Computational Modelling of Plurality and Definiteness in Chinese Noun Phrases

Yuqi Liu, **Guanyi Chen**, Kees van Deemter

Central China Normal University
Utrecht University

Overview

Background

Data

Computational Modelling

Overview

"Cold-Hot" Division of Languages

James Huang[1] distinguished languages as:

A language is
"Cool"
(e.g., Chinese, Japanese and Korean)
if understanding a sentence requires some work on the reader's or the hearer's part, which may involve inference, context, and knowledge of the world,

vs.

A language is
'Hot'
(e.g., English and French)
is "hot" if the information required to understand each sentence is largely obtainable from what is overtly seen and heard in it;

[1] Huang, "On the distribution and reference of empty pronouns".

Coolness: Anaphora

Speaker A: Did John see Bill yesterday?
(Chinese) 张三 看见 李四 了 吗?

(1)
a. Yes, he saw him
b. * Yes, *e* saw him
c. * Yes, he saw *e*
d. * Yes, *e* saw *e*
e. * Yes, I guess *e* saw *e*
f. * Yes, John said *e* saw *e*

(2)
a. 他 看见 他 了
b. *e* 看见 他 了
c. 他 看见 *e* 了
d. *e* 看见 *e* 了
e. 我 猜 *e* 看见 *e* 了
f. 张三 说 *e* 看见 *e* 了

Coolness: from Anaphora to other Categories[2]

(3) a. 狗 很 聪明 。

Dogs are intelligent.

b. 我 看到 狗 。

I saw a dog/dogs.

c. 狗 跑走 了 。

The dog(s) ran away.

[2] Auwera and Baoill, *Adverbial constructions in the languages of Europe*.

Coolness: two Levels

1. Grammar of cool languages allows the absence of these components;
2. (Comprehension) The meaning of these absences can be inferred from their contexts;
   (Production) A component can be dropped and is often dropped if its meaning can be inferred from its context.

Coolness: two Levels

1. Grammar of cool languages allows the absence of these components;
2. (Comprehension) The meaning of these absences can be inferred from their contexts;
   (Production) A component can be dropped and is often dropped if its meaning can be inferred from its context.

Coolness: two Levels

1. Grammar of cool languages allows the absence of these components;
2. (Comprehension) The meaning of these absences can be inferred from their contexts;
   (Production) A component can be dropped and is often dropped if its meaning can be inferred from its context.

This study

1. Grammar of cool languages allows the absence of these components;
2. (Comprehension) The meaning of these absences can be inferred from their contexts;
   (Production) A component can be dropped and is often dropped if its meaning can be inferred from its context.

Overview

Overview

Idea: Automatic Construction from a parallel corpus

- It is hard to annotate plurality and definiteness of Chinese NPs and construct a large-scale dataset;
- BUT it is easy to obtain such information from English NPs;
- IDEA: Making use of parallel corpora and annotating plurality and definiteness of Chinese NPs automatically.
- Alignment $\rightarrow$ Matching Chinese and English NPs $\rightarrow$ Annotating Chinese NPs using English NPs

A good candidate corpus should:

- be large;
- be parallel;
- be of good quality;
- use informal language. (!)

Idea: Automatic Construction from a parallel corpus

- It is hard to annotate plurality and definiteness of Chinese NPs and construct a large-scale dataset;
- BUT it is easy to obtain such information from English NPs;
- IDEA: Making use of parallel corpora and annotating plurality and definiteness of Chinese NPs automatically.
- Alignment $\rightarrow$ Matching Chinese and English NPs $\rightarrow$ Annotating Chinese NPs using English NPs

A good candidate corpus should:

- be large;
- be parallel;
- be of good quality;
- use informal language. (!)

Idea: Automatic Construction from a parallel corpus

- It is hard to annotate plurality and definiteness of Chinese NPs and construct a large-scale dataset;
- BUT it is easy to obtain such information from English NPs;
- IDEA: Making use of parallel corpora and annotating plurality and definiteness of Chinese NPs automatically.
- Alignment → Matching Chinese and English NPs → Annotating Chinese NPs using English NPs

A good candidate corpus should:

- be large;
- be parallel;
- be of good quality;
- use informal language. (!)

