The Coding Scheme for Annotating Extended Nominal Coreference and Bridging Anaphora in the Prague Dependency Treebank

Anna Nedoluzhko, Jiří Mírovský, Petr Pajas
Charles University in Prague
Institute of Formal and Applied Linguistics
{nedoluzko, mirovsky, pajas}@ufal.mff.cuni.cz

Abstract

The present paper outlines an ongoing project of annotation of the extended nominal coreference and the bridging anaphora in the Prague Dependency Treebank. We describe the annotation scheme with respect to the linguistic classification of coreferential and bridging relations and focus also on details of the annotation process from the technical point of view. We present methods of helping the annotators – by a pre-annotation and by several useful features implemented in the annotation tool. Our method of the inter-annotator agreement is focused on the improvement of the annotation guidelines; we present results of three subsequent measurements of the agreement.

1 Introduction

The Prague Dependency Treebank (PDT 2.0) is a large collection of linguistically annotated data and documentation (Hajič et al., 2006). In PDT 2.0, Czech newspaper texts are annotated on three layers. The most abstract (tectogrammatical) layer includes the annotation of coreferential links of two types: grammatical coreference (typically within a single sentence) and textual coreference (for pronominal and zero anaphora). The current paper focuses on the present annotation of extended textual coreference, where the anaphoric expression is neither personal pronoun, nor zero. Also the annotation of bridging anaphora on PDT is discussed.

In the last few years, a number of annotation schemes have been released, three of which are to be shortly presented here. The MUC is considered to be the most standard annotation scheme (Hirschman, 1997) and it is used in more than one application (MUC-6, MUC-7, ACE). The advantage of this scheme is its simplicity and a very detailed linguistically oriented coding scheme. It has been however criticized for its vague interpretation of the notion of coreference and for the limited coverage of relations (only identical relation between nouns is annotated). One of the most well known later approaches is MATE (Poesio, 2004) and its extension on the GNOME corpus. The project is meant to be multi-functional. The annotation scheme was primarily developed for dialog acts analyses, but may be easily adapted for any other investigation. In the extended GNOME scheme, the identical coreference is annotated along with some bridging relations, such as ELEMENT, SUBSET, POSSession and OTHER for underspecified relations. In PoCoS (Krasavina and Chiarchos, 2007), a two layer coreference annotation scheme was suggested: the Core Layer is general and reusable, while the Extended Layer supports a wider range of specific extensions.

In this document, we present the application of coreference annotation on a slavonic language (Czech). Czech has no definite article, so in many cases, an anaphoric relation cannot be easily identified. That's why we concentrated solely on coreference, i.e. on the case when two expressions denote the same entity. Anaphoric relation between non-coreferential objects is annotated separately, together with some other types of bridging anaphora (see 2.1).

2 Methods of coreference and bridging anaphora annotation

Subject to annotation are pairs of coreferring expressions, the preceding expression is called antecedent, the subsequent one is called anaphor.

The (mostly manual) annotation of the extended coreference and bridging anaphora proceeds basically in one phase. Unlike MUC/MATE/PoCoS projects, where annotation is divided into two phases (identifying elements that can come in coreference relation (so called "markables") and establishing anaphoric relation), we do not make preliminary annotation of "markables". Realizing the disadvantage of difficult agreement comparison, we still think that to separate identifying "markables" is unnecessary in case of a language without grammatical category of definiteness.

2.1 The annotation scheme

For the time being, we annotate textual coreference and bridging anaphora. In what follows, we briefly present the classification of these two types of context-dependences.

The cases where anaphor is a personal, demonstrative or zero pronoun are already annotated in PDT. In the present annotation, the most cases of anaphoric expressions are expressed by NP with nominal head, in some cases also by pronominal demonstrative adverbs (*there, then* etc.), adjectives (by named entities (e.g. *Germany – German*) and possessive forms)), numerals or verbs (*own – ownership*), see ex. (1).

Textual coreference is further classified into two types – coreference of NPs with specific or generic coreference. This decision is made on the basis of the expectation, that generic coreferential chains have different anaphoric rules from the specific ones. Into this group, there is also included a big number of abstract nouns, whose coreference is not quite clear in every particular case. So, the generic type of textual coreference serves as the ambiguity group too.

In **bridging anaphora** we distinguish PART, SUBSET and FUNCT traditional relations (see e.g. Clark 1977), CONTRAST for coherence relevant discourse opposites (e.g. *People don't chew, it's cows who chew*) and further underspecified group REST, which is used for capturing bridging references – potential candidates for a new bridging group (e.g. location – resident, relatives, event – argument and some others).

2.2 Annotation Principles

In order to develop maximally consistent annotation scheme, we follow a number of basic principles. Some of them are presented below:

Chain principle: coreference relations in text are organized in ordered chains. The most recent mention of a referent is marked as antecedent. This principle is controlled automatically (see 3.1.2). Chain principle does not concern bridging anaphora.

