HG3051 Corpus Linquistics

Case Studies: Pronouns and Classifiers

Francis Bond
Division of Linguistics and Multilingual Studies
http://www3.ntu.edu.sg/home/fcbond/
bond@ieee.org

Lecture 8

https://github.com/bond-lab/Corpus-Linguistics

Overview

- Revision of Lexical, Morphological and Syntactic Studies
  - Lexical Studies
  - Grammatical Studies
  - Variation
- Case Studies
  - Pronouns
  - Classifiers

Revision of Lexical, Morphological and Syntactic Studies

Corpus Studies of Lexicography

Discussion ***big*, *large*, *great***

- ***big*** mainly for concrete things
- ***large*** mainly for amounts and numbers
- ***great*** similar to ***large*** but many special senses
  - ***great deal***
  - ***great man***
  - ***great burrow***
  - ***great*** `relative`

  also use as intensifier *great big, great importance*

Corpus Studies of Morphology

Discussion

➢ ***-[ts]ion*** more common in Academic (but common everywhere)
basic use is to make an action non-agentive

  ➢ *It provides a direct indication of fuel consumption.*

➢ ***-ment*** often used for mental states
agreement, amazement, embarrassment (Fiction)

  ➢ *Patrick shrugged in embarrassment.*

➢ ***-ness*** used for personal qualities
bitterness, happiness, politeness (Fiction)

  ➢ *The bitterness in his heart was mixed with ....*

It would be good if we could automatically divide the words according to their semantic field (which we can approximate with WordNet, ...)

Corpus Studies of Syntax

Discussion

Typically ***start*** is used to show the onset of a process, often with an adverb

- *The soil formation process may <u>start</u> again in the fresh material*
- *The train <u>started</u> down the hill*

begin is used with more concrete agents

- *Then I <u>began</u> to laugh a bit.*
- *The original mass of gas cooled and <u>began</u> to contract.*

Because the corpus doesn't mark **animacy** or **concrete agent** these statements are weak: we can't really make predictions or measure correlation.

little vs ***small***: Interpretation

- Attributive much more common for both
  - Predicative relatively more common in conversation
  - Predicative relatively more common for ***small*** than ***little***
- Collocation results:
  - ***little***: concrete objects (*little boy*)
  - ***small***: amounts (*small proportion*)
- But predicative ***small*** also for physical size:
  - *She's small and really skinny*
  - *He's really small isn't he?*
- We still don't really know why ☹
  corpus linguistics gives us the what, but not the why

Where do we go from here?

- Corpora show clearly that even very similar words can show different behavior.
- But they still don't explain why
  - Hand correction limits data sizes
  - Without semantic tags, we can't generalize automatically
- Corpora with more mark-up (syntax and semantics) would help
  - But they are expensive, ...

Case Study: Pronouns

Possessive Pronouns in Japanese contrasted with English

- Introduction
- Possessive Expressions in Japanese and English

(1)

Kanji:			
Jap:	*watashi-wa*	*shita-wo*	*kanda*
Gloss:	I-TOP	tongue-ACC	bit
Eng:	*'I bit my tongue'*		

- Differences in Noun Phrase Structure
- Pragmatic Analysis
- Application to Machine Translation
  Proposed method for generating possessive pronouns
  Experimental Results
- Conclusion

Introduction

Possessive expressions

Possessive determinatives are often used as determiners in English when no equivalent would be used in a Japanese sentence with the same meaning.

Larger Problem

Japanese has no syntactic equivalent to determiners in English, no articles, and noun phrases are normally not marked for number.

Under-specified elements need to be deduced!

Corpus-based Study of Distribution

Type:		MT Test set		News reports	
		No.	%	No.	%
I	English Idiomatic Possessive	105	16%	35	19%
II	Possessive Expression in Japanese	193	30%	5	3%
III	No Possessive in Japanese	359	54%	176	78%
Total:		657		181	

- Two Corpora
  - NTT MT Test set (6,200 sentences, 15,000 NPs)
  - Nikkei News Reports (1,382 sentences, 8000 NPs)
- Matched English:
  `[Mm]y|[Yy]our|[Hh]is|[Hh]er|[Ii]ts|[Tt]heir|[Oo]ur`
  Then hand checked Japanese for translation (on paper with colored pens!)

