

Resolving Pronominal References in Chinese with the Hobbs Algorithm

Susan P. Converse

CIS, University of Pennsylvania

spc@seas.upenn.edu

Abstract

This study addresses pronominal anaphora resolution, including zero pronouns, in Chinese. A syntactic, rule-based pronoun resolution algorithm, the “Hobbs algorithm” was run on “gold standard” hand parses from the Penn Chinese Treebank. While first proposed for English, the algorithm counts for its success on two characteristics that Chinese and English have in common. Both languages are SVO, and both are fixed word order languages. No changes were made to adapt the algorithm to Chinese. The accuracy of the algorithm on overt, third-person pronouns at the matrix level was 77.6%, and the accuracy for resolving matrix-level zero pronouns was 73.3%. In contrast, the accuracy of the algorithm on pronouns that appeared in subordinate constructions was only 43.3%, providing support for Miltsakaki’s two-mechanism proposal for resolving inter- vs. intra-sentential anaphors.

1 Introduction

The goal of this study is pronoun resolution, including null/zero pronouns, in Chinese. There has been extensive research for many years into computational approaches to automatic anaphora resolution in English, and increasingly in other languages as well (Mitkov, 1999; Mitkov, 2002).

Yet although there have been countless linguistic studies in Chinese on anaphora and zero anaphora (for example, (Huang, 1984; Huang, 1994; Yang et al., 1999) just to illustrate the range), the published computational work to date is limited to just a few studies (Chen, 1992; Yeh and Chen, 2001; Yeh and Chen, 2005).

In 1978 Jerry Hobbs proposed an algorithm for the resolution of pronominal coreference in English (Hobbs, 1978). The performance of this algorithm has frequently been used as a baseline reference for computational methods in English. The most basic version of the Hobbs algorithm is subject biased, relying on a basic strategy of left-to-right, breadth-first searches, subject to a few structural constraints.

Chinese, like English, is an SVO language. Chinese has also been regarded as a topic-comment language. From either viewpoint, it is worth examining how well the left-to-right, SVO-biased process of the Hobbs algorithm works for Chinese, perhaps so it could be used as a baseline against which to measure other automated approaches to Chinese anaphora resolution.

While Chinese and English are both SVO languages, they differ in another important parameter: Chinese is a pro-drop language, while standard English is not. Thus it will be of particular interest to see how well the Hobbs algorithm performs when proposing antecedents for zero pronouns.

The Hobbs algorithm operates on parsed sentences. In order to evaluate its performance on zero pronouns as well as overt ones, it would be useful to have text that already has the locations of

the zero pronouns marked. Because the Penn Chinese Treebank has overt strings to denote the positions of dropped arguments, test sentences were selected from that corpus.

2 The Corpus and Annotations

The source texts for this study are taken from the first 100K of the CTB 5.0 release of the Penn Chinese Treebank (CTB). The CTB consists of Xinhua news articles that have been segmented, part-of-speech tagged, and bracketed with syntactic labels and functional tags (Xue et al., 2004)¹. In the corpus, zero pronouns are denoted using the string “*pro*”. An example is given in Figure 1.

In order to provide an answer key or “gold standard” against which to test automatic anaphora resolution methods, we are annotating the CTB to indicate the pronominal coreference relations. All third-person pronouns (including 其 (his, hers, its, theirs) and 之 (he/she/it/they)), reflexives, demonstratives, and *pro* are being annotated.

Only those coreference relations that are between these anaphors and *nominal* entities are being co-indexed, however. That is, only *NPs* that denote the same entity as the entity referred to by the anaphor are being co-indexed. Since not every instance of one of these anaphors necessarily refers to a nominal entity, non-corefering anaphors are being tagged with labels that categorize them roughly by type.

The categories are: **DD** (discourse deictic) for anaphors that refer to propositions or events in the text; **EXT** (existential) for *pro* in existential contexts analogous to the English “there is/are”; **INFR** (inferrable) to be put on an anaphor that refers to a specific nominal entity when that entity does not have an overt NP denoting it in the text; **AMB** (ambiguous) when the interpretation of an anaphor is ambiguous between two (or more) referents; and **ARB** (arbitrary) for anaphors that don’t fall into the other categories².

