Dropped Personal Pronoun Recovery in Chinese SMS

Chris Giannella, Ransom Winder and Stacy Petersen[1]
The MITRE Corporation, 7515 Colshire Drive, McLean, VA 22102, USA
{cgiannella,rwinder,spetersen}@mitre.org

Abstract

In written Chinese, personal pronouns are commonly dropped when they can be inferred from context. This practice is particularly common in informal genres like Short Message Service (SMS) messages sent via cell phones. Restoring dropped personal pronouns can be a useful preprocessing step for information extraction. Dropped personal pronoun recovery can be divided into two subtasks: (1) detecting dropped personal pronoun slots and (2) determining the identity of the pronoun for each slot. We address a simpler version of restoring dropped personal pronouns wherein only the person numbers are identified. After applying a word segmenter, we used a linear-chain conditional random field (CRF) to predict which words were at the start of an independent clause. Then, using the independent clause start information, as well as lexical and syntactic information, we applied a CRF or a maximum-entropy classifier to predict whether a dropped personal pronoun immediately preceded each word and, if so, the person number of the dropped pronoun. We conducted a series of experiments using a manually annotated corpus of Chinese SMS messages. Our machine-learning-based approaches substantially outperformed a rule-based approach based partially on rules developed by Chung and Gildea in 2010. Features derived from parsing largely did not help our approaches. We conclude that the parse information is largely superfluous for identifying dropped personal pronouns if independent clause start information is available.

1. Introduction

Chinese is commonly characterized as a “pro-drop” language (Baran, Yang and Nianwen 2012), (Huang 1989) since pronouns are commonly dropped when they can be inferred from context. This practice is particularly common in informal genres like Short Message Service (SMS) messages sent via cell phones (Yang, Liu and Xue 2015). Dropped personal pronouns (I, we, you, etc.) complicate information extraction since entities are commonly assumed to be explicitly mentioned. Hence, extraction opportunities can be missed when pronouns are dropped. As an example, consider

[1] Also affiliated with The Dept. of Linguistics, Georgetown University, 3700 O Street NW, Washington DC USA.

Figure 1, below, and the task of relation detection. The message expresses a contact relation between John and the author of the message. A common approach to relation detection is to identify relations within sentences and establish co-reference chains allowing relations to be detected across sentences. However, such an approach will fail to detect the contact relationship in Figure 1 without first recovering “I”.

John went to school today. Met him in his office.	约翰去上学了。遇见他在他的办公室。
John went to school today. ***I*** met him in his office.	约翰去上学了。***我***遇见他在他的办公室。

FIGURE 1. The top row contains an example Chinese SMS message along with its English translation. The bottom row contains the messages with “I” recovered.

Little research has been dedicated to automated dropped-pronoun recovery in Chinese SMS. Only in the last year have studies been published that directly address the problem.

1.1 Problem Statement and Our Contributions

Dropped personal pronoun recovery can be divided into two subtasks: (1) detecting dropped personal pronoun slots and (2) determining the identity of the pronoun for each slot. We believe the complications to information extraction caused by dropped pronouns can largely be mitigated by recovering only person number. Hence, we address a simplified version of the problem herein: only the person numbers are identified. Specifically, given the raw[2] text of a Chinese SMS message, identify all characters that are immediately preceded by a dropped personal pronoun and determine whether the dropped pronoun is first or second person.[3] We ignore third person since the vast majority (98.3 percent) of dropped personal pronouns in our data were not third person.

After we applied a word segmenter to the message, we employed a two-stage strategy. In the first stage, a linear-chain conditional random field (CRF) was used to predict which words were at the start of an independent clause.[4] In the second stage, a label 0, 1, or 2 was assigned to each word: 0 if no dropped personal pronoun were predicted to immediately precede the word; 1 or 2 if a dropped first or second person pronoun were predicted to immediately precede the word. We developed two different algorithms for accomplishing the second stage. The first applied a linear-chain CRF to

[2] We do not assume the message is word or sentence segmented.

[3] As stated, the problem excludes dropped personal pronoun at the end of the message. We found this a rare occurrence and, for simplicity, ignore this case.

[4] An independent clause is a collection of words containing a subject (perhaps a dropped pronoun) and a predicate that modifies the subject and the clause can stand alone as a complete sentence. For example: “..., 我在动用大哥你吧。” (“..., I can still mentor you.”).

the sequence of all words in the message to assign labels. The independent clause start information and lexical and syntactic information were used to create features. The second algorithm applied a maximum entropy classifier to words close to independent clause starts—all other words were assigned label 0.

Through experiments on a manually annotated corpus of Chinese SMS messages, we examined the accuracy of our two approaches (one for each algorithm used in the second stage). We examined the impact of two different word segmenters and the impact of using parse-based features. We compared our approach to a rule-based baseline approach based partially on rules described in (Chung and Gildea 2010).

2. Data

On February 3, 2012, we downloaded 29,393 Chinese SMS messages from the National University of Singapore (Chen and Kan 2013). In October and November 2014, our MITRE colleague Dr. Sichu Li annotated the first 3,495 of these messages. In each message, she added all dropped personal pronouns and marked all the characters that were at the start of an independent clause; the beginning of each sentence (and message) was always marked as an independent clause start. In the example in Figure 1, above, there are two independent clause starts and one dropped personal pronoun.

As seen in Figure 2, below, dropped personal pronouns were common in our data set. 47.7 percent of the messages contained one or more dropped personal pronouns, while 14.1 percent contained two or more. Moreover, mid-message independent clause starts were also common. Some 35.3 percent of the messages contained two or more independent clause starts, while 20.7 percent contained three or more. Interestingly, some messages had a large number of independent clause starts and dropped personal pronouns. Despite the 160 character limit for SMS messages, some messages included many independent thoughts, expressed succinctly. The following is an example message with seven independent clause starts and two dropped personal pronouns: **我是做了三十三天的**预算，**在加上你的日志之后就**发现不够了。**一天吃**饭要花快二十。一共八百元！**哀~不**过**我**还不至于前胸贴后背，**等我不行了，我在**动用大哥你吧。(I have indeed done a thirty-three day budget, but found it not enough after taking into account your journal. One-day meals will cost about twenty yuan, a total of 800 yuan! It's a pity ~ but I will not be so hungry that my empty stomach will make my chest and back stick together, I am not good enough, but I can still mentor you.)

