

Toward an Integrated Semantic Framework for Lexical and Structural Semantics

Francis Bond

and [many more](#)

Linguistics and Multilingual Studies,
Nanyang Technological University

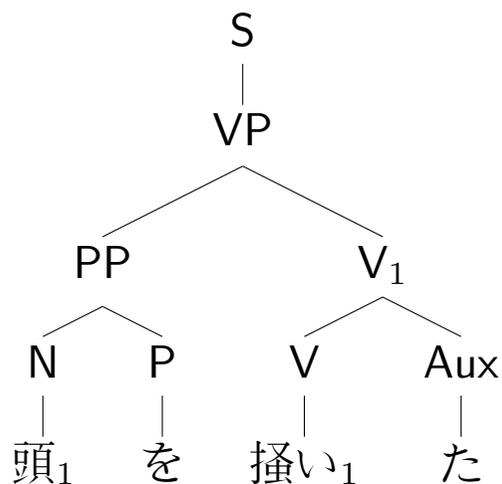
`<bond@ieee.org>`

Overview

- We are developing **rich multilingual meaning representations**
- Currently mainly sense annotation, about to start treebanking
- Goals
 - Scientific inquiry into how languages differ
 - Speeding up development of non-English by comparing analyses to English
 - Reference corpus for our Integrated Semantic Framework (MRS+WN+ α)
- **Spiral model**: annotate; improve model; re-annotate (update);
...
wordnet needs one or two more cycles of revision

Rich Representation

- (1) 頭 を 掻いた
 atama wo kaita
 head ACC scratched
 “I scratched **my** head.”



Syntax

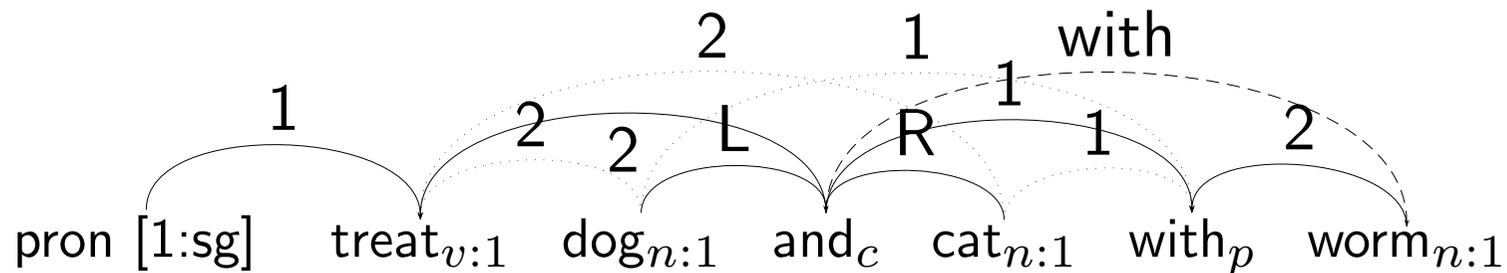
<i>atama</i> _{n:1}	is-a	bodypart
<i>kaku</i> _{v:1}	is-a	change
<i>kaku</i> _{v:1}	ARG1	Speaker
<i>kaku</i> _{v:1}	ARG2	<i>atama</i> _{n:1}
<i>kaku</i> _{v:1}	TENSE	past
Speaker	POSS	<i>atama</i> _{n:1}

Semantics

Extended features on the DMRS

I treat dogs and cats with worms.

HPSG grammars give us dependencies between (contentful) predicates (DMRS) and some information about MWEs.



Dashed lines show Preposition (P) features

Dotted lines show Conjunction (LR) features

Arc labels show the roles: 1 is ARG1, 2 is ARG2,

- Trivial for single words

$\text{cat}_{n:1} \rightarrow \mathbf{cat}_{n:1}$

- Harder for MWEs

not always the same choice in the two systems
and the coverage is patchy in both

- ERG sometimes decomposes, wordnet doesn't

$\text{here} \rightarrow \text{in}_p \text{DEF}_q \text{here}_a \text{place}_n$

$\text{here} \rightarrow \text{in}_p \text{this}_q \text{place}_n$; no need for ADV

