Developing Guidelines for the Annotation of Anaphors in the Chinese Treebank

Susan Converse
CIS, University of Pennsylvania
spc@linc.cis.upenn.edu

Abstract

This paper describes the *CTB Coreference Annotation Guidelines* for annotating pronominal anaphoric expressions in the Penn Chinese Treebank. The goals of the annotation are: to provide training data for learning-based pronoun resolution tools, and to provide a "gold" standard to be used in the evaluation of pronoun resolution algorithms. The choices that were made concerning the coindexing of pronominal anaphors and their antecedents are discussed, as are some questions that arose in trying to categorize those pronominal expressions that did *not* refer to specific nominal entities in the text.

1 Introduction

This paper describes the *CTB Coreference Annotation Guidelines* for annotating the pronominal anaphoric expressions in the Penn Chinese Treebank (CTB). The purpose of the annotation is twofold: to provide training data for learning-based pronoun resolution tools, and to provide a "gold" standard to use in evaluating the outputs of pronoun resolution algorithms.

There were two annotation tasks: coindexing anaphors with NP antecedents whenever possible, and classifying the non-coindexable anaphors into categories. While the coindexing of pronominal anaphors (including the zero pronoun) with their antecedents (or postcedents) was fairly straightforward, the categorization of those pronominal expressions that did *not* refer to nominal entities in the text raised a some questions which are discussed below.

2 What we are annotating

Annotations were made to the parsed files of the June, 2001 corrected release of the Penn Chinese Treebank (Xia et al., 2000). Example 1 shows a sample of the annotation, with both an overt pronoun 其 ("their" in this context), and a zero pronoun "*pro*" referring to a nominal entity. The coindexation (**#2**) will be discussed in Section 4 and the annotation **#EXT** on the second *pro* will be discussed below in Section 5.1.3.

```
(IP (NP#2-TPC (ADJP (JJ 国有))
              (NP (NN 企业)))
    (NP-SBJ (NN 负债))
    (VP (VV 高达)
        (QP-OBJ (CD 七千亿)(CLP (M 元)))))
(PU ， )
(IP (NP-SBJ (NP#2 (PN 其))   overt pronoun "their"
            (NP (NN 产品)(NN 质量)))
    (VP (ADVP (AD 也))
        (VP (VV 参差不齐))))
(PU ， )                                  dropped subject
(IP (NP#2-SBJ (-NONE- *pro*))             zero pronoun
    (VP (IP-CND
            (NP#EXT-SBJ (-NONE- *pro*)) zero that
            (VP (ADVP (CS 如))      does not co-refer
                (VP (VE 无)
                    (NP-OBJ (NN 改进)))))
        (PU ， )
        (ADVP (AD 很))(ADVP (AD 难))
        (PP-LOC (P 在)
                (NP (NN 海外)(NN 市场)))
        (VP (VV 参与)
            (NP-OBJ (NN 竞争)))))
```

The debt of state-owned enterprises has reached as high as 700 billion yuan, and **their** product quality is uneven. If there is no improvement, **they** will will find it very difficult to participate in the competition in overseas markets.

Example 1: Sample of the annotation.

2.1 Which anaphoric expressions

Consistent with work in English (e.g., (Dagan and Itai, 1991; Ge et al., 1998; Lappin

and Leass, 1994; Morton, 2000)), only third-person pronouns and demonstratives were chosen to be annotated. The pronouns include reflexives (e.g., 自己/"oneself"), reciprocals (e.g., 彼此/"one another"), and possessives. For the possessives, the only distinct possessive forms are 其 and 之 (both "its/his/hers/theirs"), which are annotated. Ordinary third-person pronouns in possessive constructions are annotated just as they would be in other contexts.

Unlike English, however, Chinese is a pro-drop language, and there are zero pronouns in addition to the overt ones. For pro drop, the *Bracketing Guidelines* for the CTB (Xue and Xia, 2000) specify use of the string "*pro*" (*small pro*) to explicitly denote a zero pronoun in a dropped subject or dropped object position in a parse tree.