N	Dropped Personal Pronouns	Independent Clause Starts
0	1829	0
1	1174	2261
2	352	509
3	102	363
4	23	180
5	8	82
≥6	7	100

FIGURE 2. The second and third columns show the number of messages with N dropped personal pronouns and independent clause starts, respectively. Since the beginning of each message was marked as an independent clause start, no messages contained zero starts.

As seen in Figure 3, below, the dropped personal pronouns identified by our annotator were similar to the 10 non-abstract pronouns identified by annotators and described in (Baran, Yang and Nianwen 2012) and (Yang, Liu and Xue 2015).[5]

	Chinese	Count	English Translation	Person Number
Dropped pronouns also in Yang et al and Baran et al.	我	1103	I	1st
	我们	118	we	1st
	你	941	you (singular)	2nd
	你们	12	you (plural)	2nd
	他	18	he	3rd
	她	11	she	3rd
	它	4	it	3rd
	他们	3	they (masculine)	3rd
	她们	2	they (feminine)	3rd
	俺	3	I (non-Mandarin dialect)	1st
	本人	1	myself	1st
	我俩	1	us	1st
	您	2	you (singular, courteous)	2nd
	你两	2	you two	2nd

FIGURE 3: The dropped personal pronouns identified by our annotator, along with their frequencies in our data set.

Of the 2221 dropped personal pronouns our annotator identified, only 38 (1.7 percent) were third person. Hence, we ignored all third-person dropped pronouns.

3. Our Approach: Outline

Our approach involves a pipeline consisting of several steps. Some of these steps require statistical models to be trained from a set of Chinese SMS messages annotated

[5] Baran et al. and Yang et al. used different datasets than we did.

with independent clause starts and dropped personal pronouns. We split our annotated message corpus randomly into training (80 percent) and testing (20 percent) parts.

Training procedure:

(A). Apply a word segmenter to each message in the training part, ignoring those whose resulting segmentation conflicts with the annotations. Specifically, ignore a message if a dropped personal pronoun or independent clause start annotation occurs between two characters in the same word.

(B). Build a sequence labeler which, given a word segmented message, labels each word with true or false to indicate whether the word is predicted to be the start of an independent clause. Ignore dropped personal pronoun annotations.

(C). Build a sequence labeler which, given a word-segmented message labeled with independent clause starts, labels each word 0, 1, or 2: 0 if no dropped personal pronoun is predicted to immediately precede the word; 1 or 2 if a dropped first or second person pronoun is predicted to immediately precede the word.

Application procedure: Given a new Chinese SMS message, do the following.

(A). Apply a word segmenter to the message.

(B). Apply the independent clause start sequence labeler to add the independent clause start labels.

(C). Apply the dropped pronoun person number sequence labeler to add the dropped pronoun labels.

4. Our Approach: Details

In this section, we describe the details of each step of the pipeline.

4.1 Chinese Word Segmentation

In our experiments, we compared two Chinese word segmenters. The first was the Stanford Chinese word segmenter[6] version 1.6.7 using the "ctb" model (Manning, et al. 2014). This segmenter was not specifically designed for SMS messages. The second was a segmenter specifically designed for Chinese SMS messages (Wang, et al. 2012) which we refer to as the "UMAC" segmenter. The segmenters produced substantially different results: 84 percent of the messages were segmented differently.

4.2 Independent Clause Start Prediction

[6] http://nlp.stanford.edu/software/segmenter.shtml

We built a linear-chain CRF (Lafferty, McCallum and Pereira 2001) model from the training part, using a mean zero Gaussian prior (ignoring the dropped personal pronoun annotations). The result was a sequence labeler which, given a new word segmented message, labeled words that were predicted to be the start of an independent clause. We used the following features.

- All word unigrams from two words to the left to two words to the right of the current word. The word bigrams consisting of the current word and the word to the left and the word to the right.
- All character unigrams and bigrams that are a prefix or suffix of the current word.

In training, the Gaussian prior was set using a random pure-training (80 percent) and development (20 percent) split of the training part. The prior variance was ranged over {0.3, 0.5, 0.7, 0.9, 1, 5, 10, 50, 100}. For each value, a CRF model was built using the pure-training part and applied to the development, producing an F score. Let γ denote the prior value that produced the largest F score. The final CRF model was built using the full-training part with the Gaussian prior variance set to γ.

It should be noted that for both training and application, we used the jcarafe Java library which is freely available from Sourceforge.[7] The problem of Chinese independent clause prediction is similar to that of Chinese comma classification as addressed in (Xue and Yang 2011). However, significant differences between the two problems exist, for example, independent clauses can be separated without punctuation.

4.3 Dropped Pronoun Person Number Prediction

We considered two different approaches to building a dropped pronoun person number sequence labeler from the training part.

4.3.1 CRF Approach

This approach utilizes a linear-chain CRF to label each word with a 0, 1, or 2. We used all the features described in Subsection 4.2, plus the following.

- [ICS features] A Boolean feature set to "true" if an independent clause break was three or fewer words to the left of the current word.

[7] http://sourceforge.net/projects/carafe/files/jcarafe/

- [POS features] All part-of-speech (POS) tags assigned to words two to the left through two to the right of the current word. We used the Stanford POS tagger[8] version 3.5.2, using the "chinese-distsim" model (Manning, et al. 2014) in our experiments. The POS bigram consisting of the POS assigned to the current word and to the word to the left. The POS bigram consisting of the POS assigned to the current word and to the word to the right.
- [Parse-based features] All the constituency parse patterns that are satisfied at the current word; details can be found in the Appendix.

In training, the Gaussian prior was set as described in Subsection 4.2.

4.3.2 Maximum-Entropy Classifier Approach

A classifier is applied to some of the words in messages to assign dropped personal pronoun labels. Choosing the words to which the classifier is applied is the key to this approach. As seen in Figure 4, below, the vast majority of dropped personal pronouns have an independent clause start within three words to the left: 95.2 percent when the Stanford word segmenter was used; 97.8 percent when the UMAC segmenter was used. Hence, the approach classifies only words with an independent clause start near to the left, all other words are assigned label 0.[9]

N	Stanford	UMAC
1	1577	1131
2	132	111
3	66	50
4	45	16
≥5	44	13

FIGURE 4: The second (third) column contain the numbers of dropped personal pronouns in the training part that had an independent clause start N words to the left when the Stanford (UMAC) word segmenter was used.