Idea: Automatic Construction from a parallel corpus

- It is hard to annotate plurality and definiteness of Chinese NPs and construct a large-scale dataset;
- BUT it is easy to obtain such information from English NPs;
- IDEA: Making use of parallel corpora and annotating plurality and definiteness of Chinese NPs automatically.
- Alignment $\rightarrow$ Matching Chinese and English NPs $\rightarrow$ Annotating Chinese NPs using English NPs

A good candidate corpus should:

- be large;
- be parallel;
- be of good quality;
- use informal language. (!)

Idea: Automatic Construction from a parallel corpus

- It is hard to annotate plurality and definiteness of Chinese NPs and construct a large-scale dataset;
- BUT it is easy to obtain such information from English NPs;
- IDEA: Making use of parallel corpora and annotating plurality and definiteness of Chinese NPs automatically.
- Alignment $\rightarrow$ Matching Chinese and English NPs $\rightarrow$ Annotating Chinese NPs using English NPs

A good candidate corpus should:

- be large;
- be parallel;
- be of good quality;
- use informal language. (!)

Subtitle

The Data

- ZIMUZU[3] maintained a large-scale dataset of subtitles of television episodes and movies, but it was down because of IP protection;
- We use the pre-processed version by Wang et al. (2018)[4];
- Two full episodes were selected as dev set.

[3] http://www.zimuzu.tv/

[4] Wang et al., "Translating pro-drop languages with reconstruction models".

The Data

- ZIMUZU[3] maintained a large-scale dataset of subtitles of television episodes and movies, but it was down because of IP protection;
- We use the pre-processed version by Wang et al. (2018)[4];
- Two full episodes were selected as dev set.

[3] http://www.zimuzu.tv/

[4] Wang et al., "Translating pro-drop languages with reconstruction models".

The Data

- ZIMUZU[3] maintained a large-scale dataset of subtitles of television episodes and movies, but it was down because of IP protection;
- We use the pre-processed version by Wang et al. (2018)[4];
- Two full episodes were selected as dev set.

[3] http://www.zimuzu.tv/

[4] Wang et al., "Translating pro-drop languages with reconstruction models".

The Data

- ZIMUZU[3] maintained a large-scale dataset of subtitles of television episodes and movies, but it was down because of IP protection;
- We use the pre-processed version by Wang et al. (2018)[4];
- Two full episodes were selected as dev set.

Data	$\|S\|$	$\|W\|$		$\|P\|$		$\|V\|$		$\|L\|$	
		Zh	En	Zh	En	Zh	En	Zh	En
Train	2.15M	12.1M	16.6M	1.66M	2.26M	151K	90.8K	5.63	7.71
Tune	1.09K	6.67K	9.25K	0.76K	1.03K	1.74K	1.35K	6.14	8.52
Test	1.15K	6.71K	9.49K	0.77K	0.96K	1.79K	1.39K	5.82	8.23

[3] http://www.zimuzu.tv/

[4] Wang et al., "Translating pro-drop languages with reconstruction models".

(Automatic) Annotation

1. Word Alignment: GIZA++ is used to do alignment in both directions;
2. NP Identification: CoreNLP is used to identify NPs in both languages;
3. NP Matching: a simple algorithm that works OK
   - For each direction, an NP in the source language is paired with the NP in the target language that has the most aligned tokens with it;
   - A match is done *iff* a pair of NP is paired in both directions.
4. Post-processing: Conjunctions are removed; Other inner structure are ignored;
5. Annotation:
   - Plurality: POS + number
   - Definiteness: Article + Determiner + Proper Name (?)

(Automatic) Annotation

1. Word Alignment: GIZA++ is used to do alignment in both directions;
2. NP Identification: CoreNLP is used to identify NPs in both languages;
3. NP Matching: a simple algorithm that works OK
   - For each direction, an NP in the source language is paired with the NP in the target language that has the most aligned tokens with it;
   - A match is done *iff* a pair of NP is paired in both directions.
4. Post-processing: Conjunctions are removed; Other inner structure are ignored;
5. Annotation:
   - Plurality: POS + number
   - Definiteness: Article + Determiner + Proper Name (?)