Examples

Type I: English Idiomatic Possessive (16%–19%)

(2)

Kanji:			
Jap:	*kanojo-wa*	*chie-wo*	*shibotta*
Gloss:	she-TOP	knowledge-ACC	squeezed
Eng:	*‘She racked her brains’*		

Type II: Possessive expression in Japanese (30%–3%)

(3) MT test set is not a corpus of natural text

Kanji:				
Jap:	*kanojo-wa*	*kare-no*	*kao-wo*	*mita*
Gloss:	she-TOP	he-ADN	face-ACC	saw
Eng:	*‘She saw his face’*			

Type III: No possessive expression in Japanese (54%–78%)

(4)

Kanji:			
Jap:	*kanojo-wa*	*saifu-wo*	*nakushita*
Gloss:	she-TOP	wallet-ACC	lost
Eng:	*'She lost her wallet'*		

(5) *NTT*

NTT-wa	'menber-netto'-no	meesho-de	kotoshi	nigatsu-kara
NTT-TOP	'member-net'-ADN	name-by	this year	February-from

tsune-ni	sa–bisu-o	kaishi-shite-iru
already	service-ACC	start-is

"NTT began its VPN services in February."

Distribution of possessives in English

- Possessive determinatives used relatively frequently
  — *of* POSSESSIVE PRONOUN rare

- Generally not used after verbs of **possession** or **acquisition**, except for emphasis
  I have a car vs *I have my car*

- Typically referential use, not generic or ascriptive

In particular, words which denote **work, body parts, personal possessions, attributes** and relational nouns such as **kin** and **people defined by their relation to another person** (such as *assailant, subordinate*) are often modified by possessive determinatives in English.

Distribution of possessives in Japanese

- Normally only if ‘possessor’ is not subject

(5) *watashi-wa saifu-o otoshita*
I-TOP wallet-ACC dropped
I dropped my wallet

(6) *watashi-wa jibun-no saifu-o otoshita*
I-TOP self-ADN wallet-ACC dropped
I dropped my own wallet

(7) *watashi-wa kare-no saifu-o otoshita*
I-TOP he-ADN wallet-ACC dropped
I dropped his wallet

- Use of any pronouns is rare
  All 5 uses in the newspaper corpus are common nouns (pronominalized in translation)

Two examples:

(8)

indoneshia-TOP *3.5x10*^{11}*-doru*-ADN *shikin*-ACC
Indonesia 3.5x10^{11}-dollars capital
infura-seibi-toshite *tounyuu-suru* *keikaku-da*
infrastructure-preparation-as invest-do plan-is

"*Indonesia*$_i$ *is planning to invest 300.5 billion dollars to expand its*$_i$ *infrastructure*"

(9)

mubaraku-daitouryou-ADN	*rainichi-ji-ni*	*hyoumei-suru*	*kangae-da*
President-Mubarak	japan-visit-time-in	convey-do	thought-is

“*the decision will be conveyed to President Muhammad Hosni Mubarak*$_i$ *during his*$_i$ *visit to Tokyo*”

English NP Structure

1. NP → Det (Mod)* Noun (Det is specifier)
2. Possessive determinative functions as central determiner
3. Unique
4. Contrasts with a closed set (+ integers):

   articles ZERO, *a/an, some, the*, NULL
   possessive phrases e.g. *the man's*
   demonstratives *this, these, that, those*
   pronouns *we, you, us*
   quantifiers *each, enough, much, more, most, less, a few, a little* ...
   wh-words *which, what* (interrogative or relative)
   determinatives *some, any, no, either, neither, another*

NP's headed by count nouns must have an article

Japanese NP Structure

1. NP → (Mod)* Noun (no specifier)
2. Possessive expression functions as modifier
3. Can be multiple modifiers: (rare)

(10) *watashi-no kono hon*
me-ADN this book
Lit: "*my this book*"

4. Is a member of an open set, including:

none (most common)
genitive noun phrases *Tarou-no* "Taro's", *nihon-no* "Japanese" ...
demonstratives *kono* "this", *sono* "that", *that over there* "ano"
quantifiers *koko-no* "each", *kaku* "each" ...
wh-words *dono* "which, what" ...

Analysis

Explain the differences with Grice's Conversational Maxims.

- *The Maximum of Quantity*:
  (i) make your contribution as informative as is required for the current purposes of the exchange
  (ii) do not make your contribution more informative than is required
- *The Maximum of Relevance*:
  Make your contributions relevant

The kind of information encoded by determinatives such as quantifiers and demonstratives is generally encoded in both Japanese and English. The Maxim of Relevance requires its presence if relevant.

