Dialogue focus tracking for zero pronoun resolution

Sudha Rao^{1,3}, Allyson Ettinger², Hal Daumé III^{1,3}, Philip Resnik^{2,3}

¹Computer Science, ²Linguistics, ³UMIACS

University of Maryland

raosudha@cs.umd.edu, aetting@umd.edu, hal@cs.umd.edu, resnik@umd.edu

Abstract

We take a novel approach to zero pronoun resolution in Chinese: our model explicitly tracks the flow of focus in a discourse. Our approach, which generalizes to *deictic* references, is not reliant on the presence of overt noun phrase antecedents to resolve to, and allows us to address the large percentage of "non-anaphoric" pronouns filtered out in other approaches. We furthermore train our model using readily available parallel Chinese/English corpora, allowing for training without hand-annotated data. Our results demonstrate improvements on two test sets, as well as the usefulness of linguistically motivated features.

1 Introduction

"Pro-drop" languages like Chinese, Japanese and Turkish allow for dropping of pronouns when the referents of those pronouns can be inferred. English is typically *not* pro-drop, but is unusual in that regard: two thirds of languages documented in WALS (Haspelmath et al., 2005) can be categorized as prodrop. In such languages, sentences are frequently characterized by "zero pronouns": gaps in the sentence which in English would hold an overt pronoun. In some languages, verbal morphology or clitics elsewhere in the sentence are sufficient to resolve the ambiguity of dropped pronouns; in other languages, there is *no* overt marking at all in the sentence and the referent of the dropped pronoun must be resolved using pragmatic information.

Our work departs from mainstream work on zero pronoun resolution in that we focus primarily on the resolution of *deictic* zero pronouns. Unlike an



Figure 1: A conversation between a student and a teacher. The text has been translated from Mandarin, but zero pronouns are retained and indexed with their referent: (T)eacher or (S)tudent.

anaphoric zero pronoun (Section 2), whose reference must be specified by a noun phrase occurring previously in the text, a *non-anaphoric* zero pronoun refers to an entity that is salient from larger units of discourse (such as full sentences or passages) or from the extralinguistic environment (outside of the text altogether). Although anaphoric zero pronoun resolution has been the focus of most past work (Yeh and Chen, 2007; Chen and Ng, 2014), 50% or fewer of zero pronouns in natural Chinese text are anaphoric (Zhao and Ng, 2007; Kong and Zhou, 2010). Our approach allows for generalization to non-anaphoric pronouns, focusing in particular on *deictic* non-anaphoric zero pronouns, which refer to salient entities in the environment (such as the speaker, hearer or pragmatically accessible referent) without requiring any introduction in the preceding text. Figure 1 shows an example conversation in which zero pronouns are frequently used to refer to speaker or listener, and would be translated to English as "I" or "you."

We propose a model for resolving deictic zero pronouns that draws inspiration from ideas in Centering Theory (Grosz et al., 1995): discourses tend to settle on a particular focus for a time, before switching. Furthermore, we presume that when a switch happens, there is likely to be an overt cue of this. For example, in Figure 1, the initial focus on **T** is signaled with the overt second person pronoun in the first utterance; the switch of focus to S in the third utterance is also signaled by an overt "you." However, at that point, the focus remains on S for several utterances until "The last round..." at which point it switches away from the speakers. It is brought back to **S** in the last utterance, which can be inferred from the fact that **S** is the most recent focus that fits the required semantic constraints.

To account for these phenomena, we develop a novel sequential model for zero pronoun resolution that explicitly tracks the conversation focus in a dialogue (Section 3). We test, using data from Chinese SMS ("texting") dialogues, the hypothesis that our model can predict the identity of pronouns (at a granularity of the person attribute: first, second, or third person—with particular focus on first and second person) based on a variety of features of the utterance context, without reference to a particular antecedent (Section 4.3). In this way, we address a much higher percentage of the zero pronouns found in Chinese texts, and particularly in dialogue.

Our second contribution is to show that **one can train a zero pronoun resolution system using supervision coming from English translations of the Chinese text** (Section 2.2). This obviates the need for expensive linguistic annotation of Chinese and allows us to use plentiful parallel data to train our model. Our results confirm that even though this "translation as annotation" process is noisy, it is still possible to learn on large amounts of "bronze standard" data.