Complex NPs abound in the CTB and present a choice for the placement of the indices and category labels. The decision was made to put the index for a complex NP referent on the entire complex NP rather than on just the head of the phrase

¹<http://www.cis.upenn.edu/~chinese>

²Linguists beware: this is far more general than the arbitrary in “arbitrary PRO”

```
(IP
  (ADVP (AD 同时))
  (PU , )
  (IP(IP(NP#2-SBJ (NP (NP (DP (DT 全))
                        (NP (NN 国)))
                        (NP (NN 城镇)
                          (NN 公房)))
                        (NP (NN 月)
                          (NN 租金)))
      (VP (VE 有)(AS 了)
        (NP-OBJ (CP (WHNP-1 (-NONE- *OP*))
                    (IP (NP-SBJ (-NONE- *T*-1))
                        (VP (ADVP (AD 较))
                          (VP (VA 大))))))
          (NP (NN 提高))))))
  (PU , )
  (IP(NP#2-SBJ (-NONE- *pro*))
    (VP(NP-LOC (QP (CLP (M 部分))
                  (NP (NN 地区)))
      (VP(VRD (VV 提高)(VV 到))
        (IP-OBJ
          (NP-SBJ (-NONE- *PRO*))
          (VP(VV 占)
            (LCP-OBJ
              (QP(DNP(NP(NP(NP(QP (CD 双))
                                (NP (NN 职工))))
                (NP (NN 家庭)))
                (NP (NN 收入)))
              (DEG 的)
              (QP (CD 百分之十))
              (LC 左右))))))))))
  (PU 。 ))
```

At the same time, there has been a comparatively large increase in **the entire country’s monthly rent for public housing in cities and townships**₂, with **that**₂ in a portion of the regions increasing to account for about 10% of the income of dual income families.

Figure 1: Sample of the annotation and example of annotating high.

(that is, to annotate “high” in the NP tree). Figure 1 has such a case. The annotation #2 is placed on the parent NP-SBJ level, rather than at the level of the head (NP (NN 月)(NN 租金) (monthly rent).

The reasoning for this choice was that the full NP unambiguously distinguishes between different nominal entities whose NPs have identical head nouns. Head nouns of complex NPs can always be algorithmically obtained.

3 The Hobbs Algorithm

The “Hobbs Algorithm” was outlined in a paper by Jerry Hobbs in 1978 (Hobbs, 1978). The algorithm is shown in the Appendix. While the algorithm is naive in that the steps proceed merely according to the structure of the parse tree, there are two meta-level points to consider in the execution of the steps. First, the algorithm counts on number and gender agreement. Second, in his paper, Hobbs proposes applying “simple selectional constraints” to the antecedents that the algorithm proposes, and illustrates their use in the sentence he uses to explain the operation of the algorithm:

“The castle in Camelot remained the residence of the king until 536 when he moved *it* to London.”

When trying to resolve the pronoun “it” in this sentence, the algorithm would first propose “536” as the antecedent. But dates cannot move, so on selectional grounds it is ruled out. The algorithm continues and next proposes “the castle” as the antecedent. But castles cannot move any more than dates can, so selectional restrictions rule that choice out as well. Finally, “the residence” is proposed, and does not fail the selectional constraints (although one might find that these “simple” constraints require a fair amount of encoded world knowledge).

In the paper, Hobbs reported the results of testing the algorithm on the pronouns “he”, “she”, “it”³, and “they”, 300 instances in total (100 consecutive pronouns each from three different genres). He found that the algorithm alone worked in 88.3% of the cases, and that the algorithm plus selectional restrictions resolved 91.7% of the cases

³excluding “it” in time or weather constructions, as well as pleonastic and discourse deictic “it”

correctly. But of the 300 examples, only 132 actually had more than one “plausible” antecedent nearby. When he tested the algorithm on just those 132 cases, 96 were resolved by the “naive” algorithm alone, a success rate of 72.7%. When selectional restrictions were added the algorithm correctly resolved 12 more, to give 81.8%.