(Automatic) Annotation

1. Word Alignment: GIZA++ is used to do alignment in both directions;
2. NP Identification: CoreNLP is used to identify NPs in both languages;
3. NP Matching: a simple algorithm that works OK
   - For each direction, an NP in the source language is paired with the NP in the target language that has the most aligned tokens with it;
   - A match is done *iff* a pair of NP is paired in both directions.
4. Post-processing: Conjunctions are removed; Other inner structure are ignored;
5. Annotation:
   - Plurality: POS + number
   - Definiteness: Article + Determiner + Proper Name (?)

(Automatic) Annotation

1. Word Alignment: GIZA++ is used to do alignment in both directions;
2. NP Identification: CoreNLP is used to identify NPs in both languages;
3. NP Matching: a simple algorithm that works OK
   - For each direction, an NP in the source language is paired with the NP in the target language that has the most aligned tokens with it;
   - A match is done *iff* a pair of NP is paired in both directions.
4. Post-processing: Conjunctions are removed; Other inner structure are ignored;
5. Annotation:
   - Plurality: POS + number
   - Definiteness: Article + Determiner + Proper Name (?)

(Automatic) Annotation

1. Word Alignment: GIZA++ is used to do alignment in both directions;
2. NP Identification: CoreNLP is used to identify NPs in both languages;
3. NP Matching: a simple algorithm that works OK
   - For each direction, an NP in the source language is paired with the NP in the target language that has the most aligned tokens with it;
   - A match is done *iff* a pair of NP is paired in both directions.
4. Post-processing: Conjunctions are removed; Other inner structure are ignored;
5. Annotation:
   - Plurality: POS + number
   - Definiteness: Article + Determiner + Proper Name (?)

The Resulting Corpus

		Plurality		Definiteness	
	Size	Singular	Plural	Definite	Indefinite
train	103686	79158	24528	48471	55215
dev	10368	7894	2474	4777	5591
test	10369	7925	2444	4844	5525

Overview

Human Evaluation

- We random sampled 400 samples for human evaluation;
- Each sample was judged by at least 2 annotators;
- We evaluated the correctness of: NP identification, plurality annotation and definiteness annotation;

Human Evaluation

- We random sampled 400 samples for human evaluation;
- Each sample was judged by at least 2 annotators;
- We evaluated the correctness of: NP identification, plurality annotation and definiteness annotation;

Human Evaluation

- We random sampled 400 samples for human evaluation;
- Each sample was judged by at least 2 annotators;
- We evaluated the correctness of: NP identification, plurality annotation and definiteness annotation;

Human Evaluation

- We random sampled 400 samples for human evaluation;
- Each sample was judged by at least 2 annotators;
- We evaluated the correctness of: NP identification, plurality annotation and definiteness annotation;

	$\text{Acc}_{=2}$	$\text{Acc}_{\geq 1}$	IAA
NP Identification	79.50	96.25	0.8335
Plurality	84.00	96.75	0.8725
Definiteness	81.00	97.25	0.8375

BUT THIS IS NOT QUITE RIGHT!

- As a native speaker, deciding definiteness is hard;
- It is abnormal to obtain such a high IAA;
- Maybe because it is hard, subjects simply agree on what they see?
- We, therefore, conducted experiment 2, where we sampled 200 samples and asked each subject to directly annotate plurality and definiteness; and compare the results with the corpus.

BUT THIS IS NOT QUITE RIGHT!

- As a native speaker, deciding definiteness is hard;
- It is abnormal to obtain such a high IAA;
- Maybe because it is hard, subjects simply agree on what they see?
- We, therefore, conducted experiment 2, where we sampled 200 samples and asked each subject to directly annotate plurality and definiteness; and compare the results with the corpus.

BUT THIS IS NOT QUITE RIGHT!

- As a native speaker, deciding definiteness is hard;
- It is abnormal to obtain such a high IAA;
- Maybe because it is hard, subjects simply agree on what they see?
- We, therefore, conducted experiment 2, where we sampled 200 samples and asked each subject to directly annotate plurality and definiteness; and compare the results with the corpus.