2 Linguistic motivation

Handling zero pronouns in Chinese (or other prodrop language) involves two separate tasks: (1) Zero pronoun identification: locating and marking the gaps corresponding to zero pronouns; and (2) Zero pronoun resolution: determining the entity referred to by the zero pronoun. Our focus is the latter task.

Zero pronoun resolution, like general pronoun resolution, is almost universally approached as a problem of linking a pronoun to an overt noun phrase antecedent in the text. However, while some zero pronouns do have overt noun phrase antecedents, many other zero pronouns do not. In fact, (Zhao and Ng, 2007) report that just 52% of zero pronouns in their training set (and 46% of zero pronouns in their test set) are "anaphoric." Kong and Zhou (Kong and Zhou, 2010) report just 41%. Some zero pronouns fail to link to an antecedent because they refer to facts or events described by larger phrases or full sentences earlier in the text, preventing coreference with a single noun phrase. Other zero pronouns, particularly in dialogue settings, are *deictic*, pointing to salient entities in the environment without requiring introduction by an overt mention in the text.

2.1 Dialogue focus

A central principle of document cohesion that underlies frameworks such as Centering Theory (Grosz et al., 1995) states that discourses tend to settle on a particular focus for a time, before eventually switching to a new one. The status of a particular focus within this flow of discourse is typically signaled by the form of the expression chosen to point to it. When a focus is introduced (or returned to), a full (overt) noun phrase is generally used to indicate it. While that entity remains in focus, subsequent mentions can be realized with less explicit forms. In English, these less explicit forms are overt pronouns. In Chinese (and pro-drop languages more generally), these focus continuations are generally realized as zero pronouns.

We see in this example an illustration of these discourse principles:

1. In pro-drop languages (Chinese), overt pronouns introduce switches in focus, while zero pronouns are used while an established focus continues.

- 2. In non-pro-drop languages (English), overt pronouns serve the focus-continuation function.
- 3. There are "deictic" exceptions to these rules, licensed by environmental salience of the referent and inferable from the meaning of the utterance (final question of the example).

Importantly, these continuations and switches of foci occur for the most part at the level of the syntactic clause. This is thus the level at which we model, assigning labels to individual clauses, which will in turn indicate the identity of any dropped subject pronoun in that clause.¹ In identifying focus, we remain at the granularity of the "person" attribute (first, second, or third person). This is the most relevant granularity for deictic pronoun resolution, as the intent is to capture the alternation between speakers within that dialogue (first and second person), along with switches of focus to any referents external to the dialogue (third person).

2.2 Translation as annotation

Currently, most state-of-the-art machine learning systems for Chinese zero pronoun resolution are supervised, requiring manually resolved pronouns for training. We hypothesize comparable distribution between zero pronouns in a pro-drop language, and overt pronouns in a *non*-pro-drop language. More specifically, because non-pro-drop languages lack zero pronouns, the discourse functions that are served by zero pronouns in pro-drop languages *must* in non-pro-drop languages be served by overt pronouns.

To be more concrete, the original Mandarin SMS conversation from Figure 1 is reproduced in Table 1, together with a human translation into English. Indeed, we see in the example in Figure 1 that the zero pronouns on the Chinese side correspond to overt pronouns on the English side. For this reason we make use of a parallel (Chinese/English) corpus for training of our sequence labeling model, deriving the identities of missing pronouns from the English translation of the Chinese text rather than from coreference relations with antecedents in the Chinese text. Our model thus does not rely on the availability of hand-annotated data for training.

3 Our focus tracking model

Given a Chinese dialogue, our goal is to identify zero pronouns and resolve them either as deictic (first or second person) or non-deictic (third person). We use off-the-shelf tools for the identification of the zero pronouns (described in Section 4.1) and focus on the *resolution* task.

In our implementation, we *jointly* predict the *focus* and identify the number of the pronoun that would be used. For instance, when **S** is speaking about herself, we consider this a "1" label; when **S** is speaking about her conversation partner, we consider this a "2" label. This numbering corresponds to which pronominal form would be required in English.²

3.1 Supervision via bronze standard data

We obtain "bronze" standard (as opposed to "gold" standard) data by looking at human-produced English translations of Chinese utterances, such as those seen in Table 1. Our label set consists of two properties: the *person* being referred to (first person, second person or third person), and whether the reference is overt or not (visible or hidden). The "visible" three labels correspond to clauses in which an overt subject pronoun appears *on the English side*. Chinese clauses bearing this label may have an overt or a zero pronoun subject—if the Chinese side contains a zero pronoun subject, then this label will be used to determine the correct person attribute (first, second, or third) of the unseen pronoun.