English:

1. Possessive determinative contrasts with articles
   — equivalent effort

2. Use of indefinite article implicates not owned
   — unless ‘possession’ predicated by verb

3. Use of definite article implicates more restricted reference

4. ⇒ Use possessive determinative if relevant
   — unless ‘possession’ predicated by verb
   (don’t be more informative than is required)

Japanese:

1. Possessive expression requires extra effort

2. Don't use by default
   — interpretation is that subject is antecedent

3. ⇒ Use possessive expression to contradict default

4. ⇒ Use possessive expression to emphasize default

A complicated example

The word *keijourieki* “pretax profit” appeared 29 times. In Japanese it was only pre-modified by time expressions (12 times).

The English equivalents were more varied:

Det	Freq	Comment
ϕ	12	Prepositional phrase
ϕ	4	Direct Object (3 x *post*, 1 x *expect*)
ϕ	4	Subject
its	1	Subject
its	4	COMPANY *said/announced that its …*
A	1	*A one billion yen pretax profit*
both	1	very free translation
Toyobo's	1	Subject (Toyobo from other sentence)
their	1	Direct Object of (*post*) Subject is many companies

A complicated example (cont)

(11) $\textsc{Company}_i$ announced Wednesday it$_i$ has posted ϕ_i pretax profits of ...

(12) $\textsc{Company}_i$ announced Tuesday that its$_i$ pretax profit rose ...

(13) $\textsc{Company}$'s 11 [...] subsidiaries$_i$ are expected to post their$_i$ first-ever combined pretax profits of ...

(14) $\textsc{Company}_i$ will post a rise of 6% in ϕ_i pretax profits ...

(15) $\textsc{Company}_i$ will post 28 billion yen in ϕ_i pretax profits ...

The direct object of *post* implies 'possession' by its subject, the direct object of *announce* doesn't. But what about the PPs?

Should we put this in the lexicon?

Application to Machine Translation

- Mark nouns that head English noun phrases with possessive determinatives where there is no possessive expression in the Japanese in the lexicon (possessed-nouns)
  - 205 different possessed-nouns (MT test set)
  - heading 825 noun phrases
  - 359 (44%) translated with possessive pronouns
- Mainly nouns that denote **`kin, body parts, work, personal possessions, attributes`** and **`people defined by their relation to another person`**
- Which nouns need to be marked is language specific, and probably register and domain specific as well.

Translating NPs headed by possessed-nouns

1. A noun phrase that fulfills all of the following conditions will be generated with a default possessive determinative with deictic reference determined by the modality of the sentence it appears in*.

   (a) The noun phrase is headed by a possessed-noun that denotes `kin` or `body parts`
   (b) The noun phrase is the subject of the sentence
   (c) The noun phrase is referential
   (d) The noun phrase has no other determiner

*First person for declarative, second person for imperative or interrogative.

2. A noun phrase that fulfills all of the following conditions will be generated with a default possessive determinative whose antecedent is the subject of the sentence the noun phrase appears in.

   (a) The noun phrase is headed by a possessed-noun
   (b) The noun phrase is not the subject of the sentence
   (c) The noun phrase is referential
   (d) The noun phrase has no other determiner
   (e) The noun phrase is not the direct object of a verb of `possession` or `acquisition`

Effects of noun phrase referentiality

Only for Referential NPs:

(16)	Kanji:		
	Jap:	*hana-ga*	*kayui*
	Gloss:	nose-NOM	itch
	Eng:	‘My nose itches’	
	MT-93	`A nose itches`	
	MT-94	`My nose itches`	

Not for Generic NPs:

(17) Kanji:

Jap:	*hana-wa*	*kankakukikan*	*da*
Gloss:	nose-TOP	sensory organ	is

Eng: ‘The nose is a sensory organ’

MT-93: `A nose is a sensory organ`

MT-94: ϕ `Noses are sensory organs`

Restrictions from verbs

- If a noun phrase headed by a possessed-noun is the direct object of a verb of **possession** or **acquisition** then do not generate a possessive pronoun.

(18)	Kanji:		
	Jap:	*kuruma-wo*	*motteimasu-ka*
	Gloss:	car-OBJ	have-Q
	Eng:	‘Do you have a car?’	
	MT-93:	`Do you have a car?`	
	MT-94′:	`Do you have your car?`	
	MT-94:	`Do you have a car?`	

Experimental Results

Results of the generation of all noun phrases headed by possessed-nouns in the MT test set (Total 752 noun phrases).

