Annotating dropped pronouns in Chinese newswire text

Elizabeth Baran, Yaqin Yang, Nianwen Xue

Computer Science Department, Brandeis University
415 South Street, Waltham MA, USA
ebaran@brandeis.edu, yaqin@brandeis.edu, xuen@brandeis.edu

Abstract

We propose an annotation framework to explicitly identify dropped subject pronouns in Chinese. We acknowledge and specify 10 concrete pronouns that exist as words in Chinese and 4 abstract pronouns that do not correspond to Chinese words, but that are recognized conceptually, to native Chinese speakers. These abstract pronouns are identified as "unspecified", "pleonastic", "event", and "existential" and are argued to exist cross-linguistically. We trained two annotators, fluent in Chinese, and adjudicated their annotations to form a gold standard. We achieved an inter-annotator agreement kappa of .6 and an observed agreement of .7. We found that annotators had the most difficulty with the abstract pronouns, such as "unspecified" and "event", but we posit that further specification and training has the potential to significantly improve these results. We believe that this annotated data will serve to help improve Machine Translation models that translate from Chinese to a non pro-drop language, like English, that requires all subject pronouns to be explicit.

Keywords: pro-drop, Chinese text annotation, Chinese language processing

1. Introduction

Chinese allows a pronoun to be dropped subject to various syntactic and pragmatic constraints and this is a phenomenon that has attracted substantial interest in theoretical linguistics (Huang, 1984; Huang, 1989). In a morphologically rich language like Italian, the pro-drop in the subject position is often licensed by agreement morphology on the verb. The lack of such agreement morphology in Chinese only adds to the intrigue of the very possibility of pro-drop in such a typologically different language. Our interest in pro-drop is motivated by practical natural language applications such as Machine Translation. When a pro-drop language such as Chinese is translated into a non pro-drop language like English, a null (or dropped) subject would have to be made explicit, and this has proved to be a challenging problem for current statistical machine translation systems which do not model this problem explicitly due to lack of resources annotated with pro-drop.

In this paper we describe an annotation project in which we mark up (i) the locations in a sentence where a pronoun is dropped, and (ii) the actual pronoun that is dropped at that location. The pronoun is chosen from an inventory of possible pronouns in Chinese, and in cases where no Chinese pronoun can plausibly occur in that position, we posit abstract pronominal types based on their referential properties. It turns out that these abstract referential types often have explicit counterparts in other languages. Our ultimate goal is to model the distribution of these dropped pronouns automatically, in the hope that they will contribute to improving Machine Translation and other natural language applications.

This rest of the paper is organized as follows. Section 2 describes the procedure and Section 3 presents the annotation scheme and discusses a few complications that make this task particularly challenging. In Section 4 we discuss our results and agreement scores. Section 5 concludes this paper.

2. Annotation Procedure

We used a portion of the manually parsed articles from the newswire section of the Chinese Treebank (CTB) (Xue et al., 2005) to carry out this task. Please see Table 1 for a complete list of the currently annotated data set.

File Numbers
41-43, 45-52 , 54-70 , 72-73 , 75 , 77-80,
283-290, 292-300 , 302 , 304-306, 309-310 ,
314-320 , 322-324 , 326-404 406-410 , 412-414 ,
421-423 , 425-438 , 440-519 , 521-525 ,
527-528 , 530 , 532 , 535-536, 600-605 ,
607-614 , 618-621 , 624-640 , 642-644 ,
646-657 , 659-660 , 662-664 , 666-676 ,
679-684 , 686-687 , 690 , 692-711,
713-729 , 736-749 , 751

Table 1: Chinese Treebank files used for annotation.