1v: Overt English first person pronoun: "I" or "we"

- 2v: Overt English second person pronoun: "you"
- **3v:** Overt English third person pronoun: "he", "she", "it", "they"

¹Although Mandarin does license dropped object pronouns, we focus in this paper only on subject pronouns, as the syntactic subject is (a) more consistently dropped in Mandarin, and (b) more tightly tied to the notion of focus of conversation that motivates our model; see also a discussion of the centering hierarchy in Chinese (Wang, 2011). Relatedly, we filter out possessive pronouns in subject position, as they do not point to the topical entity represented by the full noun phrase.

²We ignore morphological issues in English dealing with possession and grammatical role, since these are exogenous to the resolution task.

	Original Mandarin	English Translation	Label
1) Student	陈老师你还好吗	How are you, Teacher Chen?	2v
2) Teacher	Ø好啊,你还没走?	I am fine. You have not left yet?	1v, 2v
3) Student	Ø回来有一个月 Ø不敢给你说话	I have been back for a month.	1v, 1v
		I didn't dare to chat with you.	
4) Student	没呢	Not yet.	1h
5) Student	美国的面试 🖉 进行了 4 轮了	I have gone through 4 rounds of interviews for the	1v
		American (company).	
6) Teacher	为什么	Why?	2h
7) Student	最后一轮是跟总经理	The last round of interview is with the general manager.	3v
8) Student	还有两次网上测验	There are also two online tests.	3h
9) Teacher	◎还在面试阶段吗	Are you still in the interview phase?	2v

Table 1: Sample Chinese SMS conversation with English translation and derived labels.

However, there are plenty of utterances (e.g., Table 1 lines 4 and 6) in which the English translation does not contain an overt subject. This can happen in English in imperative constructions, (some) questions, and general informal communication.³ In these cases, we introduce "hidden" person labels whose role is to carry forward the focus from the previous utterance. For instance, in utterance 4, even though there is no subject on the English side, we carry forward the fact that the most-recent referent was "first person" and denote this with "1h."

Because we are jointly modeling the focus shift and the pronoun realization aspects, when the speaker shifts, the "hidden" person must flip. For example, in utterance (5) the **Student** overtly refers to herself, yielding a label of "1v." The next utterance is by the **Teacher** but lacks an English subject. The focus remains on the **Student** and therefore this utterance is labeled "2h" meaning that the focus is on the *other* speaker, and it is non-overt in English.

1h: subject being continued is first person2h: subject being continued is second person3h: subject being continued is anything else

Finally, we introduced a seventh label for instances in which no overt subject pronoun appears on the English side, and no focus has yet been established from prior clauses (this applies only at the beginning of a discourse).

None: no subject and no focus yet established

In Table 1, the rightmost column shows the label assignment for the sample SMS exchange. (The utterances on lines 2 and 3 contain two clauses each, and thus two labels each.)

3.2 Features

We included in our model the following features. Note that these features are based solely on the *Chinese side*. Linguistic motivations for each feature category are described.

Subject continuation: a value indicating the *person* (1, 2, 3) of the most recent overt NP that was a direct descendent of an IP node (the most recent overt NP in structural subject position—including, if overt, the subject of the current clause). *The most recent overt NP subject is a strong candidate for coreference with a zero pronoun. This feature comes closest to attempting antecedent selection.*

Verb: the first verb in the VP that is sister to the subject NP (the VP of which that NP is the subject). *The nature of the verb can provide information relevant to inferring the identity of deictic forms.* For example: the Chinese verb *guji* ("reckon") is intuitively biased toward first-person subject; our training data accordingly show 68% of clauses with *guji* as verb feature were assigned first-person subject labels.

Participant index: a value indicating the index of the conversational participant. *To capture regular-ities, if any exist, in the pronoun use of a speaker.*

Participant switch: a binary value indicating whether the current utterance represents a change of speaker relative to the previous clause. *Switches in*

³This can also happen due to imperfect zero pronoun identification (Section 4.1).

speaker may, particularly in tandem with other features, be informative about topic.

