Boot-strapping a WordNet using multiple existing WordNets

Francis Bond,
Hitoshi Isahara, Kyoko Kanzaki, Kiyotaka Uchimoto

Language Infrastructure Group
National Institute of Information and Communications Technology
bond@ieee.org,
{bond,isahara,kanzaki,uchimoto}@nict.go.jp

LREC-2008

Overview

- We are building an open Japanese WordNet, based on Princeton wordnet
  - First version built automatically (62,000 synsets)
    Disambiguated with French, Spanish and German wordnets
  - $\approx$ 20,000 synsets hand checked
  - Linking to Japanese translation of SemCor
- Will release it in June 2008 (Coming soon!)
  - Similar license to Princeton WordNet
- Building a community of users to extend/maintain

WordNet

- Princeton WordNet®: is a large lexical database of English.
- Nouns, verbs, adjectives and adverbs grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept.
- Synsets interlinked
  - hypernym/hyponym (is-a)
  - meronym/part (has-a)
  - domain
- Free license

European WordNets

- WordNets have been made for many languages . . .

Part of Speech	Number of Synsets English	French	Spanish	German
Noun	82,115	17,826	7,902	9,951
Verb	13,767	4,919	3,775	5,166
Adjective	18,156	0	3,879	15
Adverb	3,621	0	0	0
Total	117,659	22,745	15,556	15,132

- EuroWordNet, BalkaNet, . . . (not all free)

WordNet: Relations

- Noun relations:
  - hypernym, hyponym, coordinate, holonym, meronym
- Verb relations:
  - hypernym, troponym, coordinate, entailment
- Adjectives
  - related noun, similar to, antonym, participle of verb
- Adverbs
  - root adjective
- Other Relations
  - domain, derivationally related form

Usability and Accessibility

Usability :

- Originally designed for psycholinguistic experiments
- Widely used in NLP
  - PP attachment
  - WSD - senseval

Accessibility :

- downloadable
- redistributable
- actively maintained

Previous Work — Japanese

- Many good thesauruses/lexicons
  - Bunrui Goihyou
  - Nihongo GoiTaikei
  - Iwanami MRD
  - Lexeed

- Usable but not accessible

Previous Work — Japanese (WN)

- Much work on building a Japanese WordNet
  - Noun part — synsets and glosses — translated into Japanese (Hayashi, 1999)
  - Multi-lingual Semantic Network put on-line (**E5-3**)
    `http://two.dcook.org/software/mlsn/main.php`
  - Some entries translated using context (Kaji and Watanabe, 2006)
  - Translation of (English) WordNet and EDR into RDF (Koide et al., 2006)

