Recovering dropped pronouns from Chinese text messages

Yaqin Yang
Paypal Inc.
yaqin276@gmail.com

Yalin Liu
Brandeis University
yalin@brandeis.edu

Nianwen Xu
Brandeis University
xuen@brandeis.edu

Abstract

Pronouns are frequently dropped in Chinese sentences, especially in informal data such as text messages. In this work we propose a solution to recover dropped pronouns in SMS data. We manually annotate dropped pronouns in 684 SMS files and apply machine learning algorithms to recover them, leveraging lexical, contextual and syntactic information as features. We believe this is the first work on recovering dropped pronouns in Chinese text messages.

1 Introduction

Text messages generated by users via SMS or Chat have distinct linguistic characteristics that pose unique challenges for existing natural language processing techniques. Since such text messages are often generated via mobile devices in informal settings and are limited in length, abbreviations and omissions are commonplace. In this paper, we report work on detecting one particular type of omission in Chinese text messages, namely dropped pronouns.

It is well-known that Chinese is a pro-drop language, meaning pronouns can be dropped from a sentence without causing the sentence to become ungrammatical or incomprehensible when the identity of the pronoun can be inferred from the context. Pronouns can be dropped even in formal text genres like newswire, but the extent to which this happens and the types of pronouns that are dropped in text messages and formal genres like newswire are very different. For example, the most frequently dropped pronouns in Chinese newswire is the third person singular 它("it") (Baran et al. 2012), and one reason is that first and second person pronouns are rarely used in newswire in the first place. In contrast, in text messages, the first person singular 我 and the second person singular 你 are commonly found in text messages due to their conversational style, and they are often dropped as well when their referent is understood in the context. This is illustrated in (1), where there are instances of dropped first person singular, second person singular and third person singular pronouns. There is also an instance where the dropped pronoun in Chinese does not have any actual referent, translating to the English pleonastic "it". Dropped pronouns are in parentheses:

(1) A 你们那 下雪 了 , 你 怎么去 上班
your area snow ASP , you how go work
"It snowed in your area. How do you go to work?"

B (我) 步行 或坐车
(I) walk or take the bus
"(I) walk or take the bus."

A (pleonastic) 看来 交通业 还是
(it) look like transportation
比较 发达 的.
relatively developed
"(It) looks like you have a relatively developed transportation system."

B (pleonastic) 下雪 (我) 就 不 能
(it) snow (I) then not can
上班 了
go to work ASP
"When (it) snows, (I) cannot go to work."

B (它) 还可以
(it) OK
"(It) is OK."

Detecting dropped pronouns involves first of all determining where in the sentence pronouns are

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 309–313,
Beijing, China, July 26-31, 2015. ©2015 Association for Computational Linguistics

dropped and then determining what the dropped pronoun is, i.e., whether the dropped pronoun should be 我, 你, 他, etc. The dropped pronoun could either correspond to one of possible pronouns in Chinese, or it can be an abstract pronoun that does not correspond to any of the Chinese pronouns. For example, Chinese does not have a pronoun that is the equivalent of the pleonastic "it" in English, but there are sentences in which a dropped pronoun occurs in a context that is similar to where "it" occurs. In this case we label the dropped pronoun as a type of abstract pronoun. Note that we do not attempt to resolve these pronouns to an antecedent in this work. We think there is value in just detecting these pronouns. For example, if we translate Chinese sentences with dropped pronouns into English, they may have to be made explicit.

We approach this as a supervised learning problem, so first we need a corpus annotated with the location and type of dropped pronouns to train machine learning models. We annotated 292,455 words of Chinese SMS/Chat data with dropped pronouns and we describe our annotation in more detail in Section 2. We then present our machine learning approach in Section 3. Experimental results are presented in Section 4, and related work is described in Section 5. Finally we conclude in Section 6.