- We have added decomposed pronouns

$\text{here}_{n:1}$ has-hypernym $\text{location}_{n:1}$; quantified-by $\text{this}_{a:1}$

$\text{this}_{a:1}$ has-hypernym $\text{proximal}_{a:1}$



Mapping Type	#	%	ERG	WN
unknown no match	48	0.3	comedians/nns	comedian
MWE	114	0.7	a+little	a_little
unknown match	136	0.9	flannel/nn	flannel
morphy	239	1.6	animate	animated
lemma+sense	274	1.8	look_v_like	look_like
ADJ+ly-ADV	405	2.6	usual	usually
mismatch	636	4.1	foul	foul-smelling
exact (ignore sense)	3,603	23.4	story_n_of	story
exact	9,948	64.6	depravity	depravity
Total	15,403	100		

➤ Not trivial to match lemmas (6% not matched at all)

Mismatches: A long and lovely tail

take	v	of-i	take_advantage
rest	v	1	rest_on
step	n	1	steps
join	v	1	join_forces
hold	v	1	hold_out
come	v	1	come_off
well	x	deg	well-kept
troop	n	1	troops
stair	n	1	stairs
fasten	v	cause-to	unfasten
grey	a	1	gray
moral	n	1	morals
let	v	go-of	let_go_of
late	a	for	later

Our main research focus

- create a representation that gets the best of ERG+WN
- tackle hard MWEs
 - *Linus coppers his bets.*
 - *Sherlock knocked her up.*
 - *They are few and far between.*
 - *They spilled the government beans.*
- Have a tool to parse automatically and rank to get the preferred reading: for Chinese, English, Indonesian and Japanese

Concrete next steps

- Compare treebank and sense annotation
 - *Cathedral and the Bazaar*
 - *Speckled Band*
 - *Dancing Man*

- Decide how to annotate senses:
 - Ambiguate MRS?
 - Map MRS to concept table and then tag?

- First attempt:
 - Text → DMRS → synsets (ntumc tagger) → tagged (UKB)

Known problems

- Mapping MWEs
 - compound nouns suddenly become hard :-)
- Solution(s)?
 - Add non-compositional MWEs to the ERG
hot-dog; knock up; copper one's bets
 - Match with MT rules
 - ?

How to do this?

- Assign MRS to all wordnet entries
parse (expecting fragments)
should parse the glosses/examples too, just for fun
- Find the **head** and see if it is a hypernym
- use the parse to make the LHS of the MT rule
- Do this in CEJI (at least)
encourage Bulgarian, Spanish, Norwegian, . . .
automate and cross link

- Consistent Granularity
 - ERG (DELPH-IN) aims to only differentiate when there is a syntactic difference.
 - We want to have the same distinctions made consistently
 - * Argument structure of nominals
 - * Cross-POS links in wordnet

- The ontology/SEM-I is part of the model
 - We don't need to replace *destruction* with *destroy*
It is linked already (derivation)
So we can sidestep the granularity consistency problem of AMR

- Cross lingual ISF matching (DMRS matching but with synset nodes)
gives us MT rules; multilingual treebanking; MT
same technique can be used for symptom/guideline matching
- Various paraphrases
 - The literalizer: replace idioms/MWEs with literal paraphrases
She knocked Holmes up → *She woke Holmes*
 - The generalizer: replace concepts with hypernyms
She knocked Holmes up → *A woman changed the state of the fictional character* → *An entity changed an entity*
have a slider to change the depth

- The Joke generator
 - *When is a job not a job?*
When it is a nose job.

Differences with AMR

- We have a pipeline to build them automatically although the mapping needs to be refined
- Some expectation that our semantics will be more consistent under the assumption that the ERG/Wordnet are consistent
- We have less money but more existing work in multiple languages
- We don't do (yet)
 - Semantic Role Labeling
 - Named Entities (although if we use Babelnet we can)
 - Co-reference resolution
- We need to make time to do the shared task
annotate the same text with different representations

➤ How to represent idioms?

(2) *bite one's tongue* “refrain from speaking”

(3) *bite the dust* “die”

➤ What is the correct syntactic representation?

pretty much the same as non-idiomatic — different lexical items

➤ What is the correct wordnet representation

bite the dust is in the *die*_{v:1} synset?

bite one's tongue is a hyponym of *silent*_{a:2} (or see-also)?