Not all empty subject positions in the CTB parses are the result of pro drop, however. The *Bracketing Guidelines* specify use of the string "*PRO*" (*big PRO*) when there is a non-overt subject in a clause that is the complement of a control verb, such as 决定/"decide" (subject control) or 使/"cause" (object control)[1].

The *Guidelines* (Section VI.4) also state that *PRO* is to be used as the non-overt subject in "non-finite clauses," typically clauses that are themselves sentential subjects. Big *PRO* is in complementary distribution with lexicalized constituents[2], while *pro* is not so restricted.

The coreference annotation task only deals with the zero pronouns, or small *pro*, in the CTB parses. The big *PRO*s are ignored, just as they are in a coreference task in English.

2.2 What kind of antecedents

We specify that anaphors are to be coindexed only with entities that are represented in the text by lexicalized noun phrases, or in some cases quantifier phrases, QPs.

Antecedent entities that must be inferred, or antecedents that are events or propositions, for example, cannot be coindexed with any overt NP, and are to be labeled with one of the categories described in Section 5.

[1]Note that *pro* and *PRO* do not represent movement. Movement is denoted using other empty category labels. Refer to the *Bracketing Guidelines*, Section VI: "Null Elements."

[2]This is consistent with Government and Binding Theory. See, for example, (Haegeman, 1994).

3 Two tasks: coindexing and categorizing

As in other languages, pronouns not only are used to refer to explicit nominal phrases in the text, but also are used discourse deictically and in an existential manner. In addition, they can refer to nominal entities that are not explicitly mentioned in the text, but that are available through inference.

The approach taken to annotation was therefore twofold.

The first task was to classify each overt pronoun, demonstrative, and small pro either as a coindexable anaphor that was coreferential with a noun phrase that explicitly appeared in the text, or as an anaphor that belonged to one of a set of categories of non-coindexable anaphors. The second task was to assign indices to the coindexable anaphors and their antecedents.

We will first describe the questions that arose in the coindexation task, then discuss the categorization of the non-coindexable pronouns.

4 Choices in coindexing overt nominal antecedents

4.1 Nesting level

When NPs are nested, such that an embedded NP refers to the same entity as the NP that contains it, there is a choice between annotating the head NP at the lowest level *vs.* annotating the top NP. The choice to annotate high, rather than at the head NP, was made for two reasons. One was that the head may be automatically recovered from the parent. The second was that including the modifying material uniquely identifies entities that are different but that have identical lexical heads.

In Example 2, the overt NPs that are tagged with the indices #3 and #4 have identical head NPs, but one is actually a subset of the other.

4.2 NP apposition

The only exception to annotating high was in a particular kind of construction with NP apposition, namely those NP-APP that had the pattern

```
(NP (NP-APP (title/descriptive phrase))
    (NP (name)))
```

In that case, either the title or full descriptive NP alone referred to the same entity denoted by the name, and the taggers were asked

to annotate both the NP-APP and the head NP containing the name.

In more complex appositive constructions, the NP-APP usually was a list, ending in **etc.**, of sample members of a set that was denoted by the head noun. Since the list was exemplary, rather than definitive, NPs containing such lists were to be annotated only at the parent level. That is, the NP-APP constituent was treated as another modifier, rather than as a referent.

4.3 NP-Predicate constructions

When the antecedent of a pronoun, demonstrative, or *pro* is a subject that is followed by the copula 是 or 为 and a predicate NP (NP-PRD) with *definite* reference, then the annotators were instructed to put the entity's index not only on the subject, but also the NP-PRD. Example 3 illustrates this.

Analogous to this, there are verbs that establish identity between their subjects and objects, usually after a change of state. These verbs are words like "become" (成为) or "establish" (建成). When subjects in these constructions were antecedents, the annotators were asked to index any NP-OBJ with definite reference as well.

Example 4 shows one such instance with the verb 成为/"become."

4.4 Quantifier phrases

Although syntactically, quantifier phrases (QPs) occur in the same contexts as NPs in the CTB, they are not necessarily equivalent in their semantic usage. While an NP denotes an entity, a QP usually denotes the value of a quantifiable property or attribute of an entity, rather than the entity itself.

