi to add the forms
- Many new extensions require extended software, documentation, corpora, ...

Faculty of Arts

Growth over time

- Initially assume that the semantic structure is the same
  - *dog* ⊂ *animal* ⇒ **犬** ⊂ **動物**
- Added Japanese words to Princeton wordnet synsets

Date	Ver	Concepts	Words	Senses	Misc
2009-02	0.90	49,190	75,966	156,684	initial release
2009-08	0.91	50,739	88,146	151,831	linked to SUMO
2009-11	0.91	49,655	87,133	146,811	
2010-03	1.00	56,741	92,241	157,398	+ definitions, examples
2010-10	1.10	57,238	93,834	158,058	
2012-01					Japanese Semcor
2014-02					NLTK module
2023-02		58,527	90,320	148,676	262,196 forms

Faculty of Arts

What's new?

1. identify gaps during corpus annotation
2. add semantic fields in one coherent set (for multiple languages)
3. re-annotate

- Richer Information
  - Orthographic variants
  - Frequencies
  - Grammatical Notes
- New Entries
  - Pronouns
  - Exclamatives
  - Classifiers
  - Time Expressions
  - Kinship Terms
  - ...
- More Accessible

Faculty of Arts

Orthographic Variants

- Japanese uses many writing systems
  - Kanji: from China
    - ★ some native Japanese variants *tōge* (峠) “mountain pass”
    - ★ Chinese and Native readings: 動 *dō* and *ugo(ku/kasu)*
  - Hiragana for inflections, function words and onomatopoeia
  - Katakana for foreign words and emphasis
- Same word can be written in many ways:
  ugoita (動いた) “moved (intrans.)”, うごいた or ウゴイタ

Faculty of Arts

Display Form

1. If there is an entry in **jumandic** we use their canonical form
2. Prefer kanji to hiragana
3. Prefer new forms to old forms
   (we compiled our own table of new and old forms)
4. If there are multiple katakana variants, prefer the longest
5. 動く, ウゴク, うごく, ugoku

Faculty of Arts

Frequencies

- Added sense frequencies from the NTU-MC and Japanese SemCor
  - **対策**$_3$, **策**$_3$, **措置**$_2$,**方略**, **方策**, **術**, **打つ手**.
- Used to order senses in the display and for a MFS baseline for WSD
- Corpus annotation takes a long time, but really strengthens the resource

Faculty of Arts

Grammatical Notes

- Add a note for Sino-Japanese verbs: `sahen`

```
<LexicalEntry id="wnja-v-74345" note="sahen"> <!-- 読書 0 v -->
      <Lemma writtenForm="読書" partOfSpeech="v"/>
        <Form writtenForm="ドクショ" script="kana"/>
        <Form writtenForm="どくしょ" script="hira"/>
        <Form writtenForm="dokusyo" script="latn"/>
        <Form writtenForm="dokusho" script="latn-hepburn"/>
       <Sense id="wnja-00625119-v-74345" synset="wnja-00625119-v"
              confidenceScore="1.0"/>
 </LexicalEntry>
```

- This helps to determine the inflection, we may add more specialized types for verbs and adjectives (from Jacy: Siegel et al., 2016)

Faculty of Arts

Classifers I

- Denumerated nouns in Japanese must have a classifier, and the choice of classifier is a mixture of semantic and cultural factors.
  - *tsuru ichi.-wa* (鶴一羽) “one-CL stork”
  - *saru ippiki* (猿一匹) “one-CL monkey”
  - *usagi ichi-wa* (兎一羽) “one-CL rabbit”
- For the classifier, we link it to the synsets it classifies (or the hypernyms of the synsets it classifies)
- We follow Bond and Paik (2000); Morgado da Costa et al. (2016)

Faculty of Arts

Classifers II

(1)