The advantage of starting from the CTB parses is that the locations of the dropped pronouns have already been annotated as part of the treebanking process. Two types of dropped pronouns are annotated in the Chinese Treebank: ∗PRO∗ and ∗pro∗s. Differentiating the two types of dropped pronouns is a difficult task, and it has been argued in the literature that their distributions are governed by similar linguistic principles (Huang, 1989), so we have made the decision to annotate both types of dropped pronouns. Since the locations of dropped pronouns can be automatically read off the syntactic parses, our focus in this project is to identify the overt pronoun that can plausibly substitute for each dropped pronoun in the CTB articles. We gave these articles to two Chinese speakers to annotate. Both annotators were provided with guidelines and trained on a small set of files. The annotators used an annotation interface we developed in-house and they were asked to se-

lect from a list of pronouns. Their annotations were then adjudicated when there was a disagreement. Out of 273 files that we have annotated so far, there are a total of 2,221 dropped pronouns.

3. Annotation Scheme

In our annotation framework we propose 14 types of pronouns. Of these 14 pronouns, 10 of them are actual pronouns that are used commonly in Chinese speech and writing. These pronouns and their corresponding distributions in the data are listed below:

1. 我(I) first person singular 1.4%
2. 我们(we) first person plural 1%
3. 你(you) second person singular $<$ 1%
4. 你们(you) second person plural $<$ 1%
5. 他(he) third person masculine singular 7.4%
6. 他们(they) third person masculine plural 10.8%
7. 她(she) third person feminine singular $<$ 1%
8. 她们(they) third person feminine plural $<$ 1%
9. 它(it) third person inanimate singular 37.8%
10. 它们(they) third person inanimate plural 13.9

Note that within the news genre, the third person inanimate pronouns followed by the third person masculine pronouns are significantly more common than any of the other pronouns in our dataset. Example 1 demonstrates a context for the most common type of pronoun, the inanimate third person singular pronoun 它 .

(1) 他 说 , 该 项目 的 实施
He said , this program DE implementation
将 推动 中 叙 两 国
will push forward China Syria two countries
经贸 关系 进一步 发展 ,
trade relation progressive development ,
标志 着 中国 纺织 机械
indicates ZHE China textile machinery
成套 设备 制造 水平
complete set equipment production level
已 提高 到 一 个 新 的 水平
already raised reach one CL new DE level
, 具备 了 *[它]* **参与** 国际
, possess LE [it] **participate** international
竞争 的 能力 。
competition DE ability.

He said, the implementation of this program will push forward the progressive development of Sino-Syrian trade relations, indicating that the level of production of China's textile machinery and full equipment has already reached a new level, and that it possesses the ability to **participate** in international competition.

In this example, it is clear that the one that is "participating in international competition" would be 中国 (China) which would be referred to with the pronoun 它 .

Besides these, we introduced 4 abstract pronominal types that we believe exist in Chinese but do not correspond to a specific Chinese word. These pronouns are expressible in non pro-drop languages like English.

1. existential 3.7%
2. unspecified 17.7%
3. event 1.6%
4. pleonastic 2.9%

As shown by the distributions, these abstract pronouns are fairly common in Chinese news corpora. The following subsections elaborate on these abstract pronouns and highlight some of the other intricacies related to this type of annotation.

3.1. Abstract Pronouns

existential The context that may suggest an **existential** subject is very limited in scope. An existential subject appears in front of a small number of "existence" verbs. In Chinese, an existential subject is most commonly paired with the verb 有 (to have) which, in existential contexts, can be translated as "there is" or "there are" in English. However, other verbs, like 存在, meaning "to exist" , can clearly also suggest the existence of an implicit existential subject. Example 2 shows an existential subject preceding the verb 有 (to have).

(2) 目前 *[existential]* 已 **有** 一 批
Currently [existential] already **have** one group
特大型 的 工业 投资 项目
extra large DE industrial investment programs
落户 新 区 .
settle in new area .

At present, **there is** already a number of mega industrial investment programs settling into the new area.

unspecified An **unspecified** subject occurs in one of two situations. The first situation is when there is no one specific thing or person that should be interpreted as doing the action, but rather anyone could be a possible subject. This type of subject can sometimes be translated to "one" in English. An unspecified subject is different from a subject that is difficult to explicitly identify. In other words, the unspecified subject should not be the result of the annotator's inability to identify a clear subject, but rather the annotator's conscious understanding that the subject could in fact be anyone and that this generality was intentional by the writer. In example 3 an unspecified subject precedes the verb 可以 (can).