2 Dropped pronoun annotation

We annotated 684 Chinese SMS/Chat files following the dropped pronoun annotation guidelines described in (Baran et al. 2012). The original guidelines are mainly designed for annotating dropped pronouns in newswire text, and we had to extend the guidelines to accommodate SMS/Chat data. For example, (Baran et al. 2012) identify 14 types of pronouns, which include four *abstract pronouns* which do not correspond to any actual pronouns in Chinese. To accommodate SMS/Chat data, we add one more type of abstract pronoun that refers to the previous utterance. The full list of pronouns that we use are listed below:

1. 我(I): first person singular
2. 我们(we): first person plural
3. 你(you): second person singular
4. 你们(you): second person plural
5. 他(he): third person masculine singular
6. 他们(they): third person masculine plural
7. 她(she): third person feminine singular
8. 她们(they): third person feminine plural
9. 它(it): third person inanimate singular
10. 它们(they): third person inanimate plural
11. Event: abstract pronoun that refers to an event
12. Existential: abstract pronoun that refers to existential subject
13. Pleonastic: abstract pronoun that refers to pleonastic subject
14. generic: abstract pronoun that refers to something generic or unspecific
15. Previous Utterance: abstract pronoun that refers to previous utterance
16. Other: cases where it is unclear what the correct pronoun should be

3 Learning

We have formulated dropped pronoun recovery as a sequential tagging problem, following (Yang and Xue. 2010). We check each word token in a sentence and decide if there is a pronoun dropped before this word. If there is one, then we further identify what type of pronoun it should be. Instead of doing this in two separate steps, we trained a 17-class Maximum Entropy classifier with the Mallet (McCallum et al. 2002) machine learning package to tag each word token with one of the pronouns or *None* in one run. *None* indicates that there is no dropped pronoun before this word.

We leveraged a set of lexical features from previous work (Yang and Xue. 2010). To our knowledge, the work we report here represents the first effort on dropped pronoun recovery on Chinese SMS/Chat data. As described in Section 2, SMS data is different from newswire data which is commonly used in previous work (Converse. 2006; Zhao and Ng. 2007; Peng and Araki. 2007; Kong and Zhou. 2010; Chung and Gildea 2010; Cai et al. 2011; Xiang et al. 2013) in many aspects. The frequency of pronoun being dropped is much higher in SMS/Chat data compared to newswire data. The distribution of dropped pronoun types in SMS data is also very different from that of newswire data. In SMS/Chat data, the identities of the participants who send the messages are critical in identifying the dropped pronoun type, while there is no participant information in newswire data. Thus, we also design a new set of context based features to capture the stylistic properties of text messages.

Lexical Features: Information embedded in the target and surrounding words provide clues for identifying dropped pronouns, e.g.,

(2) (它) 坏 了 .
(it) broken ASP .
"(It) is broken."

In (2), a pronoun is dropped at the beginning of the sentence. The follwing words "坏了" means "is broken", and it indicates that the subject refers to a thing, not a person. Part-of-speech tags are also crucial in finding the location of a dropped pronoun. Just like pronouns are usually located before verbs, it is more likely to have a pronoun dropped before an verb than a noun. We implemented a set of lexical features along with part-of-speech tags within a sliding window of five words to capture such information. The contextual features are listed below:

- unigrams within current window;
- previous and following (including current word) bigrams;
- POS tags of unigrams within current window;
- POS tags of the previous and following (including current word) bigrams;
- POS tags of the following (including current word) trigram;
- combination previous word and POS tag of current word;
- combination of POS tag of previous word and current word;
- POS tag sequence from the previous word to the beginning of a sentence or a punctuation mark.

Context-based Features: It is hard to recover dropped pronouns without understanding the context. In SMS data, one sometimes needs to trace back a few sentences to figure out what a dropped pronoun refers to.

(3) a. 我想 买 个 单反 .
I want buy CL SLR camera .
"I want to buy a SLR camera."

b. (我) 国庆节 出去 玩 啊.
(I) Independent Day go out travel .
"(I) will travel on Independent Day."

In (3), the two sentences are attributed to the same person, and a pronoun is dropped at the beginning of the second sentence. While we could easily understand the dropped pronoun refers to "我(I)" from the previous sentence, it is difficult to make this determination by just looking at the second sentence independently. Thus, we propose a list of novel context-based features tailored towards SMS/Chat data to capture such information:

- previous pronoun used by the same participant;
- previous pronoun used by the other participant;
- all previous pronouns till the beginning of a sentence or a punctuation mark used;
- next punctuation mark;
- if it is a question;
- if the POS tag of the last word is SP;
- for the first word in a sentence, use first two nouns/pronouns from the previous sentence.