76100129-x (羽)	
lemmas:jpn	羽
def:jpn	ツバメやタカやペンギンなどの鳥、またウサギに対しても用いられる分類辞
exe:jpn	日本では、月で一羽のウサギが餅を搗いていると考えられています; 彼は４羽のオウムを飼っています
def:eng	a sortal classifier used for birds such as a swallow, a hawk or a penguin, and also specifically for rabbits
exe:eng	in Japan, people think a rabbit is making rice cake on the moon; he has 4 parrots
exemplifies	06308436-n (classifier)
classifies	01503061-n (bird)
classifies	02324045-n (rabbit)

Faculty of Arts

Pronouns

- Japanese personal pronouns also encode politeness and formality, but are not marked for case
- Japanese demonstrative pronouns have a three-way system (+interrogatives)
  - proximal, medial, distal, interrogative
  - これ、それ、あれ、どれ *kore, sore, are, dore* “this, that, yon, which”
  - ここ、そこ、あそこ、どこ *koko, soko, asoko, doko* “here, there, yonder, where”
  - こんな、そんな、あんな、どんな *konna, sonna, anna, donna* “this kind of, that kind of, that other kind of, which kind of”
  - こう、そう、ああ、どう *kou, sou, aa, dou* “this way, that way, that other way, which way/how”
- We follow Seah and Bond (2014); Morgado da Costa and Bond (2016)
- Also did for English, Chinese and Indonesian

Faculty of Arts

Exclamatives I

- When tagging corpora, or teaching, we found we needed exclamatives and greeting, so we added them (Morgado da Costa and Bond, 2016)
- Many exclamatives are common across languages
  konnichiwa “good day”, *sayonara* “good bye”
- But some are purely Japanese
  as *onegai-shimasu*, *otsukaresama*
- Also added for English and Chinese

Faculty of Arts

Exclamatives II

(2)

80002404-x (お願いします)	
lemmas:jpn	お願いします, お願い
def:jpn	よくしてくれることを求める意味合いの発話
def:eng	an expression that is uttered when you ask for a favor
exemplifies	07109847-n (utterance)
see also	00903098-v (wish)
similar to	80001988-x (please)

Faculty of Arts

Exclamatives III

(3)

80002405-x (お疲れ様)	
lemmas:jpn	お疲れ様, ご苦労様
def:jpn	相手の苦労をねぎらう発話
def:eng	an expression that is uttered when you appreciate someone's work; typically used when someone leaves work
exemplifies	07109847-n (utterance)
see also	01805982-v (appreciate)
similar to	80000666-x (thank you)

g

Faculty of Arts

Time Expressions I

- Many complex times expressions are lexicalized in Japanese (compositional in Chinese)
  - 今朝 *kesa* “this morning”, 今週 *konshuu* “this week”, 来週 *raishuu* “next week”
- We also included days of the month *1st, 2nd*
- Japanese lexicalizes tomorrow, the day after tomorrow and the day after the day after tomorrow (Bond et al., 1997)
  - *ashita* (明日) “tomorrow”
  - *asatte* (明後日) “the day after tomorrow”
  - *shiasatte* (明々後日) “the day after the day after tomorrow”
- for many time expressions there are formal and nonformal variants
  - *ashita* (明日) “tomorrow ”
  - *myounichi* (明日) “tomorrow (formal)”

Faculty of Arts

Time Expressions II

(4)

90000501-n (last year)	
lemmas:jpn	昨年, 去年
lemmas:eng	last year
lemmas:cmn	去年
def:jpn	現在の属する年の直前の年
exe:jpn	去年は盛りだくさんな年だった
def:eng	the year before this year
exe:eng	last year was an eventful one
def:cmn	今年的前一年
hypernym	15203791-n (year)

Faculty of Arts

Kinship terms

- Japanese distinguishes older and younger siblings
- Also older and younger aunts and uncles
  - *oba* (伯母) “an aunt who is older than one’s parent”
  - *oba* (叔母) “an aunt who is younger than one’s parent”.
- We should also distinguish formal and informal variants
  - *oniisan* (お兄さん) “(your) older brother”
  - *ani* (兄) “(my) older brother”