➤ What about ISF?

one predicate for *bite the dust*

multiple for *bite one's tongue*? = “remain silent”

bite one's foolish tongue need somewhere to link it

Superposition

- Have a packed structure in which both parts are simultaneously true

bite _{$v:i$} –*bite*–*ones*–*tongue* *tongue* _{$n:i$} –*bite*–*ones*–*tongue*

\wedge *keep* _{$v:1$} *silent* _{$n:i1$}

mark them as being in the same group

- For e.g. generation, make sure you only generate from one

- The same predicate may take part in two deep things

Poisson and Gaussian distributions

Poisson distribution _{$n:1$} and *Gaussian distribution* _{$n:1$}

Model multiple languages

- to be able to make knowledge available in any language
 - machine translation
 - cross-lingual information retrieval
- to exploit translations to bootstrap learning
 - translation sets can pinpoint concepts
 - translations can disambiguate structure
 - different languages pick out different things

Model a text and its translation(s)

- When do translations differ (translation shift)?
- How do we measure it?
- Resources
- Results
- Discussion
- Future work

Translation Shift

- **Transposition:** syntactic change
scared A → びびる *bibiru* “feel frightened” V.
- **Modulation:** semantic change
thumb → 指 *finger* “finger”
- **Equivalence:** different expression, but the meaning is still apparent
have no umbrella → 傘がない *kasa ga nai* “not have umbrella”
- **Adaptation:** change of situation due to disparities in culture
all one’s Christmases come at once →
お盆とお正月がいつぺんに来る *obon-to oshougatu-ga ippenni kuru* “Summer Festival and New Year come together”
- **Loose Translation:** possibly unmotivated change

Little Quantitative Study

- The amount of translation shift determines the difficulty of translation
- What kinds of phenomena occur (and why) are studied in Translation Studies
 - Often with fine grained analysis
- Strategies for translating developed in Machine Translation
- Which phenomena are more common and why?
 - Depends on the language pair and genre

- Mark the meanings of open class words
 - Tag them with senses from wordnet
 - Plus pronouns and interrogatives
- Link them between the languages
 - Add new entries to wordnet as needed
 - Text, ontology and grammar are all linked
- Categorize the unlinked concepts
- Eventually link this to full semantic representations (MRS)

Example (from News Corpus)

(4) Jpn: 大臣₁ が 離党₂ した
daijin ga ritou shita
minister SBJ leave-party did

(5) Eng: *The minister₄ left₈ the party₁*

(6) Cmn: 官员₁ 离开₃ 了 政党₁
guanyuan likai le zhengdang
minister leave already political-party

How are meanings linked?

Type	Example
=	same concept <i>say</i> ↔ 言う <i>iu</i> “say”
⊃	hypernym <i>wash</i> ↔ 洗い落とす <i>araiotosu</i> “wash out”
⊃ ²	2nd level <i>dog</i> ↔ 動物 <i>doubutsu</i> “animal”
⊂	hyponym <i>sunlight</i> ↔ 光 <i>hikari</i> “light”
⊂ ⁿ	nth level
~	similar <i>notebook</i> ↔ メモ帳 <i>memochou</i> “notepad” <i>dull_a</i> ↔ くすむ <i>kusumu</i> “darken”
≈	equivalent <i>be <u>content</u> with my word</i> ↔ わたくしの言葉を信じ-て “ <u>believe</u> in my words”
!	antonym <i>hot</i> ↔ 寒く=ない <i>samu=ku nai</i> “not cold”
#	weak ant. <i>not propose to <u>invest</u></i> ↔ <u>思いとどまる</u> <i>omoi=todomaru</i> “hold back”

Genre	Text	Sentences				Words	Concepts
		Eng	Cmn	Jpn	Ind	Eng	Eng
Story	Dancing Men	599	606	698	—	11,200	5,300
	Speckled Band	599	612	702	—	10,600	4,700
Essay	Cathedral and the Bazaar	769	750	773	—	18,700	8,800
News	Mainichi News	2,138	2,138	2,138	—	55,000	23,200
Tourism	Your Singapore (web site)	2,988	2,332	2,723	2,197	74,300	32,600