Syntactic Features: Syntactic features have been shown to be useful in previous work (**?**). We also implemented the following syntactic features:

- if it is the left frontier of the lowest IP antecedent;
- if current word is "有", then find it's subject;
- path from current word to the root node.

4 Experiments and discussion

4.1 Data split

Table 1 presents the data split used in our experiments.

data set	# of words	# of files
Train	235,184	487
Dev	24,769	98
Test	32,502	99

Table 1: Training, development and test data on SMS data set.

4.2 Results

As mentioned in Section 3, we extract lexical, context and syntactic features from SMS data and train a 17-class classifier to automatically recover dropped pronouns. To obtain syntactic features, we divided 684 SMS files into 10 portions, and parsed each portion with a model trained on other portions, using the Berkeley parser (Petrov and Klein 2007). The parsing accuracy stands at 82.11% (F-score), with a precision of 82.57% and a recall of 81.65%. The results of our experiments are presented in Table 2.

tag	pre.(%)	rec.(%)	f.	count
NE	99.1	95.7	97.3	28963
我	48	53.1	50.4	1155
你	34.4	48.1	40.1	787
它	12.1	54.6	19.8	488
prev_utterance	87.6	65.3	74.8	314
pleonastic	7	10.2	8.3	172
她	4.3	27.8	7.4	117
他	11	22.2	14.7	109
我们	24	41	30.3	104
generic	6.6	17.1	9.5	91
他们	2.7	11.1	4.4	73
event	4.3	25	7.3	47
它们	4.7	100	8.9	43
other	0	0	0	16
你们	0	0	0	13
existential	12.5	2	3.4	8
她们	0	0	0	2

Table 2: precision, recall and f-score for different dropped pronoun categories on test set. The combination of "我(I)", "你(singular you)" and "utterance" accounts for 63.7% of the overall dropped pronoun population. The overall accuracy is 92.1%. "NE" stands for None, meaning there is no dropped pronoun.

4.3 Error Analysis

From Table 3, which is a confusion matrix generated from results on the test set, showing the classification errors among different types, we can see that the classifier did a better job of recovering "我(I)", "你(singular you)" and "previous utterance", the combination of which accounts for 63.7% of the total dropped pronoun instances. However, it is hard for the classifier to recover "它(it)", e.g.,

"*pro* 这种? (*pro* that kind?)"

SMS sentences are usually short. To understand what the dropped pronoun stands for, one needs to look at its previous context. But it is hard for machine to capture such long distance information.

5 Related Work

One line of work that is closely related to ours is zero pronoun resolution. In zero pronoun resolution (Converse. 2006; Zhao and Ng. 2007; Peng and Araki. 2007; Kong and Zhou. 2010), pronouns are typically resolved in three steps: zero pronoun detection, anaphoricity determination, and antecedent resolution. In the work we report here, we are more interested in detecting dropped pronouns and determining what types of pronoun they are.

	NE	我	你	它	ut	pl	她	他	我们	ge	他们	ev	它们	ot	你们	ex	她们
NE	28695	130	77	9	8	7	2	10	9	8	1	.	.	.	1	6	.
我	433	554	101	11	5	13	2	5	10	4	1	1	.	.	1	14	.
你	327	135	271	6	3	16	1	6	6	9	1	1	.	.	.	5	.
它	199	85	49	59	23	42	1	10	1	3	5	1	.	.	1	9	.
utterance	23	7	1	4	275	4	.	.	.	.	.	.	.	.	.	.	.
pleonastic	36	17	5	5	88	12	1	1	1	.	1	.	.	.	.	5	.
她	47	21	21	5	1	6	5	5	1	.	3	1	.	.	.	1	.
他	46	23	10	2	6	5	1	12	.	1	.	.	.	.	.	3	.
我们	47	17	5	2	.	.	2	2	25	.	1	1	.	.	.	2	.
generic	52	20	5	.	.	.	.	2	2	6	.	.	.	.	.	4	.
他们	38	15	7	2	.	3	2	.	1	2	2	1	.	.	.	.	.
event	16	4	3	2	11	6	1	1	.	.	1	2	.	.	.	.	.
它们	14	11	4	.	1	3	.	.	3	2	2	.	2	.	.	1	.
other	15	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.
你们	6	2	4	1	.	.	.	.	.	.	.	.	.	.	.	.	.
existential	4	2	.	.	.	.	.	.	1	.	.	.	.	.	.	1	.
她们	1	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.