Specifically, the approach assigns dropped personal pronoun labels to a new, word-segmented, Chinese SMS message as follows. Let <W[1],W[2],…,W[m]> denote the sequence of words in the message, ordered from left to right.

(A). Assign all words label 0.

[8] http://nlp.stanford.edu/software/tagger.shtml

[9] Dropped pronouns not the subject of the independent clause they inhabit will tend to be missed by this approach. However, only 1.8 percent of the dropped pronouns in the training part were not subjects, so these type of misses seems acceptable.

(B). Apply the independent clause start sequence labeler to <W[1],W[2],…,W[m]> and let <W[i_1],W[i_2]..,W[i_s]> denote the subsequence of words predicted to have an independent clause start immediately preceding.

(C). For j = 1 to s, do

   a. For k = 0 to 3, apply the maximum-entropy classifier to W[i_j+k], if it returns 1 or 2, then change the label on W[i_j+k] accordingly and terminate the "For k = 0 to 3…" loop.

To build the classifier, we first construct a set of labeled, Boolean training vectors from the training part (after a word segmenter is applied). For each word that our annotator marks as being the start of an independent clause, we consider that word and the three words to the right. If none of these words is annotated as having a preceding dropped personal pronoun, then a training vector is built from each of these words and the label on each vector is set to 0. Otherwise, a training vector is built from the left-most word annotated as having a preceding dropped personal pronoun and its label is set to 1 (2) if the dropped pronoun was first (second) person. Each training vector has entries that indicate the presence or absence of the following features.

- [Word features] All word unigrams, bigrams, and trigrams drawn from the words two to the left through two to the right of the current word.
- [POS features] All POS unigrams, bigrams, and trigrams drawn from the POS tags assigned to words three to the left through three to the right of the current word.
- [Parse-based features] All the constituency parse patterns that are satisfied at the current word; details can be found in the Appendix.

A mean zero, Laplace prior is used in building the maximum-entropy classifier. The prior variance was set in the similar fashion as described in Subsection 4.2. We used the MALLET version 2.0.7 Java library (McCallum 2002) in our experiments.

5. Experiments

We compared our approaches DP-CRF and DP-Classifier to a baseline approach. DP-CRF refers to our pipeline using the CRF-based dropped pronoun person number prediction step. DP-Classifier refers to our pipeline using the maximum entropy classifier-based dropped pronoun person number prediction step. The next subsection describes details of the baseline approach; the subsequent subsection describes the methodology we used to make the comparison.

5.1 Baseline Approach

This is a rule-based approach utilizing constituency parse information similar to an approach described in (Chung and Gildea 2010). Their rules were not developed for informal Chinese, and we found them to work poorly by themselves on our messages. As such, we added more rules. Our experiments showed that the addition of our rules resulted in an improvement in accuracy over the rules of Chung and Gildea alone. Our rules all target the subject position in clauses with different syntactic structures. Syntactically, a subject occurs in the specifier position of an intonational phrase (IP), but occasionally additional syntactic information will be present to the left of the subject, such as additional modifiers. Hence, the left-most linear position cannot be relied upon as the position where a subject pronoun is dropped.

The baseline approach first applies the Stanford constituency parser[10] version 3.5.2 using the "chinesePCFG.ser" model (Manning, et al. 2014) to the word segmented message. Then, for each word, assigns label 0 if none of the following rules hold. If at least one of the rules holds, then the word is assigned a label 1 or 2, randomly.

(A). Any rule listed in the top right of Figure 5 in (Chung and Gildea 2010).

(B). The word is left-most in a verb phrase (VP) that (1) has an IP parent; (2) does not have an immediate left noun phrase (NP sister); (3) does not have an only child that is an adjective, punctuation mark, quantifier phrase, NP, or verb; and (4) does not have a verb descendent that has a right sister NP which has a pronoun, proper noun, or complementizer phrase descendent. This rule was implemented using the Tregex[11] (Levy and Galen 2006) regular expression.

- o (VP>IP&(!$-NP)&(!<:VA|PU|QP|NP|VV)&(!<(VV$+(NP<(PN|NR|CP)))))

This rule captures dropped pronouns (as part of a full IP) at the start of a verb phrase, that does not have an NP subject (2) and there are not any intervening elements of the type: adjective, punctuation mark, quantifier phrase, noun phrase, or verb (3). Finally, part (4) prevents the occasional error in which a pronoun, proper noun or complementizer phrase is mislabeled by the tagger as having a VV parent, and therefore a VP parent, but is not in fact a VP, and therefore cannot have a subject.

[10] http://nlp.stanford.edu/software/lex-parser.shtml

[11] http://nlp.stanford.edu/software/tregex.shtml

(C). The word is left-most in a VP phrase that (1) has a VP parent; (2) does not have an immediate left NP sister; (3) has a punctuation mark as a left sister; and (4) does not have an only child that is an adjective, punctuation mark, quantifier phrase, NP, or verb.

- `(VP>VP&(!$-NP)&$-PU&(!<:VA|PU|QP|NP|VV))`

This rule captures dropped pronouns at the start of a clause that is not necessarily a full IP.

(D). The word is left-most in a VP phrase that (1) has a VP parent; (2) does not have an immediate left NP sister; and (3) has a BA descendent with a right sister IP.

- `(VP>VP&(!$-NP)&<(BA$+IP)`

This rule captures dropped pronouns in a VP that is nested within another VP that does not already have a NP subject (2) and is a ba construction. This construction is indicated by the word 把 used to focus on the result or influence of an action. A dropped pronoun may occur before the ba, which is what this rule captures. Chinese ba sentences occur with certain action verbs and the normal SVO word order is changed to: Subject+“ba”+Object+Verb phrase.

(E). The word is left-most in an IP phrase that has a prepositional phrase (PP) descendent with an adverb phrase (ADVP) left sister.

- `(IP<(PP$-ADVP))`

This rule captures dropped pronouns before an adverb phrase in order detect a subjectless VP further right in the tree, and not predicted by the above rules.

5.2 Methodology

The baseline approach, DP-CRF, and DP-Classifier were all applied to the messages in the test part[12] and precision, recall, and F scores computed. Several variations on our approaches were compared: with and without parse-based features and using the Stanford word segmenter versus the UMAC segmenter.