BUT THIS IS NOT QUITE RIGHT!

- As a native speaker, deciding definiteness is hard;
- It is abnormal to obtain such a high IAA;
- Maybe because it is hard, subjects simply agree on what they see?
- We, therefore, conducted experiment 2, where we sampled 200 samples and asked each subject to directly annotate plurality and definiteness; and compare the results with the corpus.

	Experiment 1			Experiment 2		
	$\text{Acc}_{=2}$	$\text{Acc}_{\geq 1}$	IAA	$\text{Acc}_{=2}$	$\text{Acc}_{\geq 1}$	IAA
NP Identification	79.50	96.25	0.8335	-	-	-
Plurality	84.00	96.75	0.8725	74.00	85.50	0.8940
Definiteness	81.00	97.25	0.8375	53.00	77.50	0.7627

Limitations

1. Human's choices vary:
   - It is hard to assess the 'real' quality of the corpus;
   - Is it optimal to assign only a single label?
2. We only assessed the precision, but overlooked recall;
3. We don't know how subjects decided definiteness and plurality.
   - If syntactic features play important roles, then we might have overlooked some indefinite NPs that express definite meaning;
   - If Chinese speakers have no idea of accountability, then they may view uncountable nouns as plural.

Limitations

1. Human's choices vary:
   - It is hard to assess the 'real' quality of the corpus;
   - Is it optimal to assign only a single label?
2. We only assessed the precision, but overlooked recall;
3. We don't know how subjects decided definiteness and plurality.
   - If syntactic features play important roles, then we might have overlooked some indefinite NPs that express definite meaning;
   - If Chinese speakers have no idea of accountability, then they may view uncountable nouns as plural.

Limitations

1. Human's choices vary:
   - It is hard to assess the 'real' quality of the corpus;
   - Is it optimal to assign only a single label?
2. We only assessed the precision, but overlooked recall;
3. We don't know how subjects decided definiteness and plurality.
   - If syntactic features play important roles, then we might have overlooked some indefinite NPs that express definite meaning;
   - If Chinese speakers have no idea of accountability, then they may view uncountable nouns as plural.

