Ontology Acquisition from Definitions

David Moeljadi

Nanyang Technological University

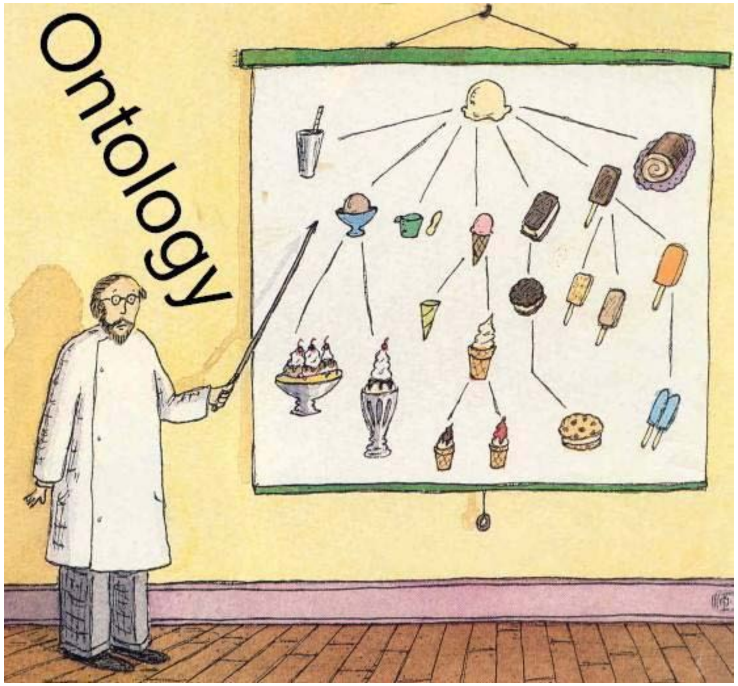
September 11, 2014

Outline

1. **Motivation**
   Why do we need ontology? How to make it?
2. **Resources used for acquiring ontology**
   Lexeed lexicon, JACY grammar, Hinoki treebank
3. **Ontology construction**
   Extract synonym, hypernym, meronym, domain
4. **Evaluation**
   Verification with GoiTaikei and WordNet,
   human evaluation
5. **Further Work**

Motivation (1/2)

- Our ultimate goal is to understand natural language
- Ontologies are an important resource in NLP:
  - machine translation,
  - question answering,
  - word-sense disambiguation

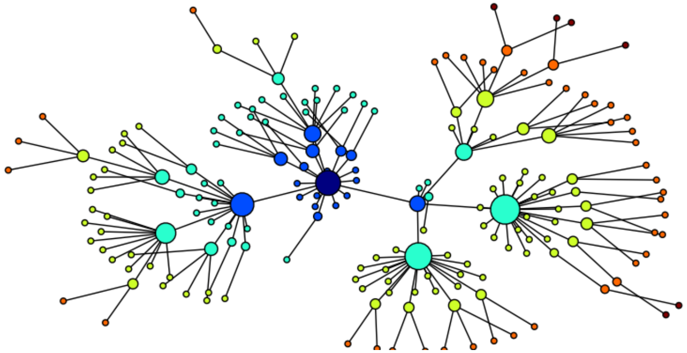


Motivation (2/2)

Manually built ontologies

- WordNet for English (Fellbaum 1998)
- GoiTaikei for Japanese (Ikehara et al. 1997)

- Difficult to construct
- Maintain entirely by hand



Automatically built ontologies

Use

Deep and shallow parsing technologies

Simple relation extractor

- Simpe rules
- Can easily be extended to cover any language

Our Resources (1/7)

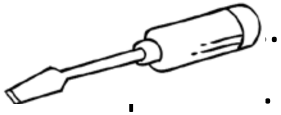
- **Japanese Semantic Lexicon (Lexeed)**
  - Most familiar 28,270 basic words
  - Familiarity is estimated by psychological experiments
  - Contains all words with familiarity ≥ 5.0 (1 – 7)
  - Covers 75% of tokens in a typical newspaper
  - Basic words (and function words) used for definitions and example sentences
  - 46,347 senses and 81,000 definition sentences
  - POS tag and morpheme analysis with ChaSen

Our Resources (2/7)

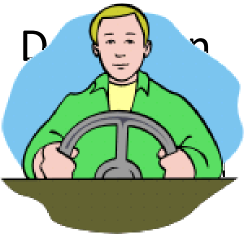
- **Japanese Semantic Lexicon (Lexeed)**

Headword	ドライバー	*doraiba-*		
POS	名詞	noun	名詞-一般	noun-lex
Familiarity	6.5 [1-7]			

Sense 1



Sense 2



D n [S1 自動車/を/運転/する/人/。]
Someone who drives a car.

S1 父/は/優良/ドライバー/として/表彰/さ/れ/た/。]
My father was awarded as an excellent driver.

