Dropped Pronoun Recovery in Chinese SMS

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

Stacy Petersen[1]
Department of Linguistics
Georgetown University
3700 O Street NW
Washington, DC 20057, USA
sjp62@georgetown.edu

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 did not help our approaches. We conclude that the parse information is largely superfluous for identifying dropped personal pronouns if reasonably accurate independent clause start information is available.

1. Introduction

Chinese is commonly characterized as a "pro-drop" language (Baran, Yang, & 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

[1] This author's work was carried while she was a summer intern at the MITRE Corporation.

Short Message Service (SMS) messages sent via cell phones (Yang, Liu, & Nianwen, 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 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 set 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. In the second stage, a label 0, 1, or 2 was assigned to each word: 0 if no dropped personal pronoun were predicted

[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.

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 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 & Gildea, 2010).

2. Data

On February 3, 2012, we downloaded 29,393 Chinese SMS messages from the National University of Singapore (Chen & 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, & Nianwen, 2012) and (Yang, Liu, & Nianwen, 2015).[4]

	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.

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

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 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[5] version 1.6.7 using the "ctb" model (Manning, et al.,

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

2014). This segmenter was not specifically designed for SMS messages. The second was a segmenter specifically designed for Chinese SMS messages (Wang, Wong, Chao, & Xing, 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

We built a linear-chain CRF (Lafferty, McCallum, & 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 bigram consisting of the current word and the word to the left
- The word bigram consisting of the current word and the word to the right
- All characters that appear as a prefix or suffix of the current word
- All character bigrams that appear as 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.[6] The problem of Chinese independent clause prediction is similar to that of Chinese comma classification as addressed in (Xue & Yang, 2011).

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—an approach based on a linear-chain CRF and an approach based on a maximum-entropy classifier.

4.3.1 CRF Approach

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

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

- A Boolean feature set to "true" if an independent clause break was three or fewer words to the left of the current word.
- 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[7] 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.

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

4.3.2 Maximum-Entropy Classifier Approach

Key to this approach is the choice of which words to apply the classifier to. 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.[8]

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.

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

[8] 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.

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.

(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 (regardless of label) has entries that indicate the presence or absence of the following features.

- All word unigrams, bigrams, and trigrams drawn from the words two to the left through two to the right of the current word.
- 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.
- 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 & 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.

The baseline approach first applies the Stanford constituency parser[9] 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 & Gildea, 2010).

(B). The word is left-most in a verb phrase (VP) that (1) has an intonational phrase (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[10] (Levy & Galen, 2006) regular expression.

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

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

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

(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.

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

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

The parser matches the special word 把, used to focus on the result or influence of an action, and marks it as BA. A dropped pronoun may occur before the BA, which is what this rule captures.

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

(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))

5.2 Methodology

The baseline approach DP-CRF and DP-Classifier were all applied to the messages in the test part[11] and F scores computed. Several variations on our approaches were compared: (1) DP-CRF and DP-Classifier using the Stanford word segmenter versus the UMAC segmenter; (2) DP-Classifier with and without parsing-based features (third in the bulleted list of features at the end of Subsection 4.3.2); (3) DP-CRF and DP-Classifier with predicted versus oracle (ground truth) independent clause start labels; and (4) DP-CRF with and without independent clause start features (first in the bulleted list of features at the end of Subsection 4.3.1).

For all approaches, precision, recall, and F 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 of the approaches 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.

[11] In the test part, our annotator added back 117 dropped personal pronouns.

(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).[12]

In order to better understand the sources of error in our pipeline, we also considered DP-CRF and DP-Classifier with an oracle ICS predictor. Namely, DP-CRF and DP-Classifier were modified to use, at testing time, the independent clause starts marked by the annotator. Step (B) in Section 3 was dropped in both the training and application procedures. The application procedure was modified to be given a new Chinese SMS message with ICS markings provided by the annotator. And, step (A) in the application procedure was modified such that any ICS marking on a character in the middle of a word (as produced by the word segmenter) was moved to the left-most character in the word.

6. Results

Several observations can be made from Figure 5, below.

- DP-CRF and DP-Classifier exhibited substantially higher scores with the Stanford word segmenter than with the UMAC segmenter. 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.
- The parse-based features did not help DP-Classifier. We conclude that the parse information is largely superfluous for identifying dropped personal pronouns, once reasonably accurate 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. The machine-learning approaches exhibited much higher precisions and moderately lower recalls.

[12] 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.

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	--	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
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	0.10	0.33	0.16

FIGURE 5: The table displays an accuracy comparison between DP-CRF, DP-Classifier, and DP-Baseline. The second column indicates whether DP-Classifier uses parse-based features (the last bullet in Subsection 4.3.2). [13]

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. As seen in Figure 6, below, the modified versions produced very little improvement with the Stanford segmenter. Thus, devoting more effort to improve the ICS predictor would likely not bear much fruit.