For all approaches, scores were computed at the character level of granularity. Specifically, let $M_1,M_2,\ldots,M_n$ denote the messages in the test part; let c_i denote the number of characters in M_i; and let $M_i[j]$ denote the j^{th} character in M_i. All approaches

[12] In the test part, our annotator added back 408 dropped personal pronouns.

operate at the word level of granularity. Hence, each approach assigns 0 to $M_i[j]$ if $M_i[j]$ is not the first character in a word. Otherwise, the label assigned to $M_i[j]$ is the label assigned to the word containing $M_i[j]$. We carried out the following procedure to compute the precision, recall, and F score of labels 1 and 2 for each approach.

(A). Set TP[1], FP[1], FN[1], TP[2], FP[2], and FN[2] to 0.

(B). For i = 1 to n, do

- a. For j = 1 to c_i, do
  - 1. For L = 1 to 2, do
    - (I.) If the approach assigned label L to $M_i[j]$ and our annotator added an L^{th} person dropped pronoun immediately before $M_i[j]$, then TP[L] is incremented.
    - (II.) If the approach assigned label L to $M_i[j]$ and our annotator did not add an L^{th} person dropped pronoun immediately before $M_i[j]$, then FP[L] is incremented.
    - (III.) If the approach did not assign label L to $M_i[j]$ and our annotator added an L^{th} person dropped pronoun immediately before $M_i[j]$, then FN[L] is incremented.

(C). For L = 1 to 2, do

- a. Compute precision = TP[L]/(TP[L]+FP[L]), recall = TP[L]/(TP[L]+FN[L[), and F = 2*precision*recall/(precision+recall).[13]

To evaluate the statistical significance between F scores we used permutation tests as discussed in (Edington and Onghena 2007) and (Yeh 2000). Specifically, consider two approaches and let $\lambda^1_i[j]$ and $\lambda^2_i[j]$ denote the label assigned to $M_i[j]$ by approach one and two, respectively. Let $F_L\{\lambda^1_1,\ldots,\lambda^1_n\}$ and $F_L\{\lambda^2_1,\ldots,\lambda^2_n\}$ denote the F score, with respect to person number L, of approach one and two, respectively. The procedure below computes, Sig, the significance level at which the following null hypothesis can be rejected: the F score of approach one is less than or equal to approach two.

(A). Let numF = 0 and do the following 100,000 times.

- a. For i = 1 to m, with probability 0.5, let $\widehat{\lambda}^1_i[j]=\lambda^1_i[j]$ and $\widehat{\lambda}^2_i[j]=\lambda^2_i[j]$ for all $1\leq j\leq c_i$. Alternatively, with probability 0.5, let $\widehat{\lambda}^1_i[j]=\lambda^2_i[j]$ and $\widehat{\lambda}^2_i[j]=\lambda^1_i[j]$ for all $1\leq j\leq c_i$.
- b. If $F_L\left\{\widehat{\lambda}^1_1,\ldots,\widehat{\lambda}^1_n\right\} > F_L\left\{\widehat{\lambda}^2_1,\ldots,\widehat{\lambda}^2_n\right\}$, then increment numF.

(B). Return Sig = 1–(numF/100,000).

[13] If TP[L]+FP[L] is zero, then precision is defined to be one; if TP[L[+FN[L] is zero, then recall is defined to be zero; if precision+recall is zero, then F is defined to be zero.

6. Results

APPROACH	PARSE-BASED FEATURES?	WORD SEGMENTER	1st PERSON PREC	1st PERSON REC	1st PERSON F	2nd PERSON PREC	2nd PERSON REC	2nd PERSON F
DP-CRF	No	Stanford	0.41	0.30	0.34	0.60	0.34	0.43
		UMAC	0.39	0.22	0.28	0.52	0.24	0.33
	Yes	Stanford	0.49	0.33	0.39	0.57	0.31	0.40
		UMAC	0.40	0.19	0.26	0.51	0.29	0.37
DP-Classifier	No	Stanford	0.38	0.31	0.34	0.49	0.34	0.40
		UMAC	0.40	0.25	0.31	0.50	0.25	0.33
	Yes	Stanford	0.37	0.31	0.34	0.50	0.32	0.39
		UMAC	0.38	0.23	0.29	0.51	0.26	0.35
DP-Baseline	--	Stanford	0.09	0.28	0.14	0.08	0.35	0.14
		UMAC	0.08	0.19	0.11	0.07	0.20	0.10

FIGURE 5: The table displays an accuracy comparison between DP-CRF, DP-Classifier, and DP-Baseline. The second column indicates whether parse-based features were used by DP-CRF or DP-Classifier. The third column indicates which word segmenter was used.

APPROACH	PARSE-BASED FEATURES?	WORD SEGMENTER	1st PERSON F DIFFERENCE SIGNIFICANCE LEVEL	2nd PERSON F DIFFERENCE SIGNIFICANCE LEVEL
DP-CRF	No	Stanford **vs.** UMAC	0.95	0.99
	No **vs.** Yes	Stanford	0.99	0.80
DP-Classifier	No	Stanford **vs.** UMAC	0.85	0.97
	No **vs.** Yes	Stanford	0.60	0.63
DP-CRF **vs.** DP-Classifier	No	Stanford	0.50	0.84
	Yes	Stanford	0.99	0.55
	No	UMAC	0.70	0.49
	Yes	UMAC	0.71	0.84

FIGURE 6: The table displays statistical significances of the difference between some of the F scores in Figure 5. The second column indicates whether parse-based features were used. The third column indicates which word segmenter was used. For example, the first row concerns the F score differences between DP-CRF without parse-based features using the Stanford versus the UMAC word segmenters. The row shows the levels at which the 1st person and 2nd person differences were found to be statistically significant.

Several observations can be made from Figures 5 and 6, above.

- The Stanford segmenter typically resulted in more accurate dropped pronoun recovery than the UMAC segmenter. In five out of eight cases, the Stanford segmenter resulted in larger F scores significant at the 0.95 level. This is a surprising result given that the UMAC segmenter was specifically designed for Chinese SMS while the Stanford segmenter was not.
- DP-CRF and DP-Classifier had very similar F scores. In only one case was the difference in F scores significant at a level greater than 0.85.
- The parse-based features typically did not help DP-Classifier. In only one case was the difference in F scores significant at a level greater than 0.8. We conclude that the parse information is largely superfluous for identifying dropped personal pronouns, once independent clause start predictions have been made.
- The machine-learning approaches (DP-CRF and DP-Classifier) substantially outperformed the rule-based approach (DP-Baseline) in terms of F score.