The Hobbs algorithm was implemented to execute on the CTB. The S label in the CTB is IP, so the two “markable” nodes from the point of view of the algorithm are IP and NP. There were two types of NPs that were excluded, however, NP-TMP and NP-ADV.

No selectional constraints were applied in this experiment. In addition, no gender or number agreement features were used.

While the written versions of Chinese third-person pronouns do have number and gender, and demonstratives have number, there is no morphology on verbs to match. Nor, without extra-syntactic lexical features, are there gender markings on nouns or proper names (the titles in this corpus as a rule do not include gender-specific honorifics).

There is a plural “suffix” (们) on some nouns denoting human groups, and one can sometimes glean number information from determiner phrases modifying head nouns, but no extra coding was done here to do so.

Zero pronouns, of course, provide no clues about gender or number, nor do 其 (his, hers, its, theirs) or 之 (he/she/it/they).

Structurally, there are many sentences in the CTB that consist of just a sequence of parallel independent clauses, separated by commas or semicolons. These multi-clause sentences were treated as single sentences from the point of view of the algorithm.

The implementation of the algorithm is one that has a core of code that can run on either the Penn Treebank (Marcus et al., 1993) or on the Chinese Treebank. The only differences between the two executables are in the tables for the part-of-speech tags and the syntactic phrase labels (e.g., PN vs. PRN for pronouns and IP vs. S for clauses), and in separate NP head-finding routines (not used in the current study).

Despite the SVO similarity between Chinese and English, we were interested to see if there

might be differences in the success of the algorithm due to structural differences between the languages that might require adapting its steps to Chinese. The most obvious place to look was in the placement of modifiers relative to the head noun in an NP. Although unplanned, it turned out that the policy of annotating complex NPs at the parent level rather than at the head noun level actually made this a moot point because of the top-down nature of the tree traversal. That is, because the algorithm proposes an NP that contains both the modifier and the head, differences between English and Chinese in head-modifier word order does not matter.

Another place in which the head-modifier ordering might come into play is in Step 6 of the algorithm. This is still under investigation, since the step did not “fire” in the set of files used here, and only proposed an antecedent once when the algorithm was run on the whole CTB.

4 The Data

As mentioned, in addition to the third person pronouns that Hobbs tested, the algorithm here was run on reflexives, possessives, demonstrative pronouns, and the zero pronoun.

A sample of 95 files, containing a total of 850 sentences (including headlines, but excluding bylines, and excluding the (End) “sentence” at the end of most articles) was used for this experiment.

In all there were 479 anaphors in the 95 files. The distribution of the anaphors for these files is shown in Table 1.

Of the anaphors in the gold standard, 331 (69.1%) were co-indexed with antecedents, while 117 (24.4%) did not corefer with entities denoted by NPs and were categorized. The remaining 6.5%, 31 anaphors (two overt and 29 ***pro***), were marked ambiguous.

Of the anaphors that were co-indexed, just over half (53.2%, 176 pronouns) were overt. In contrast, only 24.8% of the categorized pronouns were overt, and these were usually demonstratives labeled #DD.

5 Results

The performance of the Hobbs algorithm on these data varied considerably depending on the syntac-

tic position of the anaphor, and less so on whether the anaphor was overt or not.

Performance was analyzed separately for pronouns that appeared as matrix subjects (M), pronouns that appeared as subjects of parallel, independent clauses in multi-clause sentences (M2), and pronouns that were found in any kind of subordinate construction (S).

The counts for all anaphors are listed in Table 2 and the counts for third-person pronouns only in Table 3. The scores for third-person pronouns only are given in Table 4 and for all coindexed anaphors in Table 5⁴.

As shown in Table 4, the accuracy for matrix-level, third-person pronouns was 77.6%, while for all pronouns at the matrix level (Table 5) the algorithm achieved a respectable 76.3% accuracy, considering the fact that not only zero pronouns, but reflexives, possessives, and demonstratives are included.

This contrasts with 43.2% correct for third-person pronouns in subordinate constructions and 43.3% correct for all subordinate-level pronouns.

The accuracy for matrix level (M) and independent clause level (M2) combined was 75.7% for third-person pronouns, and 71.6% for all pronouns.