Table 3: Confusion matrix for each annotation category. Columns correspond to Maxent predicted values and rows refer to annotated values.

Dropped pronoun detection is also related to Empty Category (EC) detection and resolution (Chung and Gildea 2010; Cai et al. 2011; Xiang et al. 2013), the aim of which is to recover long-distance dependencies, discontinuous constituents, and certain dropped elements in phrase structure treebanks (Marcus et al. 1993; Xue et al. 2005). In previous work on EC detection (Chung and Gildea 2010; Cai et al. 2011; Xiang et al. 2013), ECs are recovered from newswire data by leveraging lexical and syntactic information from each sentence. Context information beyond the current sentence is typically not used. When recovering dropped pronouns in SMS/Chat messages, it is crucially important to make use of information beyond the current sentence.

6 Conclusion and Future Work

In this paper we report work on recovering dropped pronouns in Chinese SMS/Chat messages. Based on the properties of SMS data, we designed a set of lexical, contextual and syntactic features, and trained a Maxent classifier to recover dropped pronouns in Chinese SMS/Chat messages. We believe this is the first work on recovering dropped pronouns in Chinese text messages. This proves to be a very challenging task, and much remains to be done. In future work, we plan to experiment with applying more expressive machine learning techniques to this task.

Acknowledgments

We want to thank the anonymous reviewers for their suggestions. This work was partially 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.

References

Zhao, Shanheng and Ng, Hwee Tou 2007 *Identification and Resolution of Chinese Zero Pronouns: A Machine Learning Approach.*. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL).

Kong, Fang and Zhou, Guodong 2010 *A tree kernel-based unified framework for Chinese zero anaphora resolution.*. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing.

Fang Kong and Hwee Tou Ng 2013 *Exploiting Zero Pronouns to Improve Chinese Coreference Resolution.*. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing.

Shu Cai, David Chiang, and Yoav Goldberg. 2011 *Language-independent parsing with empty elements.*. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 212–216, Portland, Oregon, USA, June. Association for Computational Linguistics.

Tagyoung Chung and Daniel Gildea. 2010. *Effects of empty categories on machine translation*. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing.

Elizabeth Baran, Yaqin Yang and Nianwen Xue. 2012 *Annotating dropped pronouns in Chinese newswire text.*. In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey.

Andrew Kachites McCallum. 2002 *Mallet: A machine learning for language toolkit.*. http://mallet.cs.umass.edu.

Nianwen Xue and Yaqin Yang. 2013 *Dependency-based empty category detection via phrase structure trees.*. In Proceedings of NAACL HLT. Atlanta, Georgia.

Yaqin Yang and Nianwen Xue. 2010 *Chasing the ghost: recovering empty categories in the Chinese Treebank.*. In Proceedings of the 23rd International Conference on Computational Linguistics (COLING). Beijing, China.

Bing Xiang, Xiaoqiang Luo, Bowen Zhou. 2013. *Enlisting the Ghost: Modeling Empty Categories for Machine Translation*. In Proceedings of the ACL.

Converse, Susan 2006 *Pronominal anaphora resolution for Chinese.*. Ph.D. thesis.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993 *Building a large annotated corpus of english: The penn treebank.*. Computational Linguistics, 19(2):313–330.

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, 11(2):207–238.

Slav Petrov and Dan Klein. 2007. *Improved Inferencing for Unlexicalized Parsing*. In Proceedings of HLT-NAACL 2007.

Jing Peng and Kenji Araki. 2007 *Zero-Anaphora Resolution in Chinese Using Maximum Entropy.*. IEICE Transactions.