NTU Multi-lingual Corpus and cross-lingual fun*

Francis Bond, Shan Wang,
Eshley Huini Gao, Hazel Shuwen Mok, Jeanette Yiwen Tan
and many more
Linguistics and Multilingual Studies,
Nanyang Technological University
<bond@ieee.org>

[*] See also *Developing Parallel Sense-tagged Corpora with Wordnets*, (Bond et al., 2013)

DELPH-IN 2014, Tomar

Overview

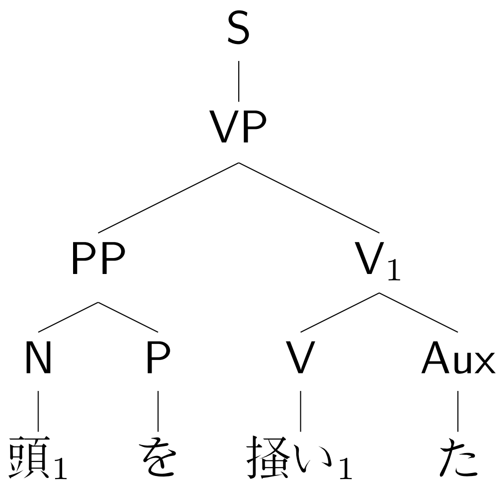
- We are developing multilingual texts
- Currently mainly sense annotation, about to start treebanking
- Goals
  - Scientific inquiry into how languages differ
  - Speeding up development of non-English by comparing analyses to English
  - Reference corpus for our Integrated Semantic Framework (MRS+WN+α)

Rich Representation

(1)

頭	を	掻いた
atama	wo	kaita
head	ACC	scratched

"I scratched my head."



Syntax

atama$_1$	is-a	bodypart
kaku$_1$	is-a	change
kaku$_1$	ARG1	Speaker
kaku$_1$	ARG2	atama$_1$
kaku$_1$	TENSE	past
Speaker	POSS	atama$_1$

Semantics

Wordnets and HPSG grammars assumed; Pragmatics yet to come

Why 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 月がいっぺんに来る *obon-to oshougatu-ga ippen-ni 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

Measure using sense tagged corpora

- 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)

Still working out what we need to represent.

Example (from News Corpus)

(2) Jpn:

大臣$_1$	が	離党$_?$	した
daijin	ga	ritou	shita
minister	SBJ	leave-party	did

(3) Eng: *The minister$_4$ left$_8$ the party$_1$*

(4) Cmn:

官员$_1$	离开$_3$	了	政党$_1$
guanyuan	likai	le	zhengdang
minister	leave	already	political-party

From the NICT multilingual corpus: licensing murky

How are meanings linked?

	Type	Example
$=$	same concept	*say* ↔言う *iu* “say”
$\supset$	hypernym	*wash* ↔洗い落とす *araiotosu* “wash out”
$\supset^2$	2nd level	*dog* ↔ 動物 *doubutsu* “animal”
$\subset$	hyponym	*sunlight* ↔光 *hikari* “light”
$\subset^n$	nth level	
$\sim$	similar	*notebook* ↔ メモ帳 *memochou* “notepad” $dull_a$ ↔くすむ *kusumu* “darken”
$\approx$	equivalent	*be content with my word* ↔ わたくし の 言葉 を 信じ-て “believe in my words”
!	antonym	*hot* ↔寒く=ない *samu=ku nai* “not cold”
#	weak ant.	*not propose to invest* ↔ 思いとどまる *omoi=todomaru* “hold back”

NTU Multilingual Corpus

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

Pilot Study

- 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 Wordnet (Xu et al., 2008)
  - Japanese Wordnet (Isahara et al., 2008)

A Sherlock Holmes short story by Arthur Conan Doyle

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

Long and difficult: done by NTU interns at NICT!

Analysis of links

Type	Eng-Jpn		Eng-Cmn	
linked	2,542		2,535	
$=$	1,416	51.58	1,712	60.07
$\sim$	990	36.07	862	30.25
$\approx$	186	6.78	128	4.49
$\supset$	75	2.73	94	3.30
$\supset^2$	8	0.81	13	1.51
$\subset$	63	2.30	39	1.37
$\subset^2$	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 $\sim$

Type	Eng-Jpn		Eng-Cmn	
Pronomilisation	0	0.00	7	0.81
Depronominalisation	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

Analysis POS mismatches (non $=$)

- 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

Idioms

(5) Said he suddenly

a.

ホームズ	が	突然	口	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?)

(6) 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

(7) 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

(8) 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

(9) 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

Pronomilization

(10) 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

Pronomilization

(11) 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

(12) 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*** not in 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

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

(14) I am

a. まったく だ。
mattaku da
absolutely COP
Absolutely

- Perfect in context

- We don’t model the discourse at all

Discussion

- 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, . . .

Thanks and more

- We would like to thank:
  - Spinoza (Piek) for bringing me to Europe
  - 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*
  - NTU URECA projects
  - HG2002 students

*

References

Bond, F., Wang, S., Gao, E. H., Mok, H. S., and Tan, J. Y. (2013). Developing parallel sense-tagged corpora with wordnets. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse (LAW 2013)*, pages 149–158, Sofia.

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.

Xu, R., Gao, Z., Qu, Y., and Huang, Z. (2008). An integrated approach for automatic construction of bilingual Chinese-English WordNet. In *3rd Asian Semantic Web Conference (ASWC 2008)*, pages 302–341.

Matching to external resources

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)	3603	23.4	story_n_of	story
exact	9948	64.6	depravity	depravity
Total	15403	100		

➢ Not trivial to match lemmas

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