6.1 The Effect of ICS Features

Since dropped Chinese pronouns tend to occur near the start of independent clauses, we believe that the ICS features are important to both DP-CRF and DP-Classifier. Hence, to better understand the effect of ICS features on overall error, we examined the impact of modifying DP-CRF and DP-Classifier to use an oracle ICS predictor and the impact of modifying DP-CRF to not use ICS features at all. Two conclusions can be drawn from Figures 7 and 8, below.

- Our use of ICF features made a substantial difference in the recovery of 1st person pronouns but not in the recovery of 2nd person pronouns. Indeed, the second and third rows of Figure 7 show that the F scores of DP-CRF with predicted ICS (and using the Stanford word segmenter) were larger than no ICS features. But, the first row of Figure 8 shows that the F score advantages were only significant at a high level for first person recovery.
- Devoting more effort to improve the ICS predictor could be helpful in improving first person pronoun recovery. As seen in Figure 7, Oracle ICS features result in large F scores for both DP-CRF and DP-Classifier (when using the Stanford word segmenter). And, as seen in the second and fourth rows of Figure 8, the F score advantages were significant at a 0.95 level.

APPROACH	WORD SEGMENTER	ICS FEATURES	1st PERSON PREC	1st PERSON REC	1st PERSON F	2nd PERSON PREC	2nd PERSON REC	2nd PERSON F
DP-CRF	Stanford	Oracle	0.46	0.34	0.39	0.57	0.37	0.45
		Predicted	0.41	0.30	0.34	0.60	0.34	0.43
		None	0.39	0.23	0.29	0.62	0.31	0.41
	UMAC	Oracle	0.33	0.25	0.29	0.47	0.28	0.35
		Predicted	0.39	0.22	0.28	0.52	0.24	0.33
		None	0.35	0.19	0.25	0.56	0.23	0.32
DP-Classifier	Stanford	Oracle	0.42	0.33	0.37	0.47	0.37	0.41
		Predicted	0.38	0.31	0.34	0.49	0.34	0.40
	UMAC	Oracle	0.37	0.27	0.31	0.46	0.25	0.32
		Predicted	0.40	0.25	0.31	0.50	0.25	0.33

FIGURE 7: The table displays accuracies of our approaches using ICS features assigned by the annotator (Oracle), ICS features assigned using an ICS predictor (Predicted), and not using ICS features at all (None). In all cases, parse-based features were not used.

APPROACH	ICS FEATURES	1st PERSON F DIFFERENCE SIGNIFICANCE LEVEL	2nd PERSON F DIFFERENCE SIGNIFICANCE LEVEL
DP-CRF	None **vs.** Predicted	0.97977	0.68462
	Predicted **vs.** Oracle	0.9966	0.80827
	None **vs.** Oracle	0.9998	0.85172
DP-Classifier	Predicted **vs.** Oracle	0.95632	0.73823

FIGURE 8: The table displays statistical significances of the difference between some of the F scores in Figure 7. In all cases, parse-based features were not used and the Stanford word segmenter was used.

6.2 Dropped Pronoun Slot Prediction

To further understand the overall error, we examined the accuracy of our approaches when distinguishing between pronoun numbers is not required. We modified our methods to detect only the dropped pronoun slots, without recovering the person number of the dropped pronouns. We simply collapsed the labels 1 and 2 in Subsection 4.2 to a single label. We denote the modified approaches DPslot-CRF, DPslot-Classifier, and DPslot-Baseline. As seen in Figure 9, below, the F scores increase substantially, (more than 35 percent). However, even detecting dropped pronoun slots is challenging with F scores not exceeding 0.6.

APPROACH	PARSE-BASED FEATURES?	WORD SEGMENTER	PREC	REC	F
DPslot-CRF	--	Stanford	0.66	0.52	0.58
		UMAC	0.64	0.38	0.48
DPslot-Classifier	No	Stanford	0.64	0.56	0.60
		UMAC	0.58	0.41	0.48
	Yes	Stanford	0.63	0.56	0.59
		UMAC	0.58	0.39	0.47
DPslot-Baseline	--	Stanford	0.18	0.59	0.28
		UMAC	0.18	0.45	0.26

FIGURE 9: The table displays an accuracy comparison between DPslot-CRF, DPslot-Classifier, and DPslot-Baseline. These approaches predict only dropped personal pronoun slots—no person number prediction is made.

6.3 Other Error Analysis

To further understand the overall error, we examined three questions.

(i) How effective is our approach at recovering pronouns dropped away from an ICS?

(ii) How prevalent is each type of prediction error, e.g. predicted 2nd when the annotator assigned 1st?

(iii) What characters most frequently appear immediately to the right of dropped pronoun recovery errors?

To address the first question, we compared DP-CRF with and without parse-based features at recovering dropped pronouns at least four characters to the right of an independent clause start. Comparing Figure 10 with the first and third rows of Figure 5 shows that dropped pronouns away from ICSs are more difficult to recover: F scores less than 0.13 for recovering dropped pronouns away from ICSs versus F scores greater than 0.3 for recovering all dropped pronouns. And, parse-based features offer little help in recovering pronouns away from ICSs.

PARSE-BASED FEATURES?	1st PERSON PREC	1st PERSON REC	1st PERSON F	2nd PERSON PREC	2nd PERSON REC	2nd PERSON F
No	0.120	0.115	0.118	0.250	0.071	0.111
Yes	0.143	0.115	0.128	0.200	0.071	0.106

FIGURE 10: The table displays the scores of DP-CRF, with and without parse-based features, with predicted ICS features, and using the Stanford word segmenter. The scores are computed only for slots that appear at least four characters to the right of an ICS assigned by the annotator.

To address the second question, we computed the prevalence of each type of prediction error as seen in Figure 11. As can be seen, the preponderance of prediction errors were false negatives, particularly for 1st person dropped pronoun recovery.

ANNOTATOR LABEL	PREDICTED LABEL	FRACTION
1st	2nd	0.09
	Null	0.61
	1st	0.3
2nd	1st	0.2
	Null	0.46
	2nd	0.34

FIGURE 11: The table refers to DP-CRF with no-parse features and using the Stanford word segmenter. Displayed are the fraction of times a label assigned by the annotator received a predicted label (null means that no predictor label was received). For example, the first row indicates: of the total number of times the annotator assigned label "1st", nine percent were predicted as "2nd".