Object (downstream): the direct object of the VP sister to the subject (if any). *This feature exploits the fact that pronouns occurring as direct objects within a clause cannot be the same as the (zero) pronoun in subject position of that clause.*

Has question particle: a binary value indicating whether the clause contains a) a question particle or *wh*-word, or b) a question mark. *This feature is likely to be a strong indicator of that the subject pronoun is* not *first person (also used by (Chen and Ng, 2013)).* For example, in our training data we found that only 16% of the clauses with question particle were marked with first person label i.e. 1v or 1h.

Bag of words: all words occurring in the clause. Apart from the verb, other words can also be highly informative about the nature of the subject.

Bag of parts of speech: all parts of speech occurring in the clause. *The structural make-up of clause may be informative about focus, for instance in the case of passive or possessive constructions.*

Hidden subject particles: a feature indicating whether the clause consists of a list of phrases consistently tagged with empty categories on the Chinese side, but consistently translated without subject pronouns on the English side (thus likely to correspond to labels 1h-3h). *This feature is intended to help the model in recognizing clauses consistently corresponding to "hidden" labels.*

In addition, for the features that consist of sequence (bag of words, bag of part of speech, object, etc.) we additionally compute bigrams and trigrams.

3.3 Structured prediction

We cast the above model as a sequence labeling problem over visible and hidden labels. We consider each conversation segment in the SMS as an input data sequence $\boldsymbol{x} = \langle x_1, x_2, \dots, x_n \rangle$ where each x_i corresponds to a clause in Chinese. Each clause in Chinese is assigned a label from the label space Y= {1v, 2v, 3v, 1h, 2h, 3h, none}. The task then is to assign labels $\boldsymbol{y} = \langle y_1, y_2, \dots, y_n \rangle$ to the input data sequence from the label space Y based on the features described in Section 3.2. At training time we assign labels to the input sequence using the "bronze standard" method described in Section 3.1.

To train the sequence labeling model, we use an online variant of the DAgger imitation learning algorithm (Ross et al., 2011) as implemented in the Vowpal Wabbit machine learning library (Langford et al., 2007; Daumé III et al., 2014). DAgger, like its predecessor SEARN (Daumé III et al., 2009), solves structured prediction problems by transforming them into sequential decision making problems. In the case of sequence labeling, the natural order for sequence decision making is left-to-right. At test time, inference is performed greedily. At training time, the learning algorithm attempts to balance between training on "oracle" states (prefixes of decisions made optimally according to the true labels) and training on "system" states (prefixes of decisions made sub-optimally according to the learned model). The online variant of DAgger balances this trade-off by slowly transitioning from making past decisions optimally to making them using the currently learned predictor.

4 Experiments

Our goal in our experiments is to answer the following questions:

- 1. How well does the bronze-standard annotation capture the underlying truth? (Section 4.2).
- 2. Is our model able to leverage both dialogue structure and semantic content to accurately resolve pronouns? (Section 4.3)
- 3. How important are the different components in our model in making effective predictions? (Section 4.4)

In the following sections, we describe the experiments we perform aimed at answering these questions. First, we describe the data we use for experimentation.

4.1 Experimental setup

For training our focus-tracking model, we use Chinese-English parallel data from the SMS/chat domain available as part of training data used in the Machine Translation task under the DARPA BOLT project. The training data consisted of 117k sentences. We test our model on heldout SMS/Chat data consisting of 1152 sentences (hand-annotated,

	Bronze	SMS	OntoNotes	
	Training	Test	Test	
# tokens	1,007,722	8104	108,531	
# sents	129, 190	1152	9607	
# dialog	3309	34	257	
# types	26,519	1747	4753	

Table 2: Dataset statistics; numbers are for the Chinese side of the data. English has 25% more tokens and roughly as many types.

as described in Section 4.2), and on telephone conversation data from the OntoNotes corpus (Hovy et al., 2006), consisting of 5000 sentences. Full data statistics are provided in Table 2.

We perform zero pronoun identification using the method of (Cai et al., 2011), which automatically recovers empty categories corresponding to dropped pronouns, integrating these empty categories into syntactic parses. Syntactic parses were obtained with the Berkeley parser (Petrov and Klein, 2007). These parses were then used to split the Chinese utterances into single-clause units, based on IP and CP clausal nodes. These clauses were aligned with clauses in the English translation, which were used to determine the identity of the clausal subject, for extracting the 1v, 2v, ... label for each utterance.⁴

For our machine learning systems, we use Vowpal Wabbit (Langford et al., 2007) with default hyperparameter settings. We train on 75% of the training data and retain 25% as development data on which to perform early stopping. We run 20 iterations by default and take the parameters with best development performance based on sequence labeling accuracy.