- All redistributable (except Mainichi: the WSJ of Japan)
- All fun to read (except Mainichi)
- Many translations exist (mainly public domain)
- Different genres

- Corpus: *The Adventure of the Dancing Men*
 - English source, Chinese and Japanese translations all public domain
 - Has both dialogue and narrative
 - Widely studied

- Lexicons
 - English Wordnet (Fellbaum, 1998)
 - Chinese Open Wordnet (Wang and Bond, 2013)
 - Japanese Wordnet (Isahara et al., 2008)

Dancing Men

	English	Chinese	Japanese
Sentences	599	680	698
Words	11,198	11,325	13,483
Concepts	6,842	5,148	5,246

POS tagged, segmented and aligned as part of the NTU Multilingual Corpus.

Wordnets

Language	Synsets	Words	Senses
English	117,659	155,287	206,941
Japanese	57,238	93,834	158,058
Chinese	111,045	115,136	168,824

- English is by far the most mature
- Japanese has more coverage of common words
- Chinese has more coverage of concepts

Annotation

- Monolingual annotation already done for each language although OK to do automatically
- Automatically match synonym, hypernym and hyponym ($=, \supset, \subset$)
- Link remaining concepts by hand (if possible) around 4 person-weeks/pair (30 sentences/day)
- Extend the wordnet/monolingual annotation as necessary
- Single annotator for each pair (Eng-Jpn, Eng-Cmn); NTU undergraduate with monolingual annotation experience

Analysis of links

Type	Eng-Jpn		Eng-Cmn	
linked	2,542		2,535	
=	1,416	51.58	1,712	60.07
~	990	36.07	862	30.25
≈	186	6.78	128	4.49
⊃	75	2.73	94	3.30
⊃ ²	8	0.81	13	1.51
⊂	63	2.30	39	1.37
⊂ ²	10	1.01	18	2.09
!	1	0.04	2	0.07
#	14	0.51	13	0.46
unlinked	2,583		1,898	

Analysis of ~

Type		Eng-Jpn		Eng-Cmn
Pronomilisation	0	0.00	7	0.81
Depronominatisation	86	8.69	22	2.55
Holonymy	12	1.12	0	0.00
Derivation	56	5.66	30	3.48

➤ We can find these automatically using wordnet relations

- 67% and 72% have the same part of speech

- Eng-Jpn:
 - 7.9% adj-noun
 - 7.4% verb-noun

- Eng-Cmn:
 - 7.3% noun-verb
 - 3.9% noun-adj

(7) Said he suddenly

- a. ホームズが 突然 □ を 開く
ho-muzu ga totsuzen kuchi wo hiraku
Holmes NOM suddenly mouth ACC open
Holmes opens his mouth suddenly

- *kuchi wo hiraku* is lexicalized but not (yet) in wordnet
- or in **Jacy** (and should it be?)

(8) I gave a start of astonishment.

a. 私 は 驚き の あまり 身
watashi wa odoroki no amari mi
1SG NOM astonishment POSS much body
を 震わせた
wo furuwaseta
ACC shook

I shook my body (due to) much astonishment

- *give a start* is lexicalized but not (yet) in wordnet
(*start* is: *wake with a start*)
- 身を震わせる is lexicalized but not (yet) in wordnet



(9) get to the bottom of it

a. 暴く こと が できます
abaku koto ga deki-masu

expose NMLZ NOM can-POL

able to expose

b. 彻底 弄 清楚
chèdǐ nòng qīngchǔ

completely make clear

to make clear completely



(10) sift the matter to the bottom

a. 最後まで調べたい

saigo made shirabe-tai

end until investigate-want

"want to investigate until the end"

b. 彻底 弄清楚

chèdǐ nòng qīngchǔ

completely make clear

"to make clear completely"

➤ *sift the matter/get to the bottom* → chèdǐ nòng qīngchǔ

➤ not a direct translation: how can we represent this?