When results are not broken down by the syntactic position of the anaphor, the performance is less impressive, with just 63.2% accuracy for just third-person pronouns and only 53.2% correct for all anaphors at all syntactic levels.

The zero anaphors alone showed the same pattern, with 73.3% correct at the matrix level and 66.7% correct for matrix and matrix2 levels combined (Table 6), but just 42.5% accuracy at the subordinate level.

6 Discussion

The difference in performance of the algorithm by syntactic level clearly suggests that a one-method-fits-all approach (at least in the case of a rule-based method) to anaphora resolution will not succeed, and that further analysis of anaphors at the non-matrix level is in order.

⁴Of the 31 anaphors marked **AMB**, in only eight cases (25.8%) did the algorithm pick an antecedent that was one of the choices given by the annotators. All eight were ***pro***.

Table 1: Distribution of all anaphors

	Indexed		AMB		Cat, no AMB		Total	
	count	%	count	%	count	%	counts	%
overt	176	53.2%	2	6.5%	29	24.8%	207	43.2%
pro	155	46.8%	29	93.5%	88	75.2%	272	56.8%
	331	69.1%	31	6.5%	117	24.4%	479	

Table 2: Counts by syntactic level for all anaphors

	Correct			Wrong			Cat(incl AMB)			Total
	M	M2	S	M	M2	S	M	M2	S	
overt	47	8	45	14	5	57	10	2	19	207
pro	11	17	48	4	10	65	15	21	81	272
	58	25	93	18	15	122	25	23	100	479

These data are consistent with the observations made by Miltsakaki in her 2002 paper (Miltsakaki, 2002). Taking a main clause and all its dependent clauses as a unit, she found that there were different mechanisms needed to account for (1) topic continuity from unit to unit (inter-sentential), and (2) focusing preferences within a unit (intra-sentential). Topic continuity was best modeled structurally but the semantics and pragmatics of verbs and connectives were prominent within a unit.

Since inter-sentential anaphoric links relate to topic continuity, structural rules work best for resolution at the matrix level, while anaphors in subordinate clauses are subject to the semantic/pragmatic constraints of the predicates and connectives.

In our results the anaphors that are subjects of matrix clauses tend to resolve inter-sententially (that is, Step 4 of the algorithm is the resolving condition), while the anaphors in subordinate constructions are more likely to have intra-sentential antecedents. That the strictly structural version of the Hobbs algorithm used here performed better for matrix-level anaphors and did not do well at all on anaphors in subordinate constructions agrees with Miltsakaki's findings.

In our data the "unit" is not always a single main clause with its dependent clauses, however. In the M2 case, the unit is a sentence containing parallel main clauses, each of which may have its own dependent clauses. An examination of

Table 3: Counts by syntactic level for third-person pronouns

	Correct			Wrong			Tot.
	M	M2	S	M	M2	S	
he	38	4	13	8	1	12	76
he-they	1	1	2	3	2	3	12
she	1	-	-	-	-	-	1
she-they	-	-	-	-	-	-	0
it	4	3	4	2	1	8	22
it-they	1	-	-	-	-	2	3
	45	8	19	13	4	25	114

the errors made for these M2 cases might show that an improvement of performance for these anaphors could be obtained by treating the independent clauses as separate sentences.

7 Future Work

As mentioned in Section 3 above, in addition to some limited information that can be obtained from just the parses to check for number agreement, a logical next step is to implement some selectional constraints. For example, the semantic categories in the Rocling dictionary could be used in combination with selectional restrictions on verb arguments. Simple features such as animacy or concreteness could prevent some incorrect choices by the algorithm.