Note that in Example 2 the QP-OBJ "109.82 billion US$" in the first clause was not annotated. Entity #3 is "the value of imports and exports", and the extent of that value is 109.82 billion at one particular time. But the measure does not uniquely denote the entity.

On the other hand, quantifier phrases may sometimes be abbreviated expressions denoting a particular set of entities, in which case the QP constituent is annotated as if it were an NP.

Example 5 illustrates such a case. The QP#1-SBJ is shorthand for 七十家企业/"70 enterprises".

```
(IP(NP#3-SBJ
      (CP(WHNP-1 (-NONE- *OP*))
         (IP (NP#2-SBJ (-NONE- *pro*))
            (VP (NP-TMP (NT 去年))                    last year
               (VP (VV 实现)                          realize
                  (NP-OBJ (-NONE- *T*-1))))))
      (NP (NN 进出口)                                 imports+exports
         (NN 总值)))                                  total value
   (VP (VV 达)                                        reach
      (QP-OBJ (CD 一千零九十八点二亿)                  109.82 B
         (CLP (M 美元)))))                            US$
(PU ， )
(IP(NP-SBJ
      (CP (WHNP-2 (-NONE- *OP*))
         (CP (IP (NP#3-SBJ (-NONE- *pro*))
            (VP (VV 占)                               stand
               (NP#4-OBJ (NP (DP (DT 全))
                             (NP (NN 国)))
                                                      entire country
                         (NP (NN 进出口)
                             (NN 总值)))
                                                      imports+exports
                                                      total value
                  (NP-EXT (-NONE- *T*-2))))
            (DEC 的)))
      (NP (NN 比重)))                                 proportion
   (VP (VRD (VV 提高)(VV 到))                         rise up to
      (QP-OBJ (CD 百分之三十九)))))                   39%
```

NP#2 is 外商投资企业/"foreign-owned enterprises"

The **value of imports and exports** that they realized last year reached 109.82 billion US$. The proportion that this represented in the **value of imports and exports** of the entire country rose to 39%.

Example 2: Annotating parent instead of head NP.

```
(IP (NP#2-PN-SBJ (NR 崇明))
   (VP (VC 是)
      (NP#2-PRD (NP-PN (NR 中国))
                (QP (OD 第三))
                (ADJP (JJ 大))
                (NP (NN 岛)))))
```

Chongming is China's third largest island

Example 3: NP-PRD with definite reference.

5 Categorizing non-coindexable anaphors

As mentioned above, not all pronominal expressions are coreferential with overt NP/QP constituents that denote specific nominal entities.

```
(IP(NP#2-SBJ (-NONE- *pro*))
   (VP(VV 成为)
      (NP#2-OBJ (DNP(NP (NP-PN (NR 无锡市))
                        (NP (NN 经济)
                            (NN 发展)))
                     (DEG 的))
                (ADJP (JJ 主))
                (NP (NN 骨架)))))
```

NP#2 is 超亿元产值的大型企业/"large-scale enterprises with production value exceeding 100 M yuan"

They have become the backbone of Wuxi Municipality's economic development

Example 4: Predicative NP-OBJ.

```
(IP (NP-SBJ (ADJP (JJ 入区))
            (NP (NN 企业)))
    (VP (QP-PRD (CD 一百二十二)
                (CLP (M 家)))))
(PU ，)
(IP (NP#EXT-SBJ (-NONE- *pro*))
    (VP (ADVP (AD 已))
        (VP (VE 有)
            (IP-OBJ
                (QP#1-SBJ (CD 七十)
                          (CLP (M 家)))
                (VP (VV 拥有)
                    (NP-OBJ
                        (DNP(NP#1 (PN 自己))
                            (DEG 的))
                        (NP (NN 产品)))))))))
```

Enterprises entering the region number 122, and already **there** are **seventy** that have their own products.

Example 5: An annotated quantifier phrase and an existential subject.

An attempt was made to classify these "non-coindexable" anaphors.

We therefore defined a set of categories with which to label those anaphors that could not be coindexed. Table 1 provides a brief summary of the categories. A discussion of each category follows.