- But still no large-scale freely available Japanese WordNet

NICT'S approach

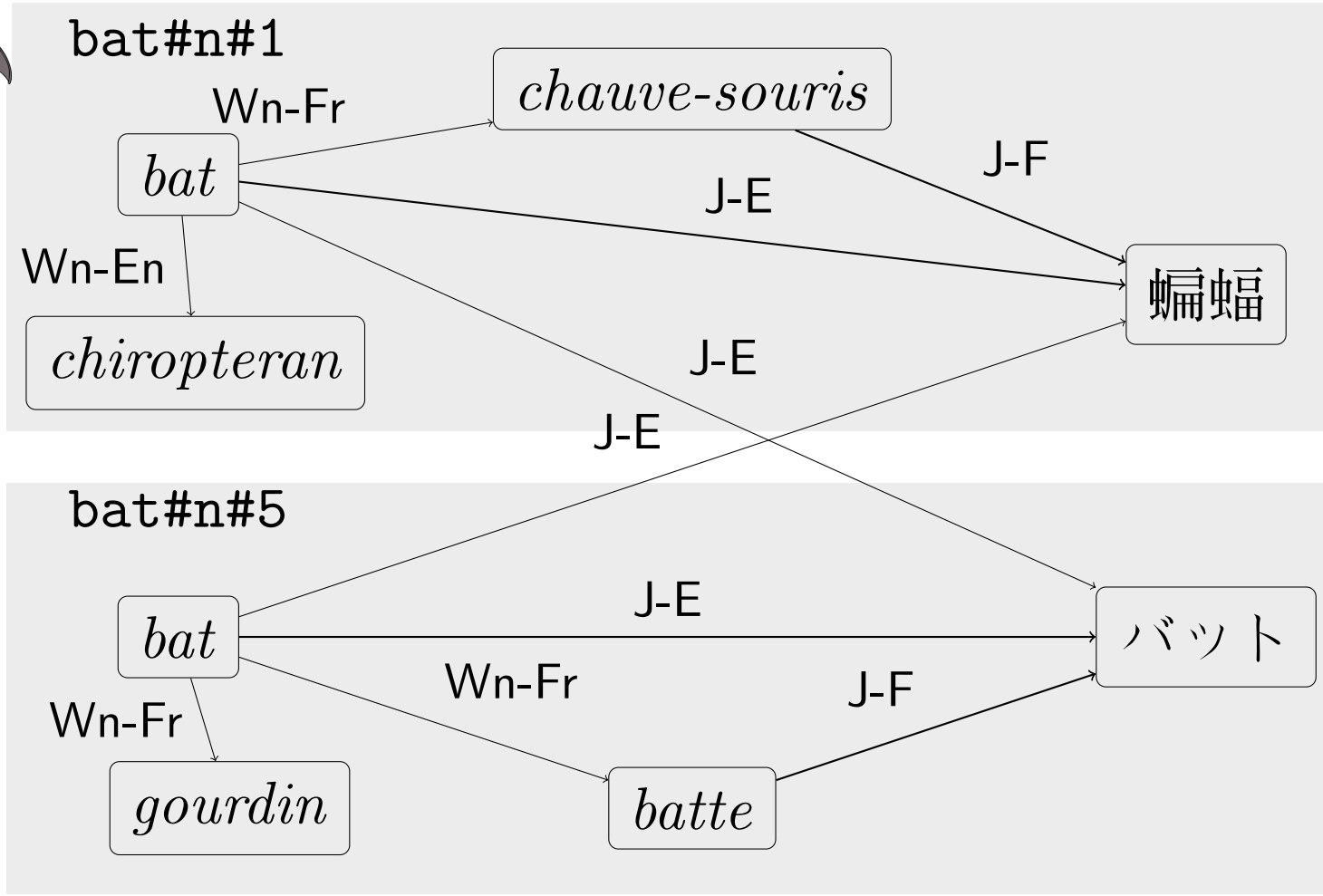
- Stage 0
  - Semi-automatically translate English WordNet (3.0)
- Stage 1
  - Manually correct the top 10,000 entries
  - This includes the 5,000 **core** synsets
- Stage 2 (in progress: poster yesterday)
  - Correct the next most frequent 15,000 entries
  - Create a Japanese version of SemCor
  - Release WordNet-ja v1.0

Semi-Automatically Translation

- For each synset in WordNet 3.0
  - Find its equivalents in WN-Fr, WN-Es, Wn-De
  - Look up translations for all equivalents $\{\mathrm{J}_e\}$, $\{\mathrm{J}_f\}$, $\{\mathrm{J}_s\}$, $\{\mathrm{J}_d\}$
  - Rank Japanese equivalents
    score $s = |\text{links}| + 10$ for links in two languages

The result is a WordNet with multiple Japanese candidates for each synset ranked by score.

Linking with Multiple WordNets

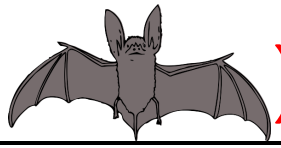


Example (bat#n#5: )

- bat#n#5 “a club used for hitting a ball in various games”

Lang	Word	Dic	Words
En	bat	JMDict	バット, 蝙蝠
	bat	EDR	バット, 蝙蝠, ラケット, 打棒, コウモリ, 蚊食い鳥, 蚊食鳥
Fr	batte	JMDict	バット
	gourdin	JMDict	棍棒

- Ranking: バット (23), 蝙蝠 (2), 棍棒, ラケット, 打棒, コウモリ, 蚊食い鳥, 蚊食鳥 (1)

Example (bat#n#1: )

- bat#n#1 "nocturnal mouselike mammal with forelimbs modified to form membranous wings . . . "

Lang	Word	Dic	Words
En	bat	JMDict	バット, 蝙蝠
	bat	EDR	バット, 蝙蝠, ラケット, 打棒, コウモリ, 蚊食い鳥, 蚊食鳥
chiropteran	—		
Fr	chauve-souris	JMDict	バット
De	Fledermaus	JMDict	バット, 蝙蝠, かわほり, コウモリ, こうもり, 蚊喰鳥

- Ranking: バット? (34), 蝙蝠 (23), コウモリ, 蚊食鳥 (22), 蚊食い鳥, 棍棒, ラケット, 打棒, かわほり, こうもり (1)

Resources: Lexicons

Part of Speech	Number of Word-Pairs					
	ja-en			ja-de	ja-fr	ja-es
	JMDict	EDR	Lifsci	JMDict	JMDict	Goihata
Noun	165,984	504,450	44,567	143,753	24,348	0
Verb	22,209	184,250	4,741	26,502	7,762	133
Adjective	16,861	44,961	11,212	17,121	4,582	70
Adverb	6,180	20,125	1,266	5,915	1,478	0
Unknown	3	0	0	0	0	3,548
Total	225,803	758,568	62,210	199,260	39,447	3,751

- Many more English dictionaries

Results: Japanese Synsets Created

Part of Speech	Number of Synsets $s > 10$	$s > 1$	All
Noun	9,243	36,432	42,725
Verb	2,991	9717	10,321
Adjective	629	6,283	8,915
Adverb	9	1,317	1,726
Total	12,872	53,749	63,687

- Candidates created for over half of WordNet
- Will try to make more with new lexicons
  - from Wikipedia (done by MLSN), Bracket-Dic, . . .