Result	Not generated	Generated
Good	I hit him in the face	I hid my face
Bad	I scratched a face	I lost my face

Result	Possessive	MT-93		MT-94	
	determinative	NPs	%	NPs	%
Good	Not generated	429	57%	346	46%
	Generated	0	0%	263	35%
	— Total	429	57%	609	81%
Bad	Not generated	323	43%	60	8%
	Generated	0	0%	83	11%
	— Total	323	43%	143	19%

Over All Results

323 NPs required possessive determinatives

Appropriately generated: 263

Inappropriately generated: 83

	MT-93	MT-94
Accuracy	57%	81%
Precision	—	88%

Improve accuracy by:

improving parsing and transfer stages

correctly identifying all possessed-nouns (use parsed aligned corpora)

Improve precision by:

improving determination of referentiality

add explicit semantic constraints:

only for possessed-nouns that denote `clothes` if the antecedent is **`human`**

Conclusions

1. Possessive determinatives are used in English even when there is no equivalent possessive expression used in Japanese

2. This can be explained by the fact that in English possessive determinatives function as determiners, while in Japanese the possessive construction is an optional modifier phrase

3. 'possessed-nouns' can be identified in English that act (imperfectly) as cues

4. Implementing an algorithm that uses possessed-nouns in the Japanese-to-English MT system **ALT-J/E** generated possessive pronouns with an accuracy of 81% (up from 57%) and precision of 88%.

5. Should also be applicable to other under-specified generation: AAC.

Gratuitous Discussion

1. Satoru Ikehara calls our approach meaning analysis as opposed to meaning understanding. We attempt to solve problems, even if not perfectly, by stepwise refinement.

2. Generally, a brute-force approach of adding information to the lexicon (which may mean checking 70,000+ common nouns ...) and adding new rules takes 3-6 months and gets an 80% solution.

3. I did this once for number/countability and articles (which took three years), then possessive pronouns, and then numeral classifiers.

4. By this stage, determiners and number were good enough that problems with prepositions and tense/aspect became more pressing.

5. The hope is that any work done will still be useful in the next version/refinement of the problem: this has proved to be the case so far.

Conclusions

1. Possessive pronouns are used in English even when there is no equivalent possessive expression used in Japanese

2. ‘possessed-nouns’ can be identified in English that act as cues

3. An algorithm is proposed that uses possessed-nouns to appropriately generate possessive pronouns in a Japanese-to-English MT system

4. Implementing the algorithm in **ALT-J/E** generated possessive pronouns with an accuracy of 81% (↑ 57%) and precision of 88%.

Annotation of Pronouns in a Multilingual Corpus of Mandarin Chinese, English and Japanese

Motivation and Overview

- Attempting to model lexical and structural semantics
  - For multiple languages — identify cross-lingual differences
  - Exploit them to learn meaning (make the implicit explicit)

- Started by annotating content words (with wordnets)

- But nouns were often translated as pronouns$_i$ — so tag them$_i$

1. Identify pronouns used in the corpus
2. Analyze in terms of components — aids matching
   - Extended wordnet gives full decompositional analysis
3. Annotate the pronouns monolingually in each language
   - Link to extended wordnet for analysis
4. Annotate their correspondences across languages
5. Analyze the distribution cross-lingually

Identifying Pronouns

- Examined words tagged as pronouns in (Mandarin) Chinese, English, Japanese (and later Indonesian) parts of the NTU Multilingual Corpus (NTU-MC) — used the POS tags
  - Different tag-sets identified quite different collections
- We took the union, and filled in missing entries by hand
  - also referred to reference grammars
  - not complete, but getting there
- 117 different types; 249 tokens:

Chinese	57	
English	68	
Indonesian	40	(in progress)
Japanese	84	

numbers out of date

- We include related determiners (demonstratives and quantifiers)

Components

Head	Person	Number	Gender	Case	Q/Type	Formality	Proximity
Quantifier	First	Dual	Feminine	Objective	Assertive	Formal	Proximal
Entity	First (I)	Plural	Masculine	Possessive	Elective	Informal	Distal
Time	First (E)	Singular	Neuter	Subjective	Negative		Medial
Manner	Second				Other	Politeness	Remote
Person	Third				Reciprocal	Polite	
Place					Universal		
Reason					Interrogative		
Thing					Reflexive		

Similative are treated as +Manner, +Proximity

Analyzing Pronouns Mono-lingually

- Decompose into:
  - **head** (HYPONYM)
  - **quantifier** (QUANTIFIER: new relation)
  - **features** (DOMAIN-USAGE)
- Also mark as INSTANCE of ***pronoun***$_{n:1}$ or its hyponyms
- E.g.

there$_{n:1}$:	HYPONYM	***location***$_{n:1}$;
	DOMAIN-USAGE	***distal***$_{a:1}$;
	INSTANCE	***demonstrative pronoun***$_{n:1}$