4.2 Gold standard test set

Although we can use "bronze standard" annotations for learning, evaluating against a bronze standard is not directly useful. Therefore, we annotated our test set (1152 utterances) by hand. In particular, for the SMS/chat test set, we recruited three linguistically-informed native Mandarin speakers to annotate Chinese clauses containing empty categories. The clauses were labeled with a person number (1,2,3) when the empty category corresponded to

Pronoun	Precision	Recall	F-measure	
1p	0.75	0.43	0.55	
2p	0.61	0.32	0.42	
3p	0.52	0.45	0.49	
Micro-avg	0.62	0.41	0.50	

Table 3: Bronze vs Gold labels

such a pronoun; or "none" in spurious cases.⁵

In our annotated data,⁶ 32% of identified zero pronouns were first person, 17% were second person, 25% were third person and 26% do not have a referent (were spurious). Of the correctly identified zero pronouns, a majority of pronouns (about 2/3) are deictic: referring either to the speaker or listener. The remainder are third person and mostly anaphoric.⁷ Since the annotators labeled the empty categories obtained from an automatic zero pronoun identification method (Cai et al., 2011), 26% spurious cases suggest that the accuracy of this method is only 74% on the SMS test data set.

We then used these annotations to evaluate our bronze standard label assignment method against the gold standard judgments. Table 3 shows the precision, recall and F-measure of the bronze annotations when evaluated against the gold annotations. We use micro-averaging to average the precision, recall and F-measure values against different sets. In this method we sum up the individual true positives (TP), false positives (FP), and false negatives (FN) of the system for different sets and then apply them to get the statistics. For example, precision across two sets 1 and 2 is given by (TP1 + TP2)/(TP1 + TP2 + FP1 + FP2). We can see a fairly significant discrepancy between our bronze labels and the gold labels. One major-and unfortunately inevitablereason for this discrepancy is a high proportion of utterances in the English translation data which have

⁴Sometimes English syntactic parses were not well-aligned with the Chinese IP/CP nodes; in practice, we split the English utterances based on end-of-clause punctuation and aligned Chinese and English clauses based on a simple order heuristic.

⁵Note that under this annotation scheme, our evaluations will be partially constrained by (Cai et al., 2011) performance, in including no zero pronouns that were missed by that method—however, use of the "none" label allows filtering out of any spuriously-identified zero pronouns.

⁶Annotations are available at www.umiacs.umd.edu/~raosudha/LDC2013E83-BOLT-P2R2-cmn-SMS-CHT-dev.annotated.

⁷In an in-person dialogue, a third person pronoun might be used in a deictic manner, as in "*She* is really smart" while pointing at someone. This rarely occurs in SMS/chat because there is no shared environment beyond the two dialogue participants.

been translated with the subject pronouns still absent. This is partially due to the casual nature of the text, and partially because the quality (fluency) of English translations in this data is at times dubious.

While there may be a systematicity to this kind of subject omission on the English side, this was not a factor taken into account by our human annotators. So while our own model may stand a chance at predicting "hidden" labels (no overt pronoun on English side) in such instances, the annotators will never assign a label of "none" to a location at which a pronoun could reasonably have been inserted.

4.3 Overall system efficacy

In this section we discuss the overall efficacy of our proposed method in comparison to a few alternatives. These alternatives are:

Random guessing baseline. A naïve system that makes predictions uniformly at random.

Subject continuation baseline. This is a rulebased approach that mimics the intuitions described in Section 2. In particular, for a Chinese utterance, we check whether the current utterance has an overt pronominal subject. If so, we assign a label of 1v, 2vor 3v depending on the person of this subject. If the current utterance has a non-pronominal subject, we assign 3h. Otherwise we "carry forward" the subject from the previous utterance, flipping the 1p/2pas necessary when the speaker changes; these are labeled as 1h, 2h or 3h.

Minimal model baseline. In the minimal model, we restrict our model to use just three features: participant index, participant switch and subject continuation feature. This is a machine learning variant of the rule-based subject continuation baseline.