Results: Base Japanese Synsets

Part of Speech	Number of Synsets $s > 10$	$s > 1$	All
Noun	2,429	3,264	3,279
Verb	656	988	993
Adjective	153	586	653
Adverb	0	0	0
Total	3,238	4,838	4,925

- Almost all 5,000 — Good coverage of the core (`www.globalwordnet.org/gwa/gwa_base_concepts.htm`)

Precision for Base Japanese Synsets

	Appropriate Translation Candidates			
	$s > 10$	$10 > s > 1$	$s = 1$	All
Base	54.40%	38.46%	20.85%	26.40%

- Automatically created synsets vs manual corrections
  - Matching through multiple languages improves precision
  - Matching in multiple lexicons is also good
  - Different language families/lexicon groups would be better
- The base synsets are more ambiguous than average
  - Precision improves as we get more specific

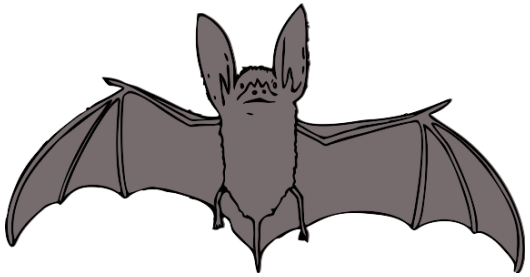
WordNet-Ja v0.9

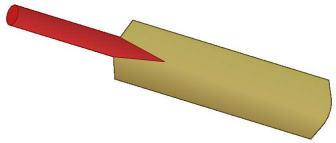
- Hand-checked Japanese words ($\approx 60,000$)
  added to English synsets ($\approx 20,000$)
- Plus high-confidence automatically translated words
  unambiguous translations of monosemous words
- No new synsets, assume Ja is similar to En
- Glosses not translated
- Large enough to be useful, good token coverage

Illustrations

- 2,000+ illustrations (959 synsets)
  - An illustration also illustrates its hypernyms
- From the Open ClipArt Library (public domain)
- SVG images (include metadata)
- Disambiguate using metadata

Illustration Example





dir	`animals/mammals/`	`recreation/sports/`
basename	`bat_orlando_karam`	`cricket_bat`
title	bat	Cricket Bat
tags	bat, mammal, animal	sports, cricket, recreation
synset	bat#n#1	cricket_bat#n#1, (bat#n#4)
Ja	蝙蝠	バット
match	hypernym	monosemous
	bat $\subset$ mammal	cricket bat

ToDo: Japanese WordNet

- Sense tag more corpora
- Add Japanese-specific synsets
  - Japanese specific concepts like *hakama*, *umami*, . . .
- Link orthographic variants
- Use WordNet-ja as the backbone for a real world knowledge base
  - Link to automatically created knowledge bases
- Cooperate with other WordNets (Thai, . . .) (**B5-3**)

Conclusions

- We have exploited existing WordNets to efficiently build a Japanese WordNet
- We will release the first version in June
  - Release early, so it can be used
    - Imperfect is more useful than non-existent
  - Accept feedback for the next version
    - Multiword Equivalents (with Kyoto University)
    - Verb Lexical Conceptual Structure (with NAIST)
  - Community maintenance (Kui, MLSN, ?)

Bibliography

References

Yoshihiko Hayashi. Translating WordNet noun part into Japanese for cross-language natural language applications. In *Technical Reports of SIG on Natural Language Processing NL130-10*, pages 73–80, 1999. (in Japanese).

Hiroyuki Kaji and Mariko Watanabe. Automatic construction of Japanese WordNet. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy, May 2006. URL `http://www.sdjt.si/bib/lrec06/summaries/439.html`.

Seiji Koide, Takeshi Morita, Takahira Yamaguchi, Hendry Muljadi, and Hideaki Takeda. OWL expressions on WordNet and EDR. In *AI society Semantic Web Ontology SIG 13*, SIG-SWO-A601-03, 2006. URL `http:`

//www.jaist.ac.jp/ks/labs/kbs-lab/sig-swo/fpapers.htm. (in Japanese).