(3) 这 次 批准 建立 的 五 家
This time approved establish DE 5 MW
保险 公司 , *[unspecified]* **可以** 说
insurance companies , [one] **can** say
是 由此 推出 的 一 大 举措 。
is from this put forth DE one big initiative .

One **can** say that the approval, this time, to establish 5 insurance companies was a big initiative put forth from here.

The second type of **unspecified** subject is one in which not just anybody could have performed the action, and in fact a particular person or organization must have performed it, but the identity of that subject has been deemed irrelevant to the reader, and it is clear from context that you do not know who it is nor that you need to know who it is. The subject of the verb, 设立 (establish) in sentence 4 is an example of this.

(4) 中国 批准 *[unspecified]* **设立** 了
China approve [someone] **establish** LE
三十万 家 外商 投资
300 thousand MW foreign merchants invest
企业 。
enterprises.
China approved the **establishment** of 300 thousand foreign-invested enterprises.

event An **event** subject is something that has occurred and is described in context (generally as a whole phrase), but has not yet been explicitly nominalized. When referring back to an event, it may feel more natural to use a demonstrative pronoun, like 这 (this) or 这件事(this thing/event) rather than the seemingly all-purpose 它.

(5) 近 五 年 来 ， 外商
Recent 5 years come , foreign merchants
投资 企业 的 进出口 在 中国
invest enterprises DE import-export in China
对外 贸易 中 所占 比重
outside trade within represent proportion
快速 增加 ， *[event]* **推动** 了
quickly increases , this push forward LE
中国 进出口 贸易 的 发展 。
China import-export trade DE development .
In the last 5 years, the proportion, in terms of foreign trade in China, of importing and exporting of foreign-invested enterprises has quickly increased. This has **pushed forward** the development of China's import-export trade.

In example , the subject of 推动 (pushed forward) is the whole preceding phrase that translates roughly to " the proportion, in terms of foreign trade in China, of importing and exporting of foreign-invested enterprises has quickly increased" . It is strange to refer to this phrase as 它 (it), because the key element"increased" is acting more verb-like. Therefore we call it an "event" . However, if this event were to be nominalized in a way that would translate e increase" , then "the increase" would be referred to with 它 .

pleonastic A subject is **pleonastic** if it is difficult to grasp any sort of notion of the semantic content of a dropped subject. This may be a case in which the dropped subject actually has no semantic content and is only necessary to satisfy the syntactic need for a subject. This often occurs with phrases that suggest the passing of time as in example 6.

(6) 从 １９９７年 *[pleonastic]* 开始 ，
From the year 1997 [pleonastic] starting ,
捷克 重新 调整 外贸
Czech start over adjust foreign investment
政策 ， 积极 鼓励 出口 ， 限制
policies , actively encourage export , limit
盲目 进口 ， 同时 大力 调整
blind imports , same time large force adjust
进出口 商品 结构 ， 努力
commodities structure , work hard increase
增加 附加值 高 的 商品
value-added high DE commodities export
出口 。
.
Beginning in 1997, the Czech Republic began to re-adjust foreign investment policies, actively encourage export, limit blind importing, and at the same time forcefully adjust the commodities structure and work hard to increase the high value-added export commodities.