Sense 3

Definition …
Ex …

Figure 1: A sample entry for the word *doraiba-*$_2$ “driver$_2$”

Our Resources (3/7)

- **JApanese grammar developed at CSLI and YY (JACY)**
  - HPSG-based grammar of Japanese
  - Developed by Melanie Siegel (+Bender, Shimada)
  - 36,000 word vocabulary
  - Integrated with ChaSen morphological analyser
  - Can download from: www.dfki.uni-sb.de/~siegel/grammar-download/JACY-grammar.html
  - Developed with Linguistic Knowledge Builder (LKB)
  - Runs with the efficient run-time engine PET
  - Profiling and Treebanking with [incr tsdb()]

Our Resources (4/7)

- **Hinoki Treebank**

  - Inspired by the Redwoods treebank of English (Oepen et al. 2002)
  - Combine the classic approaches
    - Compiling a Japanese HPSG (JACY)
    - Parsing definition sentences (Lexeed)
    - Annotating corpora for training (Hinoki)

  - Each part feeds into the others
    - Use the grammar to parse the dictionaries
      Treebank and sense tag the parsed sentences
    - Build an ontology from the parsed definitions
    - Use the ontology to enrich the language model

檜

Our Resources (5/7)

- **Parse result for driver$_2$ (tree)**

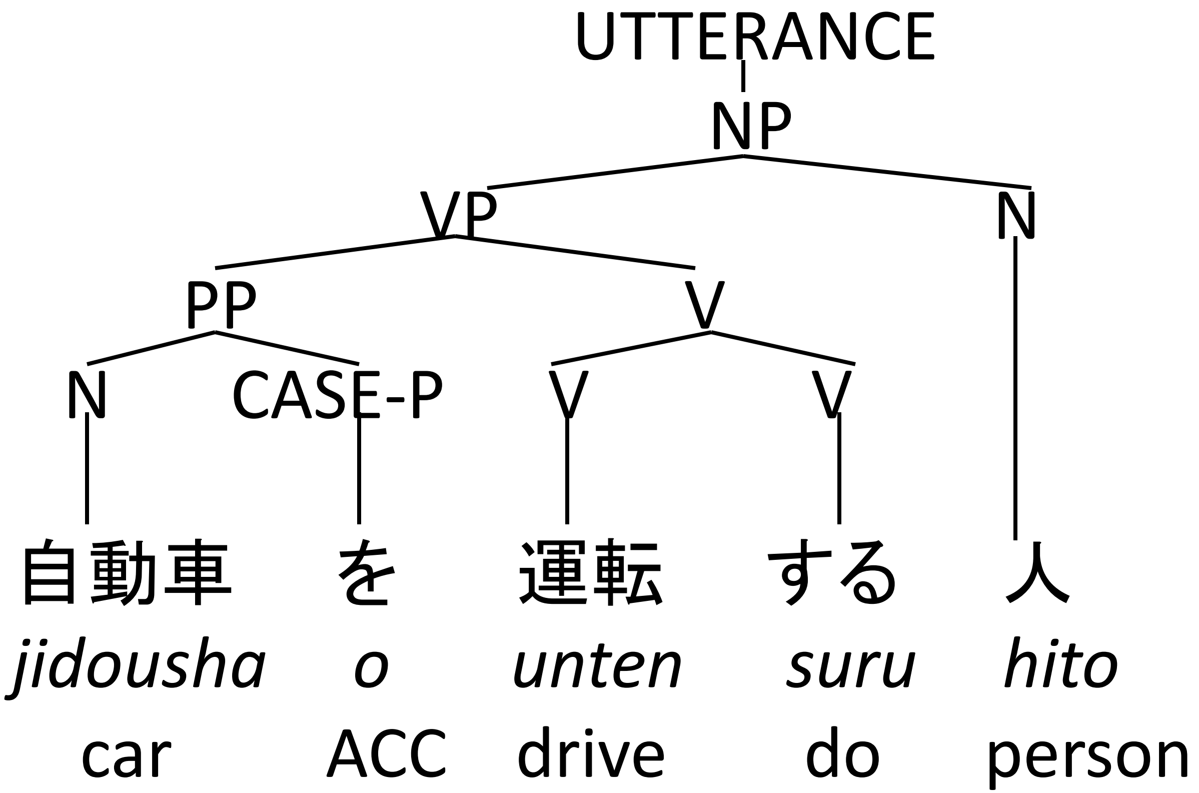


Figure 2: Phrase structure tree used for treebanking for *doraiba*-$_2$ “driver$_2$”

Our Resources (6/7)

- **Parse result for driver$_2$ (RMRS)**

```
hook(h1)
     proposition_m_rel(h1,h3:)
               qeq(h3:,h17)
     _jidousha_n(h4,x5:)
     udef_rel(h6,x5:)
               RSTR(h6,h7:)
               BODY(h6,h8:)
               qeq(h7:,h4)
     _unten_s_2(h9,e11:present:)
               ARG1(h9,x10:)
               ARG2(h9,x5:)
     _hito_n(h12,x10:)
               ING(h12:,h10001:)
     …
```

RMRS from JACY (deep)

```
hook(h9)


     _jidousha_n(h1,x2)
     o_rel(h3,u4)




     _unten_s_2(h5,e6)
     suru_rel(h7,e8)

     _hito_n(h9,x10)
```

RMRS from ChaSen (shallow)

Figure 3: Deep and shallow RMRS results for *doraiba*-$_2$ “driver$_2$”

Our Resources (7/7)

- **Lexeed + Hinoki**

Headword ドライバー *doraiba-*

POS 名詞 noun 名詞-一般 noun-lex

Familiarity 6.5 [1-7] Frequency 37 Entropy 0.79

Sense 1 …

Sense 2

$P(S_2) = 0.84$

Definition [S1 自動車$_1$/を/運転$_1$/する/人$_1$/。]

Someone who drives a car.