Oracle upper bound. None of the proposed models can hope to achieve 100% accuracy on this task because the gold annotation data consists of 26% "no pronoun" cases. Since all of our approaches *must* predict a pronoun when a zero pronoun has been identified, their performance (namely, their precision) is upper-bounded away from 100%.

The summary of results (micro-averaged across 1p, 2p and 3p) are shown in Table 5. These results show that on both the SMS data (on which the model was developed) and the OntoNotes data (on which

	SMS/chat			OntoNotes			
	Mic	ro-aver	age	Micro-average			
System	Pre Rec F			Pre	Rec	F	
Random	0.24	0.25	0.24	0.18	0.26	0.22	
Minimal	0.42	0.23	0.30	0.30	0.07	0.11	
SubjCont	0.32	0.42	0.36	0.31	0.15	0.20	
Full Model	0.59	0.31	0.41	0.30	0.36	0.33	
Upper Bound	0.74	1.00	0.85	0.55	1.00	0.71	

Table 5: Summary of results for different comparator models against the gold standard labels from SMS data (left) and OntoNotes (right).

the model was applied blindly), our full model is able to substantially outperform the baselines. In fact, on OntoNotes, despite a potential domain mismatch (from SMS/chat to telephone conversations), our full model was the only baseline to beat random guessing! Across both data sets, the minimal model tends to have high precision and low recall; the behavior of the other approaches varies across the tasks. On the SMS/chat data, our model achieves a 14% relative improvement over the best baseline; on an OntoNotes data, a 50% relative improvement.

More specific breakdowns of performance by different pronouns (1p, 2p and 3p) are shown for the subject continuation baseline and the full model in Table 4. In these tables, we also report results when evaluated on the OntoNotes test set in these Tables. As we can see, the subject continuation baseline massively overpredicts third person pronouns in the SMS data, leading to an overall low score. In comparison, our model tends to have much higher precision (at the expense of recall) across the board on the SMS data, leading to a 14% relative improvement over the subject continuation baseline.

Since, to our knowledge, no prior work (see Section 5) has focused on deictic pronoun restoration, it is not possible to directly compare our results to previously published results. Although it is an apples-to-oranges comparison, a state-of-theart *anaphoric* zero pronoun resolution system (Chen and Ng, 2014) achieves a precision of 13.3, a recall of 32.2 and an F-measure of 18.8 on the *telephone conversation* part of the OntoNotes data, but does so addressing the complementary problem of correctly choosing antecedents from previous overt noun phrases.

Another reasonable comparison would be with a

Subject Continuation Baseline								
	5	SMS Test	set	OntoNotes Test set				
Pronoun	Precision	Recall	F-measure	Precision	Recall	F-measure		
1p	0.43	0.28	0.34	0.29	0.08	0.13		
2p	0.27	0.19	0.22	0.28	0.11	0.16		
3 p	0.29	0.75	0.42	0.32	0.21	0.25		
Micro-avg	0.32	0.42	0.36	0.31	0.15	0.20		
-								
		(Our Full Mod	el				
	5	OntoNotes Test set						
Pronoun	Precision	Recall	F-measure	Precision	Recall	F-measure		
1p	0.64	0.47	0.54	0.27	0.58	0.37		
2p	0.55	0.21	0.31	0.25	0.23	0.24		
3n	0.50	0.18	0.27	0 40	0.28	0 33		

Table 4: Results across different pronoun categories for (top) subject continuation and (bottom) our full model.

0.41

0.30

0.36

	SMS/chat Micro-average			OntoNotes Micro-average			
System	Pre	Rec	F	Pre	Rec	F	
Minimal (M)	0.42	0.23	0.30	0.30	0.07	0.11	
M + question	0.44	0.23	0.30	0.31	0.07	0.12	
M + object	0.43	0.23	0.30	0.30	0.07	0.12	
M + verb	0.58	0.29	0.39	0.32	0.13	0.18	
M + pos	0.52	0.30	0.38	0.23	0.27	0.25	
M + bow	0.59	0.28	0.38	0.30	0.35	0.32	
Full Model	0.59	0.31	0.41	0.30	0.36	0.33	

Micro-avg

0.59

0.31

Table 6: Summary of results for feature ablation against the gold standard labels from SMS data (left) and OntoNotes (right).

model trained on gold annotated data (instead of bronze data). Owing to cost, we could not obtain gold annotations for the full training set; however, a leave-one-out cross validation on the gold annotated test set (of SMS data) gave an F-measure of 0.47, versus 0.41 for our model trained on bronze labels. This suggests the noisy bronze labels are indeed useful for this task.