Overview

Background

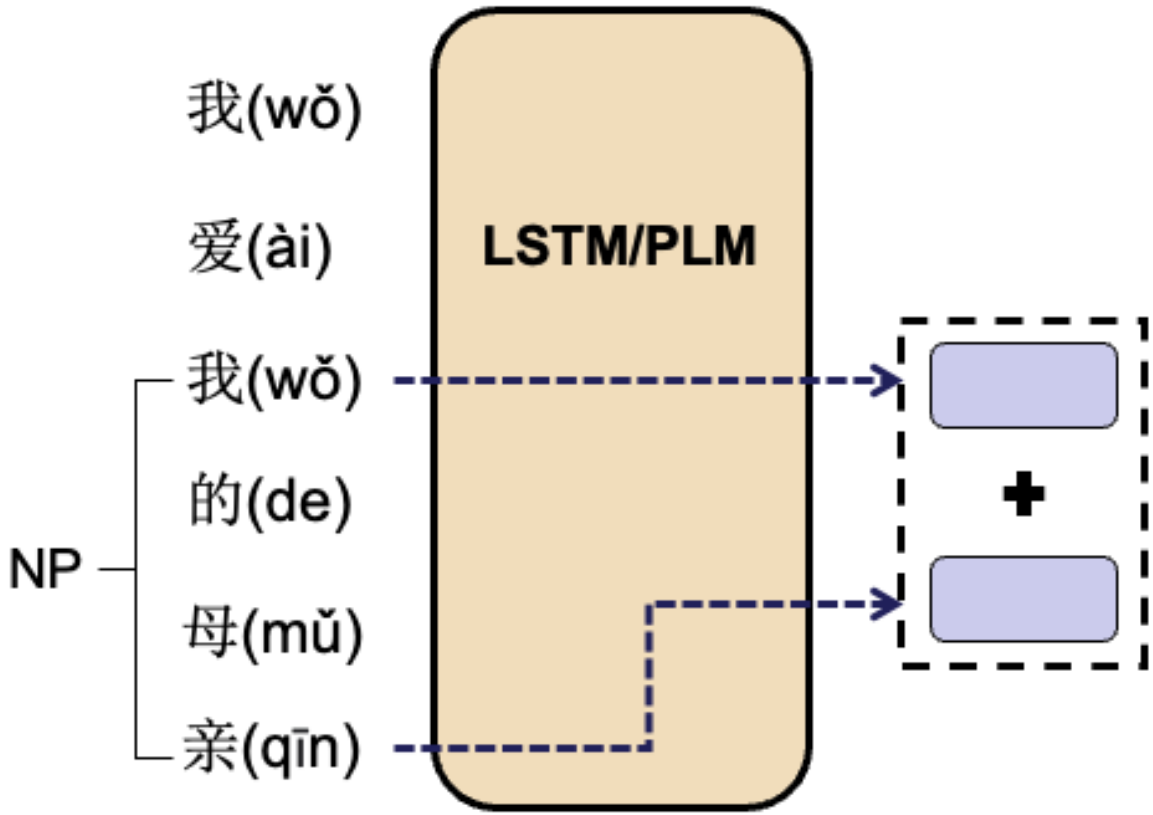
Data

Computational Modelling

Recall

1. Grammar of cool languages allows the absence of these components;
2. (Comprehension) The meaning of these absences can be inferred from their contexts;
   (Production) A component can be dropped and is often dropped if its meaning can be inferred from its context.

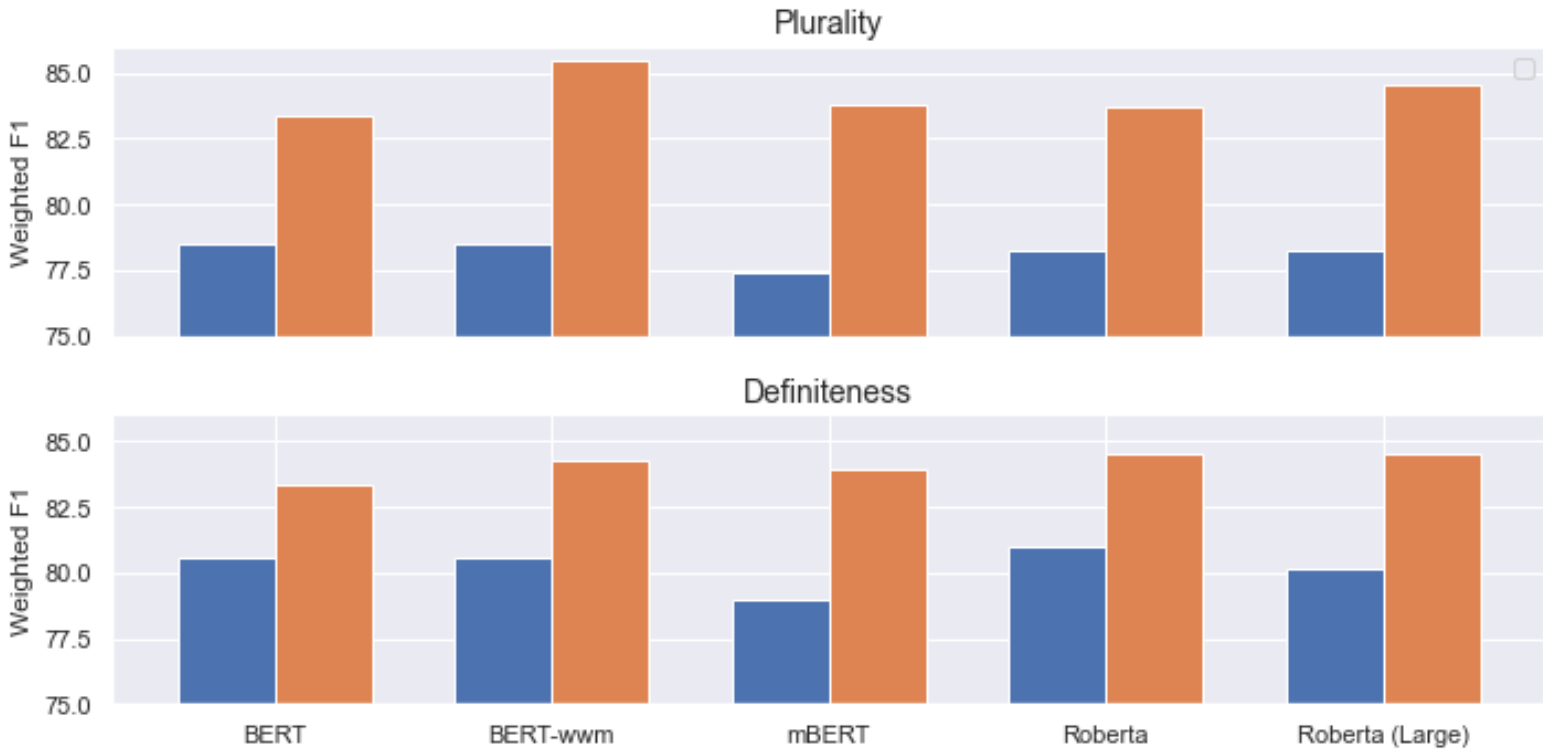
Computational Models for Plurality/Definiteness