We also plan to investigate the operation of Step 6 with respect to the original intent of the

Table 4: Performance by syntactic level: third-person pronouns

	Correct	Wrong	Total
matrix (M)	45	13	58
	77.6%	22.4%	
matrix2 (M2)	8	4	12
	66.7%	33.3%	
subord. (S)	19	25	44
	43.2%	56.8%	
All 3 levels	72	42	114
	63.2%	36.8%	
M + M2	75.7%	24.3%	

Table 5: Performance by syntactic level: coindexed overt and *pro*

	Correct	Wrong	Total
matrix (M)	58	18	86
	76.3%	23.7%	
matrix2 (M2)	25	15	40
	62.5%	37.5%	
subord. (S)	93	122	215
	43.3%	56.7%	
All 3 levels	176	155	331
	53.2%	46.8%	
M + M2	71.6%	28.4%	

Table 6: Performance by syntactic level: *pro* alone

	Correct	Wrong	Total
matrix (M)	11	4	15
	73.3%	26.7%	
matrix2 (M2)	17	10	27
	63.0%	37.0%	
subord. (S)	48	65	113
	42.5%	57.5%	
All 3 levels	76	79	155
	49.0%	51.0%	
M + M2	66.7%	33.3%	

rule, the bracketing conventions used in the CTB, and the difference in the headedness of NPs between Chinese and English.

As discussed in Section 6, the performance of this rule-based algorithm on subordinate-level anaphors confirmed Miltsakaki's observations about the need for different models for inter- vs. intra-sentential anaphora resolution. We therefore plan to investigate alternative strategies for the resolution of subordinate-level anaphors.

8 Appendix

Naive Hobbs Algorithm

This algorithm traverses the surface parse tree, searching for a noun phrase *of the correct gender and number* using the following traversal order:

- 1) begin at NP node immediately dominating the pronoun
- 2) go up the tree to the first NP or S node encountered
call this node X
call the path to reach X "p"
- 3) traverse all branches below node X to the left of path p,
in left-to-right, breadth-first manner

propose as the antecedent any NP node that is encountered that has an NP or an S node between it and X
- 4) if node X is the highest S node in the sentence
traverse the parse trees of previous sentences in order of recency (the most recent first), from left-to-right, breadth-first and
propose as antecedent the first NP encountered

else goto step (5)
- 5) from node X go up the tree to the first NP or S node encountered
call this new node 'X', and
call the path traversed to reach it from the original X 'p'
- 6) if X is an NP node AND
if the path p to X did not pass through the N-bar node that X immediately dominates,
propose X as the antecedent
- 7) traverse all branches below node X to the *left* of path p,
left-to-right, breadth-first

- propose any NP node encountered as the antecedent
- 8) if X is an S node
 traverse all branches of node X to the *right* of path p, 'left-to-right, breadth-first, but do not go below any NP or S encountered
- propose any NP node encountered as the antecedent
- 9) goto step 4

References

- Hsin-Hsi Chen. 1992. The transfer of anaphors in translation. *Literary and Linguistic Computing*, 7(4):231–238.
- Jerry Hobbs. 1978. Resolving pronoun references. *Lingua*, 44:311–338.
- C.-T. James Huang. 1984. On the distribution and reference of empty pronouns. *Linguistic Inquiry*, 5(4):531–574.
- Yan Huang. 1994. *The Syntax and Pragmatics of Anaphora. A study with special reference to Chinese*. Cambridge University Press, Cambridge.
- Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.
- Eleni Miltsakaki. 2002. Toward an aposynthesis of topic continuity and intrasentential anaphora. *Computational Linguistics*, 28(3):319–355.
- Ruslan Mitkov, editor. 1999. *Machine Translation: Special Issue on Anaphora Resolution in Machine Translation*, volume 14. numbers 3-4.
- Ruslan Mitkov. 2002. *Anaphora Resolution*. Longman, London.
- Nianwen Xue, Fei Xia, Fu-Dong Chiou, and Martha Palmer. 2004. The Penn Chinese Treebank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 10(4):1–30. <http://www.cis.upenn.edu/~chinese/>.
- Chin Lung Yang, Peter C. Gordon, and Randall Hendrick. 1999. Comprehension of referring expressions in Chinese. *Language and Cognitive Processes*, 14(5/6):715–742.
- Ching-Long Yeh and Yi-Chun Chen. 2001. An empirical study of zero anaphora resolution in Chinese based on centering model. In *Proceedings of ROCLING*.
- Ching-Long Yeh and Yi-Chun Chen. 2005. Zero anaphora resolution in Chinese with shallow parsing. *Journal of Chinese Language and Computing*, to appear.