Components: Place

Head	Type/Proximity	English	Japanese	Chinese
Place	Interrogative	*where*	, *doko*	*nǎlǐ*
	Proximal	*here*	, *koko*	*zhèlǐ*
	Distal	*there*		*nàli*
	Medial		, *soko*	
	Remote		, *asoko*	
	Universal	*everywhere*	*doko mo*	*dàochù*
	Existential		*doko ka*	*mǒuchù*
	Assertive	*somewhere*		
	Elective	*anywhere*		
	Other	*elsewhere*	*yoso*	*biéchù*

Not all lemmas shown

Tagging Pronouns Mono-lingually

- Tagged one document by hand *The Adventure of the Speckled Band*

Language	English	Chinese	Japanese
Contentful	1,370	1,177	463
Other	75	19	51
Total	1,445	1,196	514
Sentences	599	620	702
Words	11,628	12,433	13,902

- Distinguished existential ***there*** (but not dummy ***it***) with POS tags

- *other* includes relative pronouns, dummy ***it***, idioms and segmentation errors

Tagging and Analyzing Pronouns Cross-lingually

- Automatically linked by matching features

- Hand corrected:

	Linked Pronouns						Non-linked Pronouns	
	# Matching Features					Pronoun	English	Other
	5	6	7	8	9	to Noun		
# Chinese	5	19	54	789	58	134	369	215
# Japanese	15	120	114	37	32	139	943	109

- Case and politeness mismatches common

- A surprising number of non-linked pronouns in Chinese and Japanese

Interesting Cross-Linguistics Differences

(19) She$_i$ shot him$_j$ and then herself$_i$

a. - -

oku-san ga danna-san wo utte , sorekara jibun mo utta

Wife$_i$ shot husband$_j$ and then shot self$_i$ too

b. ,

tā ná qiāng xiān dǎ zhàngfū , ránhòu dǎ zìjǐ

She$_i$ took the gun to first shoot husband$_j$, and then shot self$_i$

(20) [many (cases) strange] ...but none commonplace ...

a.

Dan4shi4 que4 mei2you3 yi1li4 shi4 ping2dan4wu2qi2 de

‘But, there is not one case that is featureless.’

b.

Dore mo jinjode wa nai jiken dearu

‘Everything is a case which is not usual.’

(21) It is a swamp adder!

a.

Zhe4 shi4 yi1tiao2 zhao3di4 kui2she2 !

‘This is a swamp adder!’

b.

numahebi da !

‘ϕ is a swamp snake’

Discussion

- A new way of annotation that links wordnets to corpora
- Unresolved issues (possible ideas for project 2)
  - Further analysis of unlinked pronouns: which and why?
    In particular how and why are Japanese and Chinese different?
  - Tag more corpora (ongoing); Extend to more languages;
  - Integrate to HPSGs: ERG, Jacy, MCG, IndoGram

Classifiers

How do we count Email in Japanese?

- Japanese has two classifiers for counting messages:
  - *tsuu*: used for letters
  - *ken*: used for incidents
  - *hon*: used for phone calls
- See how they are used to count Email and SMS
  - Look at a newspaper corpus Mainichi News (CD-ROM) 1996, 1998, 2000, 2002, 2004

From Asako Iida ”The current transition of Japanese numeratives for counting digital messages” (2006).

http://www5b.biglobe.ne.jp/~aiida/iida2006-message.pdf

Change with familiarity

Year	1996	1998	2000	1002	2004
Email Usage	5%	11%	34%	81%	86%
Classifier	,			,	,
SMS Usage	—	39%	45%	67%	76%
Classifier		,			,

Change in Classifier use with increased familiarity
Classifiers listed in frequency order

has more of a one-way feeling, while is more of a conversation.

Sometimes depends on the tool (which classifier does it use).

Conclusions

- Different questions require different resources
- Good corpora are useful for multiple tasks

More SQL

- Find animals

```
select * from synset where synset in
(select synset from xlink where resource='lexnames' and xref=5)
limit 125
```

- find sentences with animals

```
attach 'eng.db' as as 'e';
select sent from e.sent join e.concept
on e.sent.sid=e.concept.sid
where tag in (
select synset from synset where synset in
(select synset from xlink where resource='lexnames' and xref=5))
limit 50
```

- Another way (without duplicate sentences)

```
select sent from e.sent
where sid in
(select distinct sid from concept
where tag in
(select synset
from synset
where synset in
(select synset from xlink
where resource='lexnames' and xref=5)))
limit 50
```