Example [S1 父$_1$/は/優良$_1$/ドライバー$_2$/として/表彰$_1$/さ/れ/た/。]

My father was awarded as an excellent driver.

Hypernym 人$_1$ *hito* person

Sem. Class < 292: chauffeur/driver > (⊂ < 4: person >)

WordNet *driver*$_1$ (⊂ person$_1$)

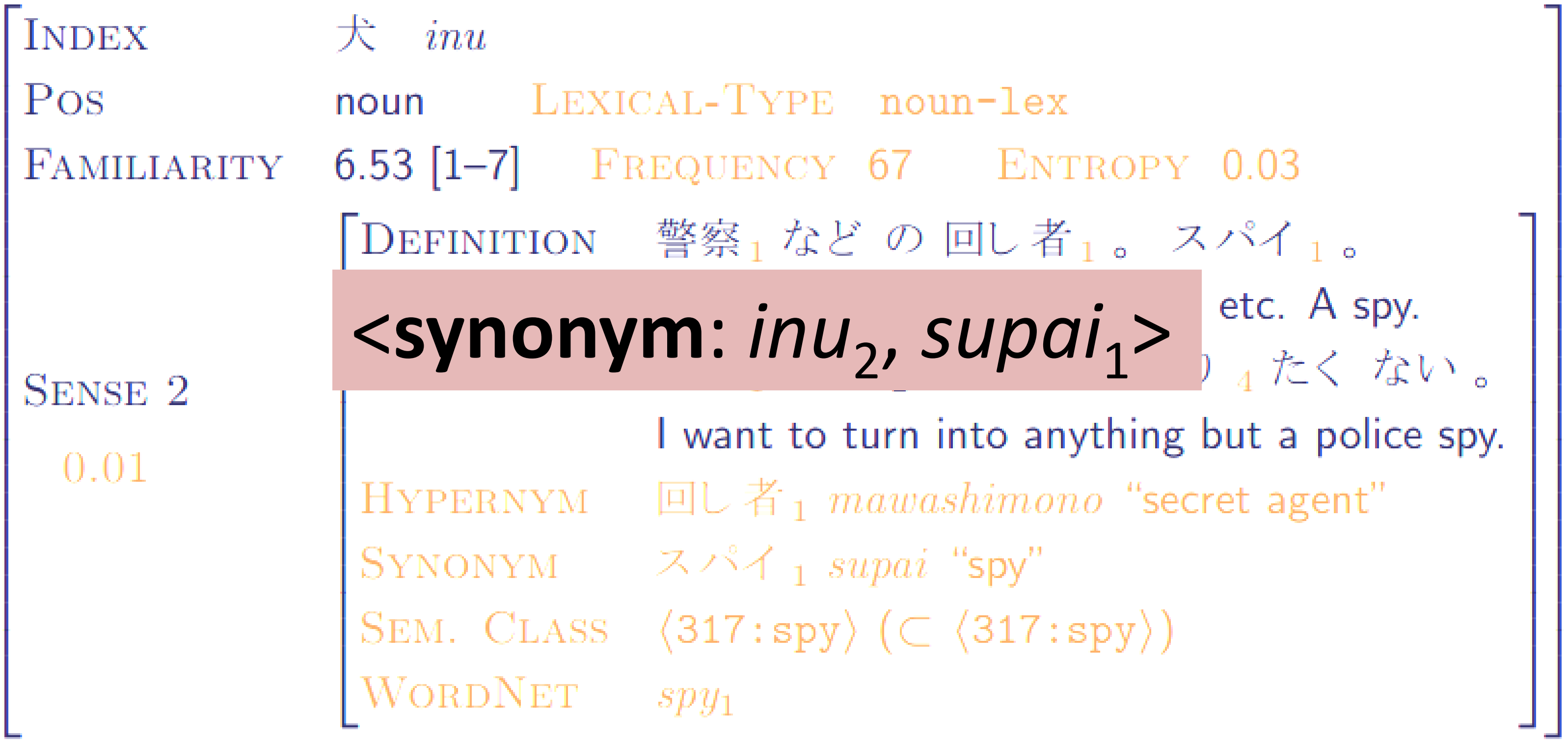
Sense 3 …

Figure 4: A sample entry for the word *doraiba-*$_2$ “driver$_2$”

Ontology Construction (1/5)

1. **If the number of real predicates = 1**

   return: <**synonym**: headword, predicate>



INDEX 犬 *inu*

POS noun LEXICAL-TYPE noun-lex

FAMILIARITY 6.53 [1–7] FREQUENCY 67 ENTROPY 0.03

SENSE 2

0.01

DEFINITION 警察$_1$ など の 回し者$_1$。 スパイ$_1$。

etc. A spy.

)$_4$ たく ない。

I want to turn into anything but a police spy.