5.1 The straightforward cases

5.1.1 AMB: ambiguous

This category is intended to identify cases in which a particular anaphor, whether an overt lexical item or a zero pronoun, could have more than one possible interpretation, depending on the understanding of the reader. The ambiguity may be between two or more possible NP antecedents in the text for example, or it may be between an entity denoted by an NP in the text and an entity that is not lexicalized but can be inferred.

5.1.2 DD: discourse deictic

Pronouns are often used to refer to events or propositions as well as to entities. When an anaphor refers to an event or proposition or to the meaning of an entire span of text, then it is labeled **DD**, for *discourse deictic*. Note that in addition to the demonstratives 这/"this" and 此/"this" that one would expect to be used for this purpose, there were situations in which *pro* was tagged with this label as well.

5.1.3 EXT: existential

Analogous to pleonastic constructions in English, the verbs (没)有/"(not)have" and 无/"without" in Chinese can be used in an "existential" manner. When they are, they have the part-of-speech tag VE[3]. The label **EXT** is used to tag *pro* when it is the subject of one of these verbs (and the verb has the reading "there are" or "there are no"). Examples 1 and 5 contain *pro*'s annotated in this way.

5.1.4 INFR: inferrable

The coindexing task is very restricted in that the only antecedents that can be assigned indices are NP constituents in the same document that denote the same entity that the anaphor refers to. The label **INFR** is to be used when the anaphor refers to an entity that is not itself explicitly lexicalized in the text, but can be inferred from the context or using real-world knowledge. The entity must be a particular nominal entity, not a generic "anyperson" or "anything" with arbitrary reading (see Section 5.2, below).

For example, the pronoun 他们/"they" might be used to refer to the people in charge of some place/institution, but the only overt NP in the text is the place/institution itself. The real antecedent of the pronoun, a reference to the people themselves, is not overtly mentioned in the text, so the anaphor would be labeled INFR.

[3]VE denotes a verb in an existential construction. Refer to the CTB *Bracketing Guidelines*, Section IV.4.1.

Table 1: Categories for pronouns that do not have a specific or clear coreferent

AMB	**ambiguous**: for anaphors that can have more than one possible antecedent, depending on the interpretation of the text
DD	**discourse deictic**: the pronominal expression refers to an event, situation, or proposition rather than to a nominal entity
EXT	**existential**: a *pro* that is the subject of an existential (VE) 有 or 无 or 存在
INFR	**inferrable**: a pronoun or *pro* that refers to a particular nominal entity that is not explicitly lexicalized in the text, but that can be *inferred* from the text
ARB	**arbitrary**: a *pro* that does not refer to a specific entity that can be inferred from or is lexicalized in the text, but instead takes an "arbitrary" reading

5.2 More subtle cases of *pro*

5.2.1 The choice of the category ARB

The above categories describe the manifestations of discourse phenomena that are familiar to English speakers as well as to Chinese speakers. The final category applies only to small *pro*, and is more specific to the phenomena observed in the parses in CTB-I.

In the first version of the guidelines, there were five categories: the four just mentioned, plus an "ANONymous" category that was meant to account for arbitrary dropped subjects in titles or slogans (analogous to the use of "one" in English). During the initial adjudication meetings, however, it became clear that there were occurrences of small *pro* that did not fall neatly into any of the five proposed categories.

Many of these *pro*'s appeared in subordinate constructions that functioned like English infinitives and gerundive clauses, and they frequently could *not* be replaced by any overt NP, even a pronoun. For these reasons, a sixth category, labeled UNTNS for untensed, was added to the guidelines in an attempt to capture a meaningful category of dropped subjects.

Unfortunately, adding this category caused more problems than it solved. First, it was too hard to clearly define. The reason it was difficult to define is the same reason that the question of a finite *vs.* non-finite distinction in Chinese is so problematic: there is no reliable diagnostic of (non-)finiteness in Chinese (and, some would argue, no non-finite/finite distinction at all (Hu et al., 2001)).

In the absence of diagnostic tests, the *Guidelines* described the category using sample syntactic structures. Although there are a few syntactic structures such as sentence-initial purpose clauses that appear to be "prototypical" untensed constructions, there are many more structures that require more subjective judgements.