To address the third question, we tabulated the characters that most frequently occur immediately to the right of slots where an incorrect dropped pronoun recovery was made (translations were obtained from Google Translate on April 7, 2016). The tabulation is depicted in Figure 12. The six characters shown account for 13 percent of the cases where an annotator added a dropped pronoun was not correctly recovered.

	ANNOTATOR ASSIGNED LABEL		
CHARACTER	1st	2nd	null
还 (also)	9	4	2
不 (do not, not)	5	4	4
一 (one, a)	7	0	4
在 (in, at)	5	4	2
今 (this, now, today)	3	3	4
就 (on)	5	3	2

FIGURE 12: The table refers to DP-CRF with no-parse features and using the Stanford word segmenter. The first column contains characters that appeared immediately to the right of slots where an incorrect dropped pronoun recovery was made. Displayed are the number of times the annotator assigned label was incorrectly predicted ("null" means that the annotator assigned no label) categorized by the character immediately after the slot. Characters whose total (sum of the three counts shown) is not greater than nine are not displayed.

7. Related Work

Our focus is limited to literature discussing computational approaches to the automatic resolution of dropped pronouns (and closely related problems) in Chinese. The reader is referred to (Huang 1989) for a detailed discussion of Chinese pronoun dropping from a theoretical linguistics perspective and to (Seki, Fujii and Ishikawa 2002), (Kawahara and Kurohashi 2005) and (Sasano and Kurohashi 2011) for computational approaches to dropped pronoun resolution (and related problems) in another language, Japanese.

We discuss literature in two groups: (1) Chinese empty category detection and (2) Chinese dropped (zero) anaphora resolution and closely related efforts.

7.1 Chinese Empty Category Detection

Dropped pronouns are but one type of empty categories. Other types include null elements in control constructions (*PRO*), traces of A movement, and the like (Xue, Xia, et al. 2000). The problem of automatically detecting empty categories involves identifying all the words in a document that are immediately preceded by one or two empty categories (unlike pronouns, multiple empty categories can immediately precede a word). Several studies have addressed this problem (or a simpler variant) on formally written Chinese text such as newswire (assumed to be word segmented and sentence-split).

Yang and Xue (2010) addressed a somewhat simpler problem: detect whether an empty category immediately precedes each word but not which category. They trained a maximum entropy classifier and applied it independently to each word. They utilized lexical and parse-based features and found the parse-based features to substantially improve accuracy for empty categories in a subject position.

Chung and Gildea (2010) addressed a somewhat simpler problem: detect whether a dropped pronoun or *PRO* (and which of the two) immediately precedes each word. They developed and compared a rule-based approach, a CRF approach, and an approach based on training a parser from manually produced parses including empty categories. They found the rule-based approach to be most effective at detecting dropped *PRO*s and the CRF at detecting dropped pronouns.

Cai, Chiang, and Goldberg (2011) trained the Berkley constituency parser on manually produced parse trees with empty categories included. They applied the parser on word lattices for each sentence. Since zero, one, or two empty categories may appear before any word, the lattice allows zero, one, or two empty category markers to appear immediately before any word. The resulting tree has words and empty category markers as terminals, and the empty category markers have a single parent specifying the specific empty category that was missing.

Kong and Zhou (2013) developed a method that recursively applies a "linear tagger" approach: each word is tagged with a single empty category or none, to various parts of the sentence. The recursion is based on the authors' definition of clause and sub-clause structure as defined from a constituency parse tree.

Xue and Yang (2013), (2014) observed that classifying individual words with a single empty category (or none) will miss cases where two empty categories immediately precede a word. To mitigate this shortcoming, these authors utilized dependency parses and classified all pairs of words and heads in each sentence (multiple empty categories that immediately precede the same word will have different heads).

7.2 Chinese Dropped Anaphora Resolution and Closely Related Efforts

Dropped (zero) anaphora resolution involves identifying all dropped noun phrase slots in a document and, for each dropped phrase, identifying its antecedent or determining that one does not exist. Pronoun resolution is a special case where the dropped noun phrase is restricted to being a pronoun. Several studies have addressed dropped anaphora and dropped pronoun resolution in formally written Chinese text such as newswire. All the studies we discuss assume the text is word segmented.

Yeh and Chen (2007) developed a rule-based procedure (using shallow parses of each sentence) to address dropped anaphora resolution. These authors used centering theory (Grosz, Joshi and Weinstein 1995) in the design of the rules component that identifies antecedents. Kong and Zhou (2010) divided dropped anaphora resolution into three subtasks: (1) dropped anaphora slot detection, (2) anaphorisity determination (determine whether a dropped anaphora has an antecedent), and (3) antecedent identification. They designed approaches for all three subtasks on the basis of tree-kernel support vector machines.

Zhao and Ng (2007) addressed dropped pronoun resolution. They used a simple rule-based procedure (based on full constituency parses) to identify dropped pronoun slots and to identify candidate antecedents. They developed a method, using a decision tree classifier, to assign dropped pronouns to antecedents. Yang, Dai, and Cui (2008) employed a similar approach, except they used a more sophisticated rule-based approach (based on verbal logic valence theory) to identify dropped pronoun slots. Chen and Ng (2013) extend Zhao and Ng's approach by utilizing more complex features and co-reference links between dropped pronouns.

Chen and Ng (2014) developed an approach to Chinese dropped pronoun resolution that does not require a training set with manually added dropped pronouns and antecedents identified. They utilize explicit pronouns to train an approach for resolving dropped pronouns.

Recently, a different version of dropped pronoun resolution has been addressed in Chinese SMS messages. In this version of the problem (which we call dropped pronoun recovery), the dropped pronouns are to be automatically restored without necessarily linking to an antecedent, if one exists. Yang, Liu, and Xue (2015) trained a 17-class maximum entropy classifier to assign words to one of 16 types of dropped pronouns or "none." The class indicated which, if any, dropped pronoun is predicted to immediately precede the word. Their classifier used lexical, part-of-speech-based, and parse-based features. They applied their classifier separately to each word in each testing message. Rao, Ettinger, Daume III, and Resnik (2015) addressed a simpler version of dropped pronoun recovery in which only the person number (first, second, or third) of the dropped pronouns need be restored. They used the approach in (Cai, Chiang and Goldberg 2011) to identify dropped pronoun slots. Motivated by centering theory, Rao et al. trained a sequence labeler which, given an SMS dialogue between multiple communicants, jointly assigns a focus label[14] to each message and a person number to each dropped pronoun. Instead of training based on manually annotated Chinese SMS dialogues, Rao et al. utilize a parallel Chinese/English SMS corpus whose English side has been manually annotated with recovered dropped pronouns. As such, they avoid the difficulty of manually annotating Chinese SMS messages. Utilizing a parallel corpus in this fashion is similar in spirit to a large number of studies that have attempted to build foreign language natural language processing tools by projecting information across a parallel corpus. In particular, the work of Rahman and Ng (2012) addresses co-reference resolution in this fashion.