HYPERNYM 回し者$_1$ *mawashimono* "secret agent"

SYNONYM スパイ$_1$ *supai* "spy"

SEM. CLASS ⟨317:spy⟩ (⊂ ⟨317:spy⟩)

WORDNET spy_1

<**synonym**: inu_2, $supai_1$>

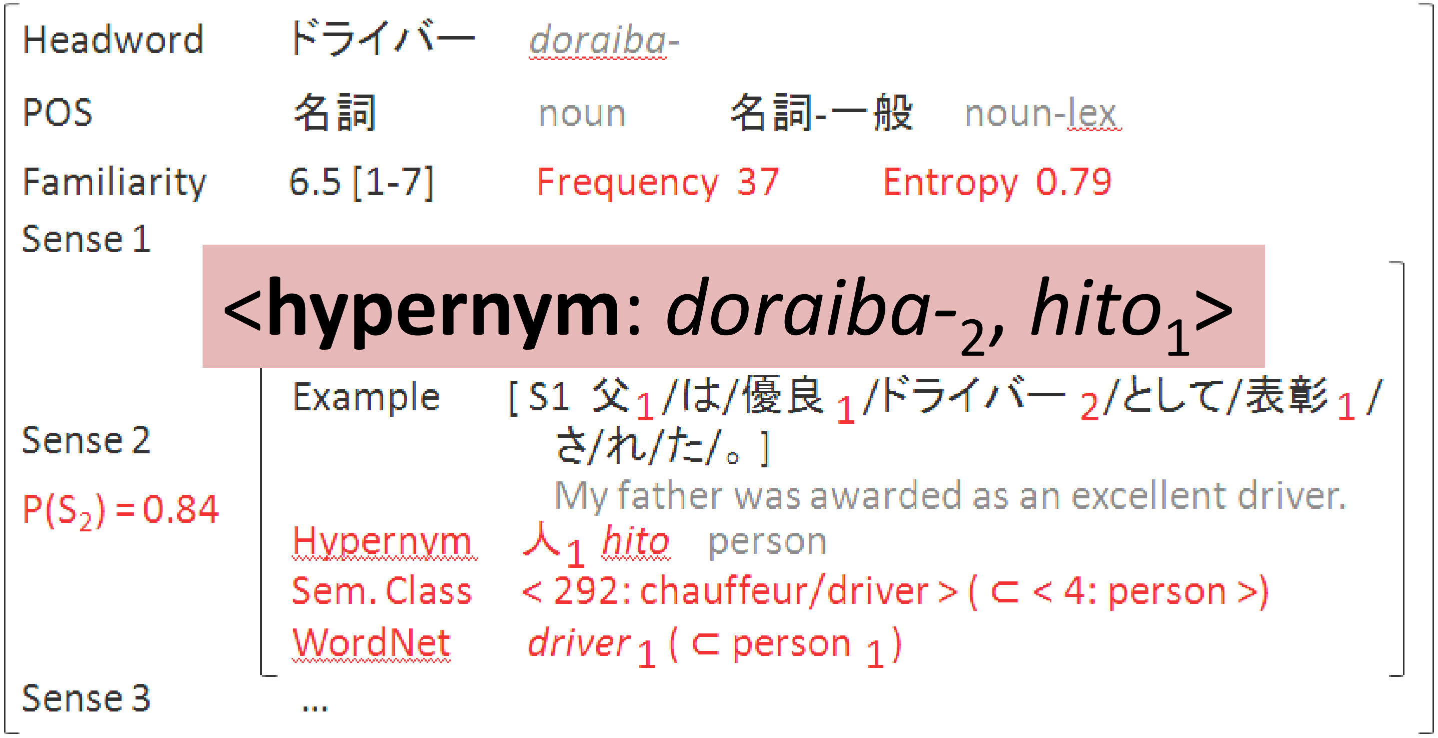
Figure 5: A sample entry for the word inu_2 "dog_2"

Ontology Construction (2/5)

2. **If the number of real predicates > 1**

   look at the predicate with the widest scope (genus)

   return: <**hypernym**: headword, predicate>



Headword	ドライバー	*doraiba-*		
POS	名詞	noun	名詞-一般	noun-lex
Familiarity	6.5 [1-7]	Frequency 37		Entropy 0.79

Sense 1

<**hypernym**: *doraiba-*$_2$, *hito*$_1$>

Sense 2

P(S_2) = 0.84

Example	[S1 父$_1$/は/優良$_1$/ドライバー$_2$/として/表彰$_1$/さ/れ/た/。] My father was awarded as an excellent driver.
Hypernym	人$_1$ *hito* person
Sem. Class	< 292: chauffeur/driver > (⊂ < 4: person >)
WordNet	*driver*$_1$ (⊂ person$_1$)

Sense 3 …

Figure 6: A sample entry for the word *doraiba-*$_2$ “driver$_2$”

Ontology Construction (3/5)

3. If the number of real predicates > 1

if the highest scoping word is an explicit relation

e.g. 略 *ryaku* “abbreviation”

return: <**abbreviation**: headword, predicate>

ア : アルプス、または 日本アルプス の 略

a : arupusu , matawa nihon-arupusu no ryaku

a : alps , or japan-alps ADN abbreviation

a : an abb ps

<**abbreviation**: a_1, *arupusu*$_2$>

<**abbreviation**: a_1, *nihon-arupusu*$_1$>

Ontology Construction (4/5)

3. **If the number of real predicates > 1**

   if the highest scoping word is an explicit relation

   e.g. 一種 *isshu* “a kind of”

   return: <**hypernym**: headword, predicate>

   e.g. 総称 *soushou* “general term”

   return: <**hyponym**: headword, predicate>

   e.g. 部分 *bubun* “part of”

   return: <**meronym**: headword, predicate>

   e.g. 敬称 *keishou* “honorific name”

   return: <**name:honorific**: headword, predicate>

   etc.