4.4 Feature ablations

In order to investigate the individual contributions of each of our features, we performed feature ablation experiments, pairing our Minimal Model with a single feature at a time and retraining the model with this pairing. The results of these experiments can be seen in Table 6. We see in this table that for the SMS data, the Verb feature creates the greatest improvement over the Minimal Model, followed by Bag of Words and Bag of POS. This supports the hypothesis that the verb is informative with respect to the nature of its subject, as are the other words of the clause, and their parts of speech. For the OntoNotes corpus, however, the Bag of Words feature performs best by a large margin. Interesting, although the Bag of Words features are clearly the most useful, the linguistically motivated features (verb/question) performing well supports our linguistic intuitions.

0.33

5 Discussion and related work

Past approaches to zero pronoun resolution focus exclusively on anaphoric zero pronouns approached as a task of antecedent identification. Almost all work makes use of syntactic structure, with differences primarily in how that structure is used. (Yeh and Chen, 2007) take a rule-based, Centering Theoryinspired approach based on a system of constraints to guide selection of zero pronoun antecedents. In the same year, (Zhao and Ng, 2007) introduced a supervised learning approach for both zero pronoun identification and antecedent selection based on engineered features; these engineered features were replaced with a tree-kernel by (Kong and Zhou, 2010), who jointly perform zero pronoun identification, anaphoricity determination, and antecedent selection. Recently, (Chen and Ng, 2013) built upon the model introduced by Zhao and Ng, introducing additional features and allowing coreference links between multiple zero pronouns. Chen and Ng (2013) also test their model on automatically identified zero pronouns and automatically generated parse trees, thus presenting the first end-to-end Chinese zero pronoun resolver.

These approaches are mostly complementary to our task, since they focus on resolving anaphoric zero pronouns (the minority!) while we focus on resolving deictic non-anaphoric zero pronouns. In particular, for an end-to-end system that resolves both deictic and anaphoric zero pronouns, one could first take our approach and whenever our model predicts "third person," which is often an anaphoric reference, one could apply one of these prior approaches for further reference resolution.

The only work we are aware of that does not require linguistically annotated data for zero pronoun resolution is that of (Chen and Ng, 2014). They hypothesize that zero pronouns and overt pronouns have *similar* distributions, and train an unsupervised model on *overt* pronouns and then apply this model to *zero* pronouns. This model performs on par with their previous (2013) supervised model. Despite this, their unsupervised model only agrees with their supervised model on 55% of zero pronoun antecedents, suggesting that this hypothesis is weak.

In particular, the complementarity of zero versus overt pronoun usage has been studied within various domains of linguistics. The Position of Antecedent Hypothesis (Carminati, 2002) states that null and overt pronouns have different antecedent selection preferences: null pronouns prefer antecedents in subject positions, while overt pronouns prefer antecedents in non-subject positions. This hypothesis has been supported by studies in a variety of prodrop languages (e.g., (Alonso-Ovalle et al., 2002) (Kweon, 2011)). Switching of reference has been identified as one of the main constraints regulating use of zero versus overt pronouns in the variationist literature (see (Cameron, 1992) for sociolinguistic studies of the phenomenon in Spanish). The importance of topically-coherent discourse sequencesand the role of linguistic and extralinguistic indicators of such sequences-has also been examined in child language acquisition, e.g., (Rohde and Frank, 2014).

Our main result shows that although our bronze standard labels are noisy (Section 4.3), they are nonetheless useful for learning to resolve deictic pronouns. Moreover, one oft-heralded advantage of the translation-as-annotation scheme (Carpuat and Wu, 2007) is that it naturally integrates into a machine translation framework, since one is learning to predict precisely what is necessary for successful translation; evaluating whether this hypothesis is true is currently an open question. One limitation of our approach is the coarseness of the labeling granularity (1p, 2p, 3p). Our ultimate plan is to provide all possibilities (e.g., both singular and plural for 1p, weighted) to a machine translation system, and let other components (e.g., language model) determine selection. For now, we believe that there is significant value in intrinsic evaluation of our approach for a problem that has not previously received significant attention.