7.3 Differences Between Our Research and the Literature

Most of the literature addresses problems related to Chinese dropped pronoun detection and recovery in formally written text. Given the substantial difference between formally written text and SMS messages, further study is warranted on SMS. Very recently, two studies (Rao, et al. 2015) and (Yang, Liu and Xue 2015) have addressed Chinese dropped personal pronoun recovery in SMS messages. These studies were conducted simultaneously to, and independently of, our research.

Rao et al. used dialogue-derived information to address a problem very similar to the one we addressed (Rao et al. included the third person label). Given that our data set did not contain SMS dialogue threads, we focused only on information that could be derived from messages treated individually. Extending our approach to utilize ideas from Rao et al. is a good direction for future work.

The primary differences between our research and that of Yang et al. are as follows:

1. We addressed a simpler problem where only the person number of dropped pronouns need be recovered. Yang et al. address the full dropped pronoun recovery problem—the specific pronouns that were dropped need be recovered.
2. We considered a CRF in addition to a word classifier. Yang et al considered only a word classifier.
3. We predicted independent clause breaks (without requiring a parser) and used them as features in our CRF. Yang et al. essentially used a full parser to extract this kind of information to use as features in their classifier.
4. Our word classifier is not applied to all words (as in Yang et al.), just those close to independent clause breaks.

[14] 1 when the focus is on the writer of the message. 2 when the focus is on another person in dialogue with the writer.

5. We examined the impact of different word segmenters on dropped pronoun recovery accuracy.

8. Conclusion

In summary, we addressed a simplified version of the problem of dropped pronoun recovery detection in Chinese SMS messages. After applying a word segmenter, we used a CRF to predict which words are at the start of an independent clause. Then, using the independent clause start information and lexical and syntactic information, we applied a CRF or a maximum-entropy classifier to predict whether a dropped personal pronoun immediately preceded each word and, if so, the person number of the dropped pronoun. We conducted a series of experiments using a manually annotated corpus of Chinese SMS messages. Our machine-learning-based approaches substantially outperformed a rule-based approach based partially on rules developed by Chung and Gildea (2010). Features derived from parsing largely did not help our approaches. We conclude that the parse information is largely superfluous for identifying dropped pronouns once independent clause start prediction has been made. Finally, ICS features were only significantly helpful at recovering 1st person pronouns and devoting more effort to improve the ICS predictor could be helpful in that case.

One avenue of future work is to use our dropped pronoun slot detection approach followed by the approach in (Rao, et al. 2015) to assign pronoun identity information to the slots. To do this, we would need a different data set that contains SMS dialogue threads. Another avenue of future work is to distinguish anaphoric dropped personal pronouns from non-anaphoric ones. Such a distinction may be useful given that anaphoric pronouns are typically more useful in downstream applications like information extraction.

Acknowledgements

We are thankful for the assistance provided by our MITRE colleagues. Dr. Sichu Li annotated, in efficient and professional fashion, a large subset of the SMS messages we downloaded from the National University of Singapore. Dr. John Prange, Mr. Rob Case, and Mr. Rod Holland provided valuable feedback on a presentation we gave describing our preliminary research findings.

We are also thankful for the assistance provided by our colleagues at other institutions. Professor Nianwen (Bert) Xue at Brandeis University, Boston, USA shared his thoughts and expertise on Chinese dropped pronoun detection, at an early stage of our research. Professor Derek F. Wong and Mr. Junwen Xing at the University of Macau, Macau, SAR PRC applied their word segmenter to the National University of Singapore corpus.

Bibliography

Baran, E., Yang, Y., and Xue, N. 2012. Annotating Dropped Pronouns in Chinese Newswire Text. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC).* European Language Resources Association (ELRA). pp. 2795-9.

Cai, S., Chiang, D., and Goldberg, Y. 2011. Language-Independent Parsing with Empty Elements. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL).* Association for Computational Linguistics. pp. 212-6.

Chen, C. and Ng, V. 2014. Chinese Zero Pronoun Resolution: An Unsupervised Approach Combining Ranking and Integer Linear Programming. *Proceedings of the 28th AAAI Conference on Artificial Intelligence.* Association for the Advancement of Artificial Intelligence Press. pp. 1622-8.

Chen, C. and Ng, V. 2013. Chinese Zero Pronoun Resolution: Some Recent Advances. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).* Association for Computational Linguistics. pp. 1360-5.

Chung, T. and Gildea, D. 2010. Effects of Empty Categories on Machine Translation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).* Association for Computational Linguistics. pp. 636-45.

Edington, E., and Onghena, P. 2007. Randomization Tests (4th ed.). Boca Raton, FL, USA: CRC Press, Taylor & Francis Group. ISBN: 978-1-58488-589-4.

Grosz, B., Joshi, A., and Weinstein, S. 1995. Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics* 21 (2): pp. 203-25.

Huang, C. T. J. 1989. Pro-Drop in Chinese: A Generalized Control Theory. Vol. 15, in *Studies in Natural Language and Linguistic Theory: The Null Subject Parameter*, by Jaeggli, O. and Safir, K., pp. 185-214. Springer Netherlands. doi:10.1007/978-94-009-2540-3_6.

Kawahara, D. and Kurohashi, S. 2005. Zero Pronoun Resolution Based on Automatically Constructed Case Frames and Structural Preference of Antecedents. Vol. 3248, in *Lecture Notes in Computer Science*, edited by Su, K.-Y., Tsujii, J., Lee, J.-L., and Kwong, O.Y., pp. 12-21. Springer Berlin Heidelberg. doi:10.1007/978-3-540-30211-7_2.

Kong, F. and Zhou, G. 2010. A Tree Kernel-Based Unified Framework for Chinese Zero Anaphora Resolution. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).* Association for Computational Linguistics. pp. 882-91.

Kong, F. and Zhou, G. 2013. A Clause-Level Hybrid Approach to Chinese Empty Element Recovery. *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI).* Association for the Advancement of Artificial Intelligence Press. pp. 2113-9.