Decomposable Predicates

(11) his long, thin back curved over

- a. 他 弯 着 瘦长 的 身子
tā wān zhe shòucháng de shēnzi
3SG curve PROG lanky de body
“he curved (his) lanky body”

- *lanky* “tall and thin” (wn)
- *shòucháng* lit: thin+**tall**
- We should link these somehow in wordnet

(12) She_i shot him_j and then herself_i

a. 奥-さん が 旦那-さん を 撃って

oku-san ga danna-san wo utte

wife-HON NOM husband-HON ACC shoot-CONJ

、それから 自分 も 撃った

, sorekara jibun mo utta

, and+then self too shoo-PST

Wife_i shot husband_j and then shot self_i too

(13) She_i shot him_j and then herself_i

- a. 她 拿 枪 先 打 丈夫 , 然后
tā ná qiāng xiān dǎ zhàngfū , ránhòu
3SG take gun first shoot husband , and+then
打 自己
dǎ zìjǐ
shoot self

She_i took the gun to first shoot husband_j, and then shot self_i

Not linkable with our current model

(14) I am sure that I shall say nothing of the kind.

- a. いやいや 、 そんな ことは
iyaiya , sonna koto wa
by+no+means , that+kind+of thing TOP
言わ-ん よ
iwa-n yo
say-NEG yo
“no no, I will not say that kind of thing”

- *sonna* in our wordnet & negation makes it hard to link
- *iyaiya* ↔ *I am sure that I shall ???*
- Decomposing pronouns gives us a lot of this, but the equivalence requires some inference

This too

(15) Now, Watson, confess yourself utterly taken aback, said he.

(16) I am

a. まったく だ。

mattaku da

absolutely COP

Absolutely

➤ Perfect in context

➤ We don't model the discourse at all

- Still many predicates not matched
 - we need more general matching
 - the wordnets are missing many idiomatic expressions
 - translations are not always faithful to the original
- Wordnet structure enables automatic links
hypernym, meronym, derivation, . . .
- But there are interesting gaps in wordnet's representation
 - Negation
 - MWEs/Phrases
 - Decomposable predicates
- The HPSGs are helpful here

Conclusions

- We have annotated 600 sentences in three languages
 - Only 27-40% of predicates translated directly
 - Many small shifts
 - Many large shifts
- Wordnets are missing many MWEs (maybe as many as 80%)
- We do not handle some common relations
 - decomposable meaning
 - negation
 - flexible idioms

Ongoing Work

- Add missing entries to the wordnets
- Improve the automatic annotation
 - link nth level hypernyms; link derivations
 - link pronouns and interrogatives
- Improve the annotation tool
- Tag and release more text: Essay, News, Tourism
(Funding for 6,000 sentences (CEJ) + 2,000 Indonesian)
- Use the data to improve machine translation
- This is Open Data: Anyone can build on this (not quite out yet)

Please Join in

- Planning to add Spanish, German, Russian, Vietnamese, . . .
- Coordinating with wordnet projects
- Will use the data to add sense-frequencies for wordnets
- Annotating *Dancing Men* in a new language is a perfect size for an undergraduate thesis
 - We hope to make our software available to do this
- Actually shifting to *Speckled Band* (less meta-text)
 - have tagged all sentences with three and checked by me
 - potential joint text with AMR, Meaning Bank, . . .

Two more things

- Discuss annotation in teaching
- Show off the new OMWx web site

Thanks and more

- We would like to thank:
 - The Creative Commons Catalyst Grant: *Assessing the effect of license choice on the use of lexical resources*
 - The JSPS-NTU grant: *Revealing Meaning Using Multiple Languages*
 - The NTU Tier 1 grant: *Shifted in Translation*
 - The MOE Tier 2 grant: *That's what you meant: A Rich Representation for Representing Meaning*
 - NTU URECA projects
 - HG2002 students



References

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

Isahara, H., Bond, F., Uchimoto, K., Utiyama, M., and Kanzaki, K. (2008). Development of the Japanese WordNet. In *Sixth International conference on Language Resources and Evaluation (LREC 2008)*, Marrakech.

Pozen, Z. (2013). Using lexical and compositional semantics to

improve HPSG parse selection. Master's thesis, University of Washington.

Wang, S. and Bond, F. (2013). Building the Chinese Open Wordnet (COW): Starting from core synsets. In *Proceedings of the 11th Workshop on Asian Language Resources, a Workshop at IJCNLP-2013*, pages 10–18, Nagoya.