3.2. Other Specifications and Challenges

Annotators are instructed to annotate from a third person perspective unless certain writing mechanisms in effect dictate otherwise. Quoted text, for example, forces a new perspectual environment and therefore modifies the annotation parameters until the end quote. In fact, direct quotations are one of the few places that first and second person pronouns can be found in news corpora, since the transition into quoted text is accompanied by a change in perspective from a third person observer to a first person experiencer.
Sometimes, the referent for a dropped pronoun is clear, but the pronoun to represent it is not. In news corpora, this occurs most frequently with nouns that represent organizations. An organization can be interpreted as one singular unit, i.e. the organization as a whole, or as the people who make up the organization. For example, 公司 (company) could be (a) the company as a unit, in which case the corresponding pronoun would be 它 or (b) the people that make up the company, in which case the corresponding pronoun would be 他们. Both of these interpretations are valid, however we posit that the default underlying representation is the single unit. In our annotation framework, this default interpretation can be overridden only if the context has supplied explicit reference to a different pronoun.
Nouns do not inflect for number in Chinese, but pronouns do. This contradiction poses an interesting challenge to annotators of dropped pronouns. Not only do they need to find the referent of the dropped pronoun, but they also need to recover number information about the noun that is not explicit on the noun itself. They must rely on context and their intuitions as a Chinese speaker to correctly decipher plural nouns from singular nouns, and still interpretations can vary.

(7) 该 **公司** 介绍 ， 在 未来 的 五
This **company** introduce , in future DE 5
年 内 **他们** 将 追加 投资 九千万
years in **they** will add investment 90 million
美元 ， 届时 ， *[他们]* **预计**
USD , at that time , [they] **estimate**
年产 值 可 达
yearly production value could reach

三亿 美元 。
three hundred million USD .

The **company** explained that within the next five years, **they** would seek to add 80 million USD in investments. At that time, **they estimate** that the yearly production value could reach three hundred million USD.

The beginning of the sentence alerts us to the fact that we are talking about 公司 (a company), but then later instantiates 公司 with the pronoun 他们, interpreting it as a group of people that represent the company. Immediately following is the verb to be annotated, 预计 which we understand to also take 公司 as its implicit subject. We follow the framework described by (Baran and Xue, 2011), so for this case we have parallel evidence to support the second interpretation of 公司 (i.e. as a group of people), this would be annotated with the subject 他们 to stay faithful to the particular context.

(8) 宋健 最后 说 , " *[我们]*
Songjian lastly said , " we
能否 把 我们 自己 的 高
can or can not BA we oneself DE high
技术 及 其 产业 搞 上去 ,
technology and our industries do upwards ;
关系 到 中国 现代化 建设
relate to China modernization construction
事业 的 成败 , 关系 到
undertaking DE success or failure ; relate to
中华 民族 的 兴衰 。 "
Chinese peoples DE rise and fall . "

Songjian finally said, "How can we boost the integration of our high technology and industries; relate them to the success and failure of the undertaking of modernization construction; how can we relate them to the rise and fall of the Chinese people?"

4. Inter-Annotator Agreement

We measure inter-annotator agreement by calculating the pairwise Kappa coefficient (Cohen, 1960) as well as the observed agreement between the two annotators. Table 2 shows these two scores as well as the accuracy of each annotator when compared to the gold standard.

	IAA Scores
Kappa	.6
Observed	.7

Table 2: Inter-Annotator Agreement Scores: Pairwise Kappa and observed agreement between annotators.

Table 3 shows the f1 scores for all annotation categories for each annotator compared to the gold standard.
From both of these tables, we can see that both annotators performed moderately well compared to the gold standard,

	Annotator 1	Annotator 2
我	82.1%	93.1%
我们	78.2%	88.9%
你	0.0%	0.0%
你们	57.1%	80.0%
他	90.1%	92.8%
他们	85.5%	86.5%
她	92.7%	100%
她们	94.1%	66.7%
它	88.5%	87.0%
它们	80.7%	80.1%
unspecified	77.8%	72.3%
existential	92.0%	87.4%
pleonastic	66.2%	62.7%
event	67.7%	50.7%

Table 3: F1 scores for each annotator compared to the adjudicated gold standard.