Ontology Construction (5/5)

4. **If there is an adpositional phrase modifies a non-expressed predicate,**
   extract the modifiers and take the head of the noun phrase to be the domain.

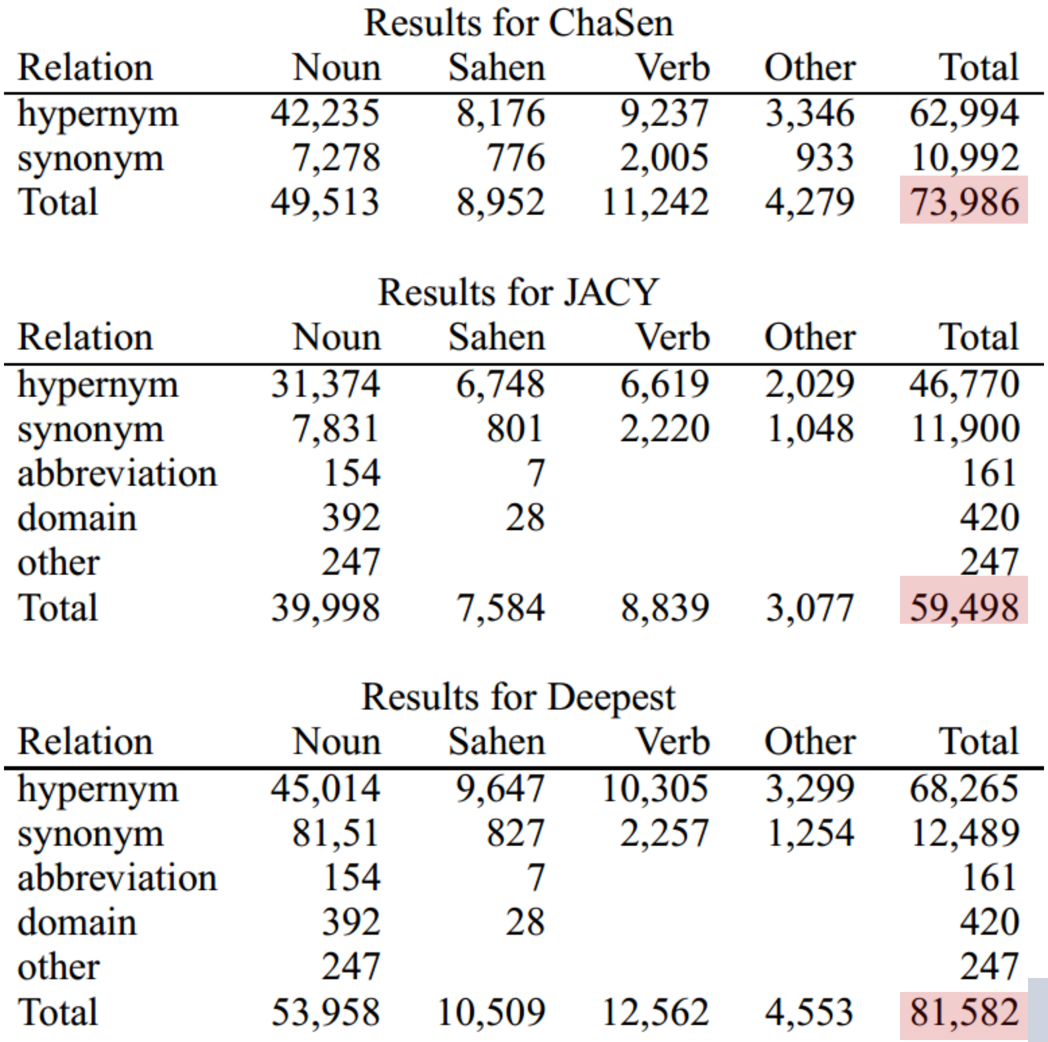
Example from driver$_3$:

ゴルフ	で	、	（ドライバー	は）	遠距離用	の	クラブ	（だ）
gorufu	de		doraiba-	wa	enkyoriyou	no	kurabu	da
golf	in		driver		long-distance	ADN	club	

In golf, (a driver$_3$ is) a club for playing long strokes.