Lafferty, J., McCallum, A., and Pereira, F. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proceedings of the 18th International Conference on Machine Learning (ICML).* Morgan Kaufmann. pp. 282–9.

Levy, R. and Galen, A. 2006. Tregex and Tsurgeon: Tools for Querying and Manipulating Tree Data Structures. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC).* European Language Resources Association (ELRA). pp. 2231-4.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., and McClosky, D. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL).* Association for Computational Linguistics. pp. 55-60.

McCallum, A. 2002. Accessed July 16, 2013. http://mallet.cs.umass.edu.

Rahman, A. and Ng, V. 2012. Translation-Based Projection for Multilingual Coreference Resolution. *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL).* Association for Computational Linguistics. pp. 1051-60.

Rao, S., Ettinger, A., Daume III, H., and Resnik, P. 2015. Dialogue Focus Tracking for Zero Pronoun Resolution. *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL).* Association for Computational Linguistics. pp. 494-502.

Sasano, R. and Kurohashi, S. 2011. A Discriminative Approach to Japanese Zero Anaphora Resolution with Large-scale Lexicalized Case Frames. *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP).* Association for Computational Linguistics. pp. 758–66.

Seki, K., Fujii, A., and Ishikawa, T. 2002. A Probabilistic Method for Analtyzing Japanese Anaphora Integrating Zero Pronoun Detection and Resolution. *Proceedings of the 19th International Conference on Computational Linguistics (COLING).* Association for Computational Linguistics. pp. 1-7.

Tao, C. and Kan, M.-Y. 2013. Creating a Live, Public Short Message Service Corpus: the NUS SMS Corpus. *Language Resources and Evaluation* (Springer Netherlands) 47 (2): pp. 299-335. doi:10.1007/s10579-012-9197-9.

Wang, L., Wong, D., Chao, L., and Xing, J. 2012. CRFs-Based Chinese Word Segmentation for Micro-Blog with Small-Scale Data. *Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing.* Association for Computational Linguistics. pp. 51-7.

Xue, N. and Yang, Y. 2013. Dependency-Based Empty Category Detection Via Phrase Structure Trees. *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL).* Association for Computational Linguistics. pp. 1051-60.

Xue, N. and Yang, Y. 2011. Chinese Sentence Segmentation as Comma Classification. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL).* Association for Computational Linguistics. pp. 631-5.

Xue, N., Xia, F., Huang, S., and Kroch, A. 2000. The Bracketing Guidelines for the Penn Chinese Treebank (3.0). Technical Report No. IRCS-00-08, University of Pennsylvania Institute for Research in Cognitive Science . http://repository.upenn.edu/ircs_reports/39/.

Yang, W., Dai, R, and Cui, X. 2008. Zero Pronoun Resolution in Chinese Using Machine Learning Plus Shallow Parsing. *Proceedings of the IEEE International Conference on Information and Automation.* Institute of Electrical and Electronics Engineers. pp. 905-10.

Yang, Y. 2014. Reading Between the Lines: Recovering Implicit Information from Chinese Texts. Ph.D. Dissertation, Department of Computer Science, Brandeis University.

Yang, Y. and Xue, N. 2010. Chasing the Ghost: Recovering Empty Categories in the Chinese Treebank. *Proceedings of the 23rd International Conference on Computational Linguistics (COLING).* Tsinghua University Press. pp. 1382-90.

Yang, Y., Liu, Y., and Xue, N. 2015. Recovering Dropped Pronouns from Chinese Text Messages. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL).* Association for Computational Linguistics. pp. 309-13.

Yeh, A. 2000. More Accurate Tests for the Statistical Significance of Result Differences. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL).* Association for Computational Linguistics. pp. 947-53

Yeh, C.-L. and Chen, Y.-C. 2007. Zero Anaphora Resolution in Chinese with Shallow Parsing. *Journal of Chinese Language and Computing* 17 (1): pp. 41-56.

Zhao, S. and Ng, H.T. 2007. Identification and Resolution of Chinese Zero Pronouns: A Machine Learning Approach. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).* Association for Computational Linguistics. pp. 541-50.

Appendix—Parse Pattern Details

Here we describe the details of the parsing-based features used by DP-Classifier (the last bullet in Subsection 4.3.2). The current word is assigned a subset of the following parse pattern IDs (based on the Stanford constituency parser). The pattern IDs indicate their source: “YX” and “YLX” indicate that the pattern was drawn from (Yang and Xue 2010) and (Yang, Liu and Xue 2015), respectively; “OURS” indicates that the pattern is our own.

YX1: the current word is the first word in the lowest IP dominating this word.

YX2: the current word starts an IP with no subject. Subject is detected heuristically by looking at left sisters of a VP node.

YX3-[POS]: the current word has POS label [POS] and starts an IP with no subject.

YX4: the current word is the first terminal child of a VP following a punctuation mark.

YX5: the POS of current word is NT, and it heads an NP that does not have a subject NP as its right sister.

YX6: the current word is a verb in an NP/VP.

YX7-[PL]: the phrasal label of the parent of the current word is [PL].

YX8: the previous word is a transitive verb, and this verb does not take an object.

YLX1: the current word is “有” and has no subject.

OURS1: the current word is leftmost (first word) in lowest IP dominating this word.

OURS2: the current word starts an IP with no subject; that is, the VP node has no NP left sisters.

OURS3: the current word is a copula and starts an IP with no subject.

OURS4: the current word is the verb “have” as a main verb and starts an IP with no subject.

OURS5: the current word is a verb (categorized as “other”) and starts an IP with no subject.

OURS6: the current word is a predicative adjective and starts an IP with no subject.

OURS7: the current word is *bei* in short passive construction and starts an IP with no subject.

OURS8: the current word starts an adverb phrase that starts an IP with no subject to the right.

OURS9: the current word is a preposition and starts an IP with no subject to the right.

OURS10: the current word is the first terminal child of a VP following a punctuation mark.

OURS11: the current word is the first terminal child of a VP following a punctuation mark, is immediately dominated by a VP, and has no left sister NP.

OURS12: the POS of current word is NT, and it heads an NP that does not have a subject NP as its right sister.

OURS13: the current word is leftmost word in a VP with a left NP sister.

OURS14: the current word is leftmost word in a VP with a right NP sister.

OURS15: the current word is the verb “有” and has no subject.