Faculty of Arts

Ontological Differences: Temperature

- English uses the same words for temperature experienced by touching or as a general feeling:
  ⟨*cold, cool, warm, hot*⟩
- Japanese distinguishes
  - the feeling: ⟨寒い, 涼しい, 暖かい, 暑い⟩
  - to-touch: ⟨冷たい, 温かい, 熱い⟩
- English uses a single word for water of any temperature: *water*
- Japanese uses different words for cold (non-hot) water and hot water:
  水 vs 湯

Faculty of Arts

Other Examples I

(5)

80001626-n (soba_noodle)	
lemmas:jpn	蕎麦
lemmas:eng	soba
def:jpn	そば粉で作られた細い麺
def:eng	narrow noodle made from buckwheat
hypernym	(noodle)

(6)

80002377-n (castle construction)	
lemmas:jpn	築城
def:jpn	城の建設
def:eng	the construction of castles
hypernym	(construction)

Faculty of Arts

Other Examples II

(7)

90000315-n (hajjah)	
lemmas:jpn	ハジャ
lemmas:eng	hajjah
def:jpn	メッカへの巡礼を行った女性
def:eng	a woman who has made the pilgrimage to Mecca
hypernym	(haji)
category	(muslim)

(8)

80001731-n (exchange student)	
lemmas:jpn	留学生
lemmas:eng	exchange student
def:jpn	海外で勉強する学生
def:eng	a student who studies abroad
hypernym	(student)

Faculty of Arts

Other Examples III

(9)

80000338-n (Shunto)	
lemmas:jpn	春闘
lemmas:eng	spring wage negotiation
def:jpn	毎年労働組合が、賃金引き上げなどの要求を掲げて行う全国的な闘争
def:eng	annual event by Japanese workers union when wages are renegotiated
hypernym	(protest)

Faculty of Arts

More Accessible

- Moved to github: `https://bond-lab.github.io/wnja/`
- LMF (XML) license and citation bundled together
- Can be read by OMW 2.0
- Can by read by python WN module

Faculty of Arts

Conclusions

- This paper presents the current state of the Japanese Wordnet: **wnja**.
- We hope that **wnja** will continue to be a useful resource not only for natural language processing, but also for language education/learning and linguistic research.
- In future work, we want to look more at the description of formality and politeness, as well as to increase the coverage.

Faculty of Arts

Acknowledgements

Some of this work was done while visiting the Humanities Center at Tokyo University, thanks to Tsuneko Nakazawa and Tsuneaki Kato.

Faculty of Arts

References I

Francis Bond, Kentaro Ogura, and Hajime Uchino. 1997. Temporal expressions in Japanese-to-English machine translation. In *Seventh International Conference on Theoretical and Methodological Issues in Machine Translation: TMI-97*, pages 55–62. Santa-Fe.

Francis Bond and Kyonghee Paik. 2000. Re-using an ontology to generate numeral classifiers. In *18th International Conference on Computational Linguistics: COLING-2000*, pages 90–96. Saarbrücken.

Luís Morgado da Costa and Francis Bond. 2016. Wow! what a useful extension to wordnet! In *10th International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož.

Luís Morgado da Costa, Francis Bond, and Helena Gao. 2016. Mapping and generating classifiers using an open Chinese ontology. In *Proceedings of the 8th Global Wordnet Conference (GWC 2016)*, pages 248–254.

Faculty of Arts

References II

Yu Jie Seah and Francis Bond. 2014. Annotation of pronouns in a multilingual corpus of Mandarin Chinese, English and Japanese. In *10th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*. Reykjavik.

Melanie Siegel, Emily Bender, and Francis Bond. 2016. *Jacy — An Implemented Grammar of Japanese*. CSLI Publications.

Faculty of Arts