<**domain**: *doraiba-*$_3$, *gorufu*$_1$>

Evaluation (1/7)



extract less relationships

Results for ChaSen

Relation	Noun	Sahen	Verb	Other	Total
hypernym	42,235	8,176	9,237	3,346	62,994
synonym	7,278	776	2,005	933	10,992
Total	49,513	8,952	11,242	4,279	73,986

high coverage

extract more relationships

Results for JACY

Relation	Noun	Sahen	Verb	Other	Total
hypernym	31,374	6,748	6,619	2,029	46,770
synonym	7,831	801	2,220	1,048	11,900
abbreviation	154	7			161
domain	392	28			420
other	247				247
Total	39,998	7,584	8,839	3,077	59,498

low coverage

many relationships

Results for Deepest

Relation	Noun	Sahen	Verb	Other	Total
hypernym	45,014	9,647	10,305	3,299	68,265
synonym	81,51	827	2,257	1,254	12,489
abbreviation	154	7			161
domain	392	28			420
other	247				247
Total	53,958	10,509	12,562	4,553	81,582

highest coverage

Table 1: Results of ontology extraction (Lexeed)

Evaluation (2/7)

- **Verification with hand-crafted ontologies**

  - Compare the extracted ontology with GoiTaikei
    - 2,710 semantic classes
    - marked for 264,312 nouns
    - we can only compare nouns

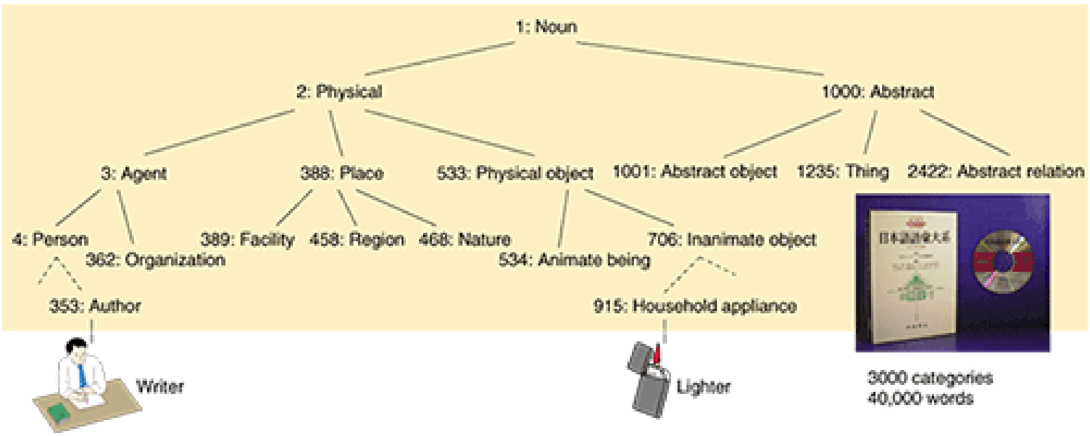


Figure 7: Common noun semantic categories of GoiTaikei (https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr200809sf5.html)

Evaluation (3/7)

- **Verification with hand-crafted ontologies**

  - Compare compatible subsumption relations

  - headword w_h, genus term w_g, semantic classes c

    $$\exists(c_h, c_g) : \{c_h \subset c_g; c_h \in C(w_h); c_g \in C(w_g)\}$$

  - If at least one of the index word’s classes is subsumed by at least one of the genus’ classes, the relationship is confirmed
  - Reverse for Hyponym: $c_g \subset c_h$

  - Headword and Genus are often in the same Goi-Taikei semantic class (45.4%)

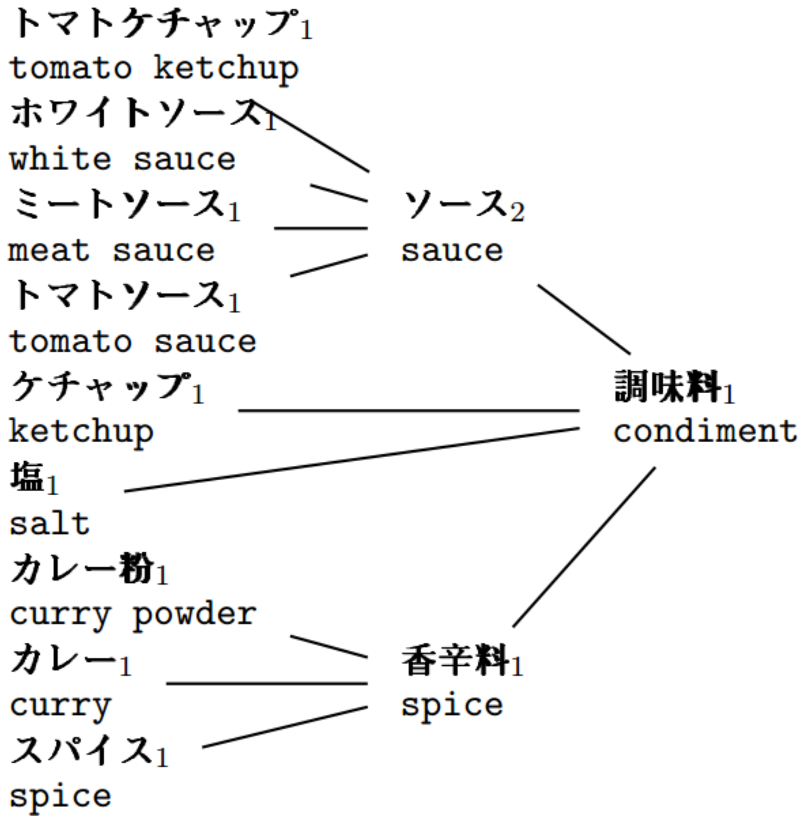
    buta niku “pork” and *niku* “meat”

    doramu “drum” and *dagakki* “percussion instrument”