Second, even if a diagnostic test for UNTNS were available, there were two basic difficulties with adding the UNTNS category. The first problem was that just as big PRO can be coindexed (e.g., in subject or object control contexts), in a similar way the *pro* in the "untensed" syntactic structures could sometimes be coindexed with an NP. This presented a problem with consistently annotating such a *pro*: should it be coindexed or marked UNTNS?

The second difficulty with just adding UNTNS to the coreference *Guidelines* was that both of the categories ANON and UNTNS contained *pro*'s with arbitrary reference, and there was no clean way to differentiate them. A diagnostic that an ANON *pro* could be substituted by an overt NP was not sufficient.

As a result of these two overlaps (UNTNS that could be coindexed, and arbitrary references in both ANON and UNTNS), there was reliably poor inter-annotator agreement in annotating *pro*'s belonging to these categories.

To simplify the categorization, therefore, the decision was made on the one hand to replace the ANON and the UNTNS categories with the single category ARB (for "arbitrary"), and on the other hand to instruct the annotators to always coindex a *pro* when possible. That is, any *pro* that can be coindexed, whether or not it appears in an "infinitival" or "gerundive" construction, and whether or not it can be replaced by an overt NP, should be coindexed. Only those *pro*'s that have arbitrary readings, with no possible specific or inferrable

```
(IP(PP (P 据)                                   according to
       (IP (NP#ARB-SBJ (-NONE- *pro*))
           (VP (VV 认为))))                      understand
   (PU ，)
   (NP-SBJ (DNP (NP (DP (DT 此)                  this
                        (CLP (M 次)))            M
                    (NP (NN 访问)))              visit
               (DEG 的))                         DE
           (NP (NN 目的)))                       goal
   (VP(VC 是)                                    is
      (PP-PRD
        (P 为了)                                 in order to
        (IP (NP#ARB-SBJ (-NONE- *pro*))
            (VP
              (VP (VV 加强)                      strengthen
                  (NP-OBJ
                      (ADJP (JJ 双边))           both sides
                      (NP (NN 经贸)              econ.+trade
                          (NN 合作))))           cooperation
              (PU ，)
              (VP (VV 扩大)                      expand
              ...                                etc.
```

It is understood that the purpose of this visit is to strengthen bilateral economic and trade cooperation and to expand ...

Example 6: Arbitrary reference and zero subject of untensed clause.

antecedent, will be marked **ARB**.

With respect to training a tool to do automatic coreference resolution of *pro* this annotation decision will overgenerate coindexed *pro*'s in the data, since some of them probably should actually be big *PRO*s. Against that negative are the pluses of easier choices for the annotators and better inter-annotator agreement.

5.2.2 Some examples

There are cases in which a small *pro* does not refer to a single, specific entity, but could be replaced by an overt, generic expression, such as 某人/"someone" or 我们/"we" in the general sense. Example 6 illustrates one such situation in the complement of a sentence-initial PP. The same example illustrates another use of ARB in a clause that could be considered "infinitival".

An example of a *pro* in an untensed construction that *can* be coindexed is shown in Example 7. It is usually possible to coindex a *pro* in these clause-initial modifiers at the sentence level with the matrix subject. The most common clauses with dropped subjects in this position are sentence-initial purpose clauses with 为/"in order to" and adverbial IPs with verbs like 利用/"utilize".

```
(IP (PP-PRP
       (P 为)                                    in order to
       (IP (NP#4-SBJ (-NONE- *pro*))             (former UNTNS)
           (VP (VV 规范)                         standardize
               (NP-OBJ (NN 建筑)                 construction
                       (NN 行为)))               behavior
   (PU ，)
   (NP#4-SBJ (NN 新区)                           new region
             (NN 管委会))                        management committee
   (VP (VV 出台)(AS 了)                          announced
       ...
```

In order to standardize the construction procedures, the management committee of the new region announced ...