APPROACH	WORD SEGMENTER	ORACLE ICS PREDICTOR?	1st PERSON PREC	1st PERSON REC	1st PERSON F	2nd PERSON PREC	2nd PERSON REC	2nd PERSON F
DP-CRF	Stanford	Yes	0.42	0.33	0.37	0.55	0.35	0.43
		No	0.41	0.30	0.34	0.60	0.34	0.43
	UMAC	Yes	0.33	0.25	0.29	0.47	0.28	0.35
		No	0.39	0.22	0.28	0.52	0.24	0.33
DP-Classifier	Stanford	Yes	0.42	0.33	0.37	0.47	0.37	0.41
		No	0.38	0.31	0.34	0.49	0.34	0.40
	UMAC	Yes	0.37	0.27	0.31	0.46	0.25	0.32
		No	0.40	0.25	0.31	0.50	0.25	0.33

FIGURE 6: The table displays an accuracy comparison between our approaches with and without an oracle ICS predictor (without an oracle ICS predictor, each approach uses an ICS predictor as described in Subsection 4.2.). In all cases with DP-Classifier, parse-based features are not used.

[13] Parse-based features are not applicable to Baseline. Parse-based features could, in principle be applicable to DP-CRF, but, we chose not to include them since they didn't help DP-Classifier much.

Finally, 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 7, 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 7: 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.

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, & Ishikawa, 2002), (Kawahara & Kurohashi , 2005) and (Sasano & 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, Huang, & Kroch, 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 (Yang & 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 (Chung & 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 et al. (Cai, Chiang, & 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 (Kong & 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 (Xue & Yang, 2013), (Yang Y. , 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 (Yeh & 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, & Weinstein, 1995) in the design of the rules component that identifies antecedents. Kong and Zhou (Kong & 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 (Zhao & 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 et al. (Yang, Dai, & 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 (Chen & Ng, 2013) extend Zhao and Ng's approach by utilizing more complex features and co-reference links between dropped pronouns.

Chen and Ng (Chen & 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 et al. (Yang, Liu, & Nianwen, 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 et al. (Rao, Ettinger, Daume III, & 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, & 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 (Rahman & 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, Ettinger, Daume III, & Resnik, 2015) and (Yang, Liu, & Nianwen, 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.

[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.

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.
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 (Chung & Gildea, 2010). Features derived from parsing did not help our approaches. We conclude that the parse information is largely superfluous for identifying dropped pronouns if reasonably accurate independent clause start information is available.

One avenue of future work is to use our dropped pronoun slot detection approach followed by the approach in (Rao, Ettinger, Daume III, & Resnik, 2015) to assign pronoun identity information to the slots. To do this, we would need a different data set that contains SMS dialogue threads.

9. 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 & Xue, 2010) and (Yang, Liu, & Nianwen, 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.

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., & Nianwen, X. (2012). Annotating Dropped Pronouns in Chinese Newswire Text. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC)* (pp. 2795-2799). European Language Resources Association (ELRA).

Cai, S., Chiang, D., & Goldberg, Y. (2011). Language-Independent Parsing with Empty Elements. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 212-216). Association for Computational Linguistics.

Chen, C., & Ng, V. (2013). Chinese Zero Pronoun Resolution: Some Recent Advances. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1360-1365). Association for Computational Linguistics.

Chen, C., & 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* (pp. 1622-1628). Association for the Advancement of Artificial Intelligence Press.

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

Chung, T., & Gildea, D. (2010). Effects of Empty Categories on Machine Translation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 636-645). Association for Computational Linguistics.

Grosz, B., Joshi, A., & Weinstein, S. (1995). Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics, 21*(2), 203-225.

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

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

Kong, F., & 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)* (pp. 882-891). Association for Computational Linguistics.

Kong, F., & Zhou, G. (2013). A Clause-Level Hybrid Approach to Chinese Empty Element Recovery. *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 2113-2119). Association for the Advancement of Artificial Intelligence Press.

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proceedings of the 18th International Conference on Machine Learning (ICML)* (pp. 282–289). Morgan Kaufmann.

Levy, R., & 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)* (pp. 2231-2234). European Language Resources Association (ELRA).

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

McCallum, A. (2002). Retrieved July 16, 2013, from MALLET: MAchine Learning for LanguagE Toolkit : http://mallet.cs.umass.edu

Rahman, A., & 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)* (pp. 1051-1060). Association for Computational Linguistics.

Rao, S., Ettinger, A., Daume III, H., & 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)* (pp. 494-502). Association for Computational Linguistics.

Sasano, R., & 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)* (pp. 758–766). Association for Computational Linguistics .

Seki, K., Fujii, A., & 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)* (pp. 1-7). Association for Computational Linguistics.

Wang, L., Wong, D., Chao, L., & 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* (pp. 51-57). Association for Computational Linguistics.

Xue, N., & Yang, Y. (2011). Chinese Sentence Segmentation as Comma Classification. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 631-635). Association for Computational Linguistics.

Xue, N., & 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)* (pp. 1051-1060). Association for Computational Linguistics.

Xue, N., Xia, F., Huang, S., & 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 . Retrieved from http://repository.upenn.edu/ircs_reports/39/

Yang, W., Dai, R., & Cui, X. (2008). Zero Pronoun Resolution in Chinese Using Machine Learning Plus Shallow Parsing. *Proceedings of the IEEE International Conference on Information and Automation* (pp. 905-910). Institute of Electrical and Electronics Engineers.

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

Yang, Y., & Xue, N. (2010). Chasing the Ghost: Recovering Empty Categories in the Chinese Treebank. *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)* (pp. 1382-1390). Tsinghua University Press.

Yang, Y., Liu, Y., & Nianwen, X. (2015). Recovering Dropped Pronouns from Chinese Text Messages. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 309--313). Association for Computational Linguistics.

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

Zhao, S., & 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)* (pp. 541-550). Association for Computational Linguistics.