Evaluation (4/7)

- **Verification with hand-crafted ontologies**

  We extracted pairs with more information than Goi-Taikei



トマトケチャップ$_1$
tomato ketchup
ホワイトソース$_1$
white sauce
ミートソース$_1$
meat sauce
トマトソース$_1$
tomato sauce
ケチャップ$_1$
ketchup
塩$_1$
salt
カレー粉$_1$
curry powder
カレー$_1$
curry
スパイス$_1$
spice

ソース$_2$
sauce

調味料$_1$
condiment

香辛料$_1$
spice

Figure 8: Refinement of the class condiment, deduced from Lexeed

Evaluation (5/7)

- **Verification with hand-crafted ontologies**

  - We tested cross-linguistically by looking up the headwords in a translation lexicon (ALT-JE [Ikehara et al. 1991] and EDICT [Breen 2004])

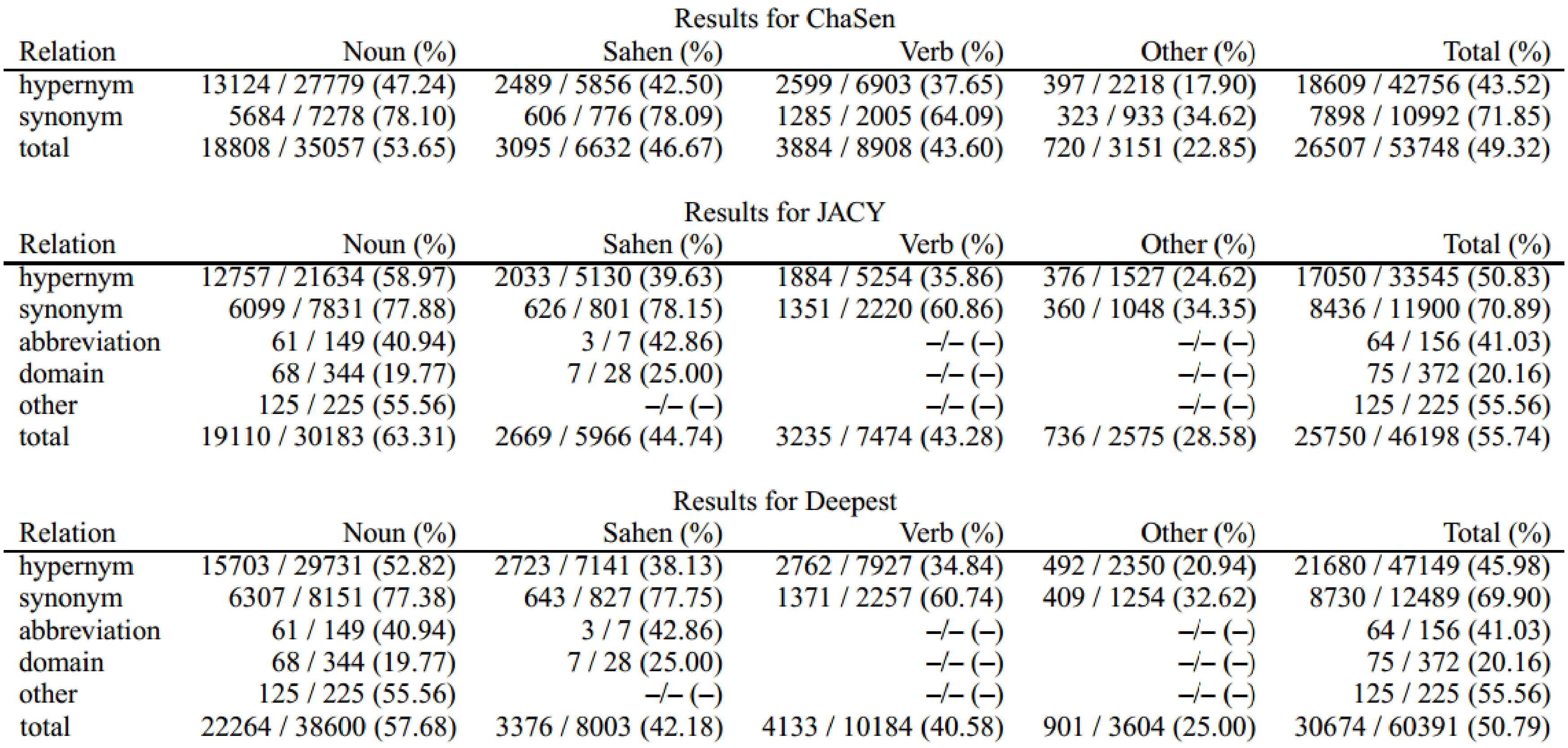
  - We linked to the appropriate WordNet synsets

  - GoiTaikei and WordNet both lack complete cover – over half the relations were confirmed with only one source, either GoiTaikei or WordNet.