Example 7: Sentence-initial purpose clause.

```
(IP (IP-SBJ (NP#ARB-SBJ (-NONE- *pro*))
            (VP (VV 利用)
                (NP-OBJ (NN 外资))))
    (VP (VV 趋向)
        (NP-OBJ (NN 多元化))))
(PU ，)
(IP (NP-SBJ (QP (CD 大批))
            (NP (NN 特区)(NN 企业)))
    (VP (VV 走向)
        (NP-OBJ (NP (NN 国际)
                    (NN 资本)
                    (NN 市场))
                (ADJP (JJ 直接))
                    (NP (NN 融资)))))
```

The use of foreign capital is showing a trend toward diversification, and a large number of enterprises in the special zone are entering the international capital market for direct financing.

Example 8: An untensed IP-SBJ.

Example 8 illustrates a *pro* subject of an IP that is itself the matrix subject.

Pre-verbal modifier prepositional phrases with IP complements, like the sentence-level adverbials, sometimes act like infinitives and sometimes like gerunds. Some have *pro* subjects that can be coindexed, often with the matrix subject (such as benefactive prepositional phrases headed by 为/"for"). Others, such as manner PPs with 以/"using" appear less likely to be coindexable.

6 Annotation process/status

Initially 40 files were annotated by two native speakers of Mandarin, both from Taipei, Taiwan. One is a trained linguist, the other is a computer science graduate student who has taken a syntax course. In the process of annotating and adjudicating, the *Guidelines* were revised, to broaden the definition of INFR from a more narrow interpretation and to add the UNTNS category and distinguish it from the ANON category.

Those *Guidelines* were used by two annotators to independently annotate all the remaining files in CTB-I for coindexation only. The annotators are two native speakers who are trained linguists. One is from the Peoples Republic of China, the other is the linguist from Taiwan who did the first 40 files.

The *Guidelines* were again recently revised to document the replacement of ANON and UNTNS by ARB and the choice to coindex whenever possible. These changes are currently being integrated into "gold" files as the files that already have been coindexed are adjudicated.

A third native speaker from the PRC is independently annotating a portion of CTB-I using this last revision of the *Guidelines*, and her annotations will be compared to the "gold" files to calculate some inter-annotator agreement numbers for this last version of the *Guidelines*.

7 Future work

The CTB-I is a rather limited corpus, both in size (100K words) and in linguistic structures. It is very declarative and concrete, and thus the range of figurative language and "messy" linguistic phenomena is relatively restricted. For example, there were no examples of object drop in the CTB-I, although there are examples in CTB-II.

The expectation is that when a more balanced corpus is tagged, the *Guidelines* will need to be refined with more specific instructions concerning the handling of such things as more complex plural antecedents, quoted speech, metonymy, and other figurative expressions.

8 Acknowledgements

Many thanks to Fu-dong Chiou, Nianwen Xue, Szu-ting Yi, Jennifer Chia, Betsy Klipple, and Martha Palmer.

References

Ido Dagan and Alon Itai, 1991. *A Statistical Filter for Resolving Pronoun References*, pages 125–135. Elsevier Science Publishers.

Niyu Ge, John Hale, and Eugene Charniak. 1998. A statistical approach to anaphora resolution. In *Proceedings of the Sixth Workshop on Very Large Corpora*, pages 161–170, University of Montreal, August.

Liliane Haegeman. 1994. *Introduction to Government & Binding Theory*. Blackwell, Cambridge, MA, 2nd edition.

Jianhua Hu, Haihua Pan, and Liejiong Xu. 2001. Is there a finite vs. nonfinite distinction in Chinese? *Linguistics*, 39:1117–1148.

Shalom Lappin and Herbert Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561. Lappin and Leass note to test.

Thomas S. Morton. 2000. Coreference for nlp applications. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 173–180, Hong Kong, October.

Fei Xia, Martha Palmer, Nianwen Xue, Mary Ellen Okurowski, John Kovarik, Fu-Dong Chiou, Shizhe Huang, Tony Kroch, and Mitch Marcus. 2000. Developing guidelines and ensuring consistency for Chinese text annotation. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000)*, Athens, Greece, May 30.

Nianwen Xue and Fei Xia. 2000. *The Bracketing Guidelines for the Penn Chinese Treebank Project*. Number 00-08. Technical Report, Institute for Research in Cognitive Science, University of Pennsylvania. **http://www.ldc.upenn.edu/ctb/**.