yet the lower IAA scores suggest that they excelled in different categories. We see that annotator 2 understood the placement of first and second person pronouns better than annotator 2, however annotator 1 was able to more accurately identify abstract pronouns.
We get a better a picture of the tendencies of the two annotators in Table 4 which is a confusion matrix between the two annotators. We see that the annotators agreed most often when it came to concrete people pronouns. This implies that the context was less ambiguous when people were the subjects, especially for first person, second person, or third person feminine pronouns. Since the masculine gender is used as a default to refer to people when the genders are mixed or not clear, an annotator choosing to annotate with a feminine pronoun tends to be sure that the antecedent is in fact feminine. Both annotators tended to confuse "event" pronouns with 它 ; this emphasizes the need to train further on distinguishing a verb-like event ("event") from a nominalized event (它). We can see that annotators have the most trouble with the most frequent pronouns, which are 它, unspecified, 它们, 他们, and even to some extent 他. This confusion is in line with the breadth the nouns that these pronouns often refer to, and also with their tendencies to step in as default choices when in doubt about the specific nature of the noun.

5. Conclusion

The annotation of dropped pronouns in Chinese is a crucial step in improving current translation models from Chinese to syntactically richer languages like English. By having data with null subject pronouns filled in, we can begin to

	它	unsp	它们	他们	他	ex	pleo	event	我	我们	她	你们	她们	你
它	**643**	90	78	15	4	11	23	11	.	.	.	.	.	.
unsp	89	**224**	32	41	19	1	19	4	1	5	2	7	1	.
它们	32	20	**179**	11	.	4	2	3	.	1	.	1	.	.
他们	8	10	8	**186**	3	1	2	.	.	.	1	1	.	.
他	6	5	.	5	**151**	.	.	.	1	.	.	.	.	.
ex	1	1	4	2	.	**70**	2	1	.	.	.	.	3	.
pleo	8	5	5	2	2	2	**29**	1	.	.	.	.	.	.
event	14	2	6	1	2	2	1	**9**	.	1	.	.	1	.
我	.	1	.	.	6	.	.	.	**20**	.	.	.	.	.
我们	2	1	.	.	2	.	.	.	1	**16**	.	.	.	.
她	.	.	.	.	.	.	.	.	.	.	**19**	.	.	.
你们	.	1	.	.	.	.	.	.	2	.	.	**5**	.	.
她们	.	.	.	.	.	.	.	.	.	.	.	.	**4**	.
你	.	.	.	.	2	.	.	.	.	.	.	.	.	**.**

Table 4: Confusion matrix for each annotation category. Columns correspond to annotator 1's annotation values and rows refer to annotator 2's annotation values.

learn ways to more accurately translate Chinese into a language like English that requires explicit subject pronouns. We show that our annotation framework is a solid foundation on which to carry out this task.

Since different categories posed problems for each annotator, we believe that inter-annotator agreement scores could be significantly improved with a second stage of annotation training that honed in on those problem categories for the individual. Also, we believe that further specifying abstract pronouns in the guidelines would also improve both accuracy and agreement scores since they are less intuitive than the commonly used set of pronouns in Chinese. In the future, it would be interesting to expand this work to different genres of Chinese text that would provide more instances of first, second, and feminine pronouns, which are underrepresented in news corpora.

Acknowledgements

This work is supported by the National Science Foundation via Grant No. 0910532 entitled "Richer Representations for Machine Translation" . All views expressed in this paper are those of the authors and do not necessarily represent the view of the National Science Foundation.

6. References

Elizabeth Baran and Nianwen Xue. 2011. Singular or plural? exploiting parallel corpora for chinese number prediction. In *Proceedings of the 13th Machine Translation Summit*, pages 207–214. Asia-Pacific Association for Machine Translation, September.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. volume 20, pages 37–46.

James C. T. Huang. 1984. On the distribution and reference of empty pronouns. *Linguistic Inquiry*, 15:531–574.

James C.-T. Huang. 1989. Pro drop in Chinese, a generalized control approach. In Jaeggli O and K. Safir, editors, *The Null Subject Parameter*. D. Reidel Dordrecht.

Nianwen Xue, Fei Xia, Fu Dong Chiou, and Martha Palmer. 2005. The penn chinese treebank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, pages 207–238.