Machine readable dictionary is a useful source of these relations

Evaluation (6/7)

- **Verification with hand-crafted ontologies**



Results for ChaSen

Relation	Noun (%)	Sahen (%)	Verb (%)	Other (%)	Total (%)
hypernym	13124 / 27779 (47.24)	2489 / 5856 (42.50)	2599 / 6903 (37.65)	397 / 2218 (17.90)	18609 / 42756 (43.52)
synonym	5684 / 7278 (78.10)	606 / 776 (78.09)	1285 / 2005 (64.09)	323 / 933 (34.62)	7898 / 10992 (71.85)
total	18808 / 35057 (53.65)	3095 / 6632 (46.67)	3884 / 8908 (43.60)	720 / 3151 (22.85)	26507 / 53748 (49.32)

Results for JACY

Relation	Noun (%)	Sahen (%)	Verb (%)	Other (%)	Total (%)
hypernym	12757 / 21634 (58.97)	2033 / 5130 (39.63)	1884 / 5254 (35.86)	376 / 1527 (24.62)	17050 / 33545 (50.83)
synonym	6099 / 7831 (77.88)	626 / 801 (78.15)	1351 / 2220 (60.86)	360 / 1048 (34.35)	8436 / 11900 (70.89)
abbreviation	61 / 149 (40.94)	3 / 7 (42.86)	–/– (–)	–/– (–)	64 / 156 (41.03)
domain	68 / 344 (19.77)	7 / 28 (25.00)	–/– (–)	–/– (–)	75 / 372 (20.16)
other	125 / 225 (55.56)	–/– (–)	–/– (–)	–/– (–)	125 / 225 (55.56)
total	19110 / 30183 (63.31)	2669 / 5966 (44.74)	3235 / 7474 (43.28)	736 / 2575 (28.58)	25750 / 46198 (55.74)

Results for Deepest

Relation	Noun (%)	Sahen (%)	Verb (%)	Other (%)	Total (%)
hypernym	15703 / 29731 (52.82)	2723 / 7141 (38.13)	2762 / 7927 (34.84)	492 / 2350 (20.94)	21680 / 47149 (45.98)
synonym	6307 / 8151 (77.38)	643 / 827 (77.75)	1371 / 2257 (60.74)	409 / 1254 (32.62)	8730 / 12489 (69.90)
abbreviation	61 / 149 (40.94)	3 / 7 (42.86)	–/– (–)	–/– (–)	64 / 156 (41.03)
domain	68 / 344 (19.77)	7 / 28 (25.00)	–/– (–)	–/– (–)	75 / 372 (20.16)
other	125 / 225 (55.56)	–/– (–)	–/– (–)	–/– (–)	125 / 225 (55.56)
total	22264 / 38600 (57.68)	3376 / 8003 (42.18)	4133 / 10184 (40.58)	901 / 3604 (25.00)	30674 / 60391 (50.79)

Table 2: Results confirmed for Lexeed (for 46,000 senses)

Evaluation (7/7)

- **Human evaluation**

  - 1,471 relations were selected using a stratified method
  - only synonyms and any relationships extracted from the first sentence
  - evaluated by native speakers of Japanese
  - The result of the judgement:
    - ▪ the relations are accurate 88.99%
    - ▪ slightly higher (91.8%) for noun relationship only (Tokunaga et al. 2001)
  - Three sources of the errors found:
    - ▪ lack of identified explicit relationship
    - ▪ lack of information from the shallow parse
    - ▪ errors in the argument structure of the deep parse

Further Work (1/1)

- Improve the grammar (and parse ranker)
  - add more grammatical phenomena
  - cover (percentage of sentences parsed)
  - precision (percentage of useful relations extracted)
- Extraction of more explicit relations
  - antonym, …
- Extend to different languages (English, …)
- Link to hand-crafted ontologies, to furtherlink together senses of words across languages
  - cross-lingual ontology for machine translation

References

- Slides borrow from Francis Bond (20th International Conference on Computational Linguistics COLING-2004) and Baldwin et al. 2010 (http://compling.hss.ntu.edu.sg/courses/hg7017/pdf/lesk.pdf)
- Eric Nichols, Francis Bond, and Daniel Flickinger (2005) “Robust ontology acquisition from machine-readable dictionaries”. In *Proceedings of the International Joint Conference on Artificial Intelligence IJCAI-2005*, 1111–1116, Edinburgh.
- Francis Bond, Eric Nichols, Sanae Fujita and Takaaki Tanaka (2004) “Acquiring an Ontology for a Fundamental Vocabulary”. In 20th International Conference on Computational Linguistics (COLING-2004), 1319–1325, Geneva.
- 笠原 要, 佐藤 浩史, Francis Bond, 田 中 貴秋, 藤田 早苗, 金杉 友子 and 天野 昭成 (2004a)「基本語意味データベース:Lexeed」の構築 [Construction of a Japanese Semantic Lexicon: Lexeed] In IPSJ SIG Technical Report 2004-NLC-159, 75–82, Tokyo.
- Sanae Fujita, Takaaki Tanaka, Francis Bond, and Hiromi Nakaiwa (2006) “An implimented description of Japanese: The Lexeed dictionary and the Hinoki treebank”. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 65–68, Sydney, 2006.