Acknowledgments

This research was supported in part by the BOLT program of the Defense Advanced Research Projects Agency, Contract No. HR0012-12-C-0015. Any opinions, findings, conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the view of DARPA. The authors would like to thank three anonymous reviewers for providing helpful comments and also acknowledge people from Computational Linguistics and Information Processing (CLIP) lab at University of Maryland for helpful discussions.

References

- [Alonso-Ovalle et al.2002] Luis Alonso-Ovalle, Susana Fernández-Solera, Lyn Frazier, and Charles Clifton. 2002. Null vs. overt pronouns and the topic-focus articulation in Spanish. *Journal of Italian Linguistics*, 14:151–170.
- [Cai et al.2011] Shu Cai, David Chiang, and Yoav Goldberg. 2011. Language-independent parsing with empty elements. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2, pages 212–216. Association for Computational Linguistics.
- [Cameron1992] Richard Cameron. 1992. Pronominal and null subject variation in Spanish: Constraints, dialects, and functional compensation. Ph.D. thesis, University of Pennsylvania.
- [Carminati2002] Maria Nella Carminati. 2002. *The processing of Italian subject pronouns*. Ph.D. thesis, University of Massachusetts at Amherst.
- [Carpuat and Wu2007] Marine Carpuat and Dekai Wu.

2007. Improving statistical machine translation using word sense disambiguation. In *In EMNLP*.

- [Chen and Ng2013] Chen Chen and Vincent Ng. 2013. Chinese zero pronoun resolution: Some recent advances. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1360–1365.
- [Chen and Ng2014] Chen Chen and Vincent Ng. 2014. Chinese zero pronoun resolution: An unsupervised probabilistic model rivaling supervised resolvers. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- [Daumé III et al.2009] Hal Daumé III, John Langford, and Daniel Marcu. 2009. Search-based structured prediction. *Machine learning*, 75(3):297–325.
- [Daumé III et al.2014] Hal Daumé III, John Langford, and Stephane Ross. 2014. Efficient programmable learning to search. *arXiv preprint arXiv:1406.1837*.
- [Grosz et al.1995] Barbara J Grosz, Scott Weinstein, and Aravind K Joshi. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational linguistics*, 21(2):203–225.
- [Haspelmath et al.2005] Martin Haspelmath, Matthew Dryer, David Gil, and Bernard Comrie, editors. 2005. *The World Atlas of Language Structures*. Oxford University Press.
- [Hovy et al.2006] Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: the 90% solution. In *NAACL*.
- [Kong and Zhou2010] Fang Kong and Guodong Zhou. 2010. A tree kernel-based unified framework for Chinese zero anaphora resolution. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 882–891. Association for Computational Linguistics.
- [Kweon2011] Soo-Ok Kweon. 2011. Processing null and overt pronoun subject in ambiguous sentences in Korean. *International Journal of Linguistics*, 3(1).
- [Langford et al.2007] John Langford, Lihong Li, and Alexander Strehl. 2007. Vowpal wabbit online learning project. http://hunch.net/?p=309.
- [Petrov and Klein2007] S. Petrov and D. Klein. 2007. Improved inference for unlexicalized parsing. In *Proceedings of NAACL HLT 2007*, pages 404–411.
- [Rohde and Frank2014] Hannah Rohde and Michael C Frank. 2014. Markers of topical discourse in childdirected speech. *Cognitive science*, 38(8):1634–1661.
- [Ross et al.2011] Stéphane Ross, Geoff J. Gordon, and J. Andrew Bagnell. 2011. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Workshop on Artificial Intelligence and Statistics (AIStats).*

- [Wang2011] Deliang Wang. 2011. Anaphora Resolution in Chinese from the Centering Perspective. Foreign Language Teaching and Research Press.
- [Yeh and Chen2007] Ching-Long Yeh and Yi-Chun Chen. 2007. Zero anaphora resolution in Chinese with shallow parsing. *Journal of Chinese Language and Computing*, 17(1):41–56.
- [Zhao and Ng2007] Shanheng Zhao and Hwee Tou Ng. 2007. Identification and resolution of Chinese zero pronouns: A machine learning approach. In Proceedings of the 2007 Joint Conference on Empirical Methods on Natural Language Processing and Computational Natural Language Learning, pages 541–550.