Results

	Plurality						Definiteness					
	Macro avg			Weighted avg			Macro avg			Weighted avg		
	P	R	F	P	R	F	P	R	F	P	R	F
RF	81.08	58.19	58.53	80.26	79.69	74.19	68.63	67.24	67.10	68.51	68.09	67.47
LR	76.08	67.39	69.79	80.11	81.58	79.77	71.73	71.53	71.58	71.78	71.82	71.75
SVM	75.56	67.37	69.69	79.88	81.40	79.65	71.34	71.04	71.10	71.37	71.40	71.29
BiLSTM	79.31	70.94	73.59	82.49	83.50	82.14	76.78	76.88	76.80	76.95	76.84	76.87
BERT	80.88	77.96	79.24	85.23	85.73	85.37	81.60	81.66	81.63	81.71	81.69	81.69
BERT-wwm	80.94	**78.34**	**79.50**	**85.38**	85.83	**85.52**	81.95	81.82	81.87	81.98	81.98	81.97
mBERT	80.07	76.96	78.30	84.58	85.15	84.74	80.70	80.41	80.50	80.68	80.66	80.62
RoBERTa	81.21	77.53	79.09	85.22	85.79	85.35	**82.27**	**82.10**	**82.16**	**82.28**	**82.28**	**82.26**
RoBERTa (large)	**81.72**	77.37	79.17	**85.38**	**85.98**	85.46	81.80	81.58	81.66	81.79	81.79	81.76

How does the explicitness impact the model's behaviours?



Do the plurality and definiteness predictions help each other?

	4-way						2-way (merged)					
	Macro avg			Weighted avg			Macro avg			Weighted avg		
	P	R	F	P	R	F	P	R	F	P	R	F
BERT	67.37	64.26	65.53	70.72	71.20	70.79	65.62	63.35	64.34	69.49	69.91	69.61
BERT-wwm	67.94	65.74	66.72	71.54	71.86	71.62	66.51	**64.23**	**65.24**	70.03	70.40	70.14
mBERT	67.73	64.58	65.69	71.12	71.46	71.01	64.19	61.51	62.62	68.11	68.59	68.21
RoBERTa	68.25	**66.42**	**67.24**	72.03	72.36	72.14	67.08	63.89	65.23	**70.29**	**70.74**	**70.36**
RoBERTa (large)	**68.73**	65.51	66.87	**72.09**	**72.55**	**72.18**	**67.11**	63.36	64.90	69.90	70.35	69.92

Findings

1. We constructed a Chinese NP corpus in which plurality and definiteness are annotated. The corpus is available at: `https://github.com/andyzxq/chinese_np_def`
2. We build computational models to use plurality and definiteness in NPs as two examples to check one pragmatic aspect of the "coolness" hypothesis: *in a "cool" language, whether the meaning of an omittable component is predictable or not*.
    - The answer is YES!
    - The information for predicting plurality and definiteness benefits from each other.

Many Thanks!