Principle of the **maximum length of coreferencial chains** also concerns only the case of coreference. It says that in case of multiple choice, we prefer to continue the existing coreference chain, rather than to begin a new one. To satisfy this principle, grammatical coreferential chains are being continued by textual ones, and already annotated textual coreferences are continued by currently annotated non-pronominal links in turn.

The principle of **maximal size of an anaforic expression**: subject to annotation is always the whole subtree of the antecedent/anaphor. This principle is partially directed by the dependency structure of tectogrammatical trees and may be sometimes counter-intuitive. See ex. (1):

(1) Henry's brother Nicholas has owned the Hall for 27 years. On Nicholas' death, it passed into the ownership of his nephew, Yarburgh Greame

The principle of **cooperation with the syntactic structure of a given dependency tree**: we do not annotate relations, which are already caught up by the syntactic structure of the tectogrammatical tree. So, unlike most schemes, we do not annotate predication and apposition relations.

Preference of coreference over bridging anaphora: in case of multiple choice, we prefer coreference.

3 The Tool and Data Format

The primary format of PDT 2.0 is called PML. It is an abstract XML-based format designed for annotation of treebanks. For editing and processing data in PML format, a fully customizable tree editor TrEd has been implemented (Pajas & Štěpánek 2008).

TrEd can be easily customized to a desired purpose by extensions that are included into the system as modules. In this section, we describe some features of an extension that has been implemented for our purposes.

The data scheme used in PDT 2.0 has been slightly extended to support the annotation of the extended textual coreference (that has – unlike the originally annotated textual coreference – a type) and the bridging anaphora (that has not been annotated before and also has a type). Technically, various kinds of non-dependency relations between nodes in PDT 2.0 use dedicated referring attributes that contain unique identifiers of the nodes they refer to.

3.1 Helping the Annotators

We employ two ways of helping the annotators in their tedious task. First, we pre-annotate the data with highly probable coreference relations. The annotators check these links and can remove them if they are wrong. This approach has proved to be faster than letting the annotators annotate the data from scratch. Second, we have implemented several supporting features into the annotation tool (the TrEd extension) that help during the annotation process.

3.1.1 Pre-Annotation

We use a list of pairs of words that with a high probability form a coreferential pair in texts. Most of the pairs in the list consist of a noun and a derived adjective, which are different in Czech, e.g. Praha – pražský (in English: Prague – Prague, like in the sentence: *He arrived in Prague and found the Prague atmosphere quite casual*). The rest of the list is formed by pairs consisting of an abbreviation and its one-word expansion, e.g. ČR – Česko (similarly in English: USA – States). The whole list consists of more than 6 thousand pairs obtained automatically from the morphological synthesizer for Czech, manually checked and slightly extended.

3.1.2 Annotation

Several features have been implemented in the annotation tool to help with the annotation.

Manual pre-annotation: If the annotator finds a word in the text that appears many times in the document and its occurrences seem to co-refer, he can create a coreferential chain out of these words by a single key-stroke. All nodes that have the same tectogrammatical lemma (t_lemma) become a part of the chain.

Finding the nearest antecedent: The annotation instructions require that the nearest antecedent is always selected for the coreferential link. The tool automatically re-directs a newly created coreferential arrow to the nearest one (in the already existing coreferential chain) if the annotator selects a farther antecedent by mistake. However, the rule of the nearest antecedent can be broken in less clear situations. For example, if there are three coreferential words in the text, A, B and C (ordered from left to right), and the annotator connects A and C (overlooking B), and later realizes that B is also coreferential with A and creates the arrow from B to A, the tool re-connects the C→A arrow to C→B. Thus, the chain C→B→A is correctly created.

Preserving the coreferential chain: If the annotator removes an arrow and a coreferential chain is thus interrupted, the tool asks the annotator whether it should re-connect the chain.

Text highlighting: The annotation of the extended textual coreference and the bridging anaphora is performed on the tectogrammatical layer of PDT. However, the annotators prefer to work on the surface form of the text, using the tectogrammatical trees only as a supporting depiction of the relations. After selecting a word in the sentences (by clicking on it), the tool determines to which node in the tectogrammatical trees the word belongs. Then, the projection back to the surface is performed and all words on the surface that belong to the selected node are highlighted. Only one word of the highlighted words is a lexical counterpart of the tectogrammatical node (which is usually the word the annotator clicked on – only in cases such as if the annotator clicks on a preposition or other auxiliary word, the lexical counterpart of the corresponding tectogrammatical node differs from the word clicked on). Using this information, also all words in the sentences that have the same t_lemma (again, we use only the lexical counterparts) as the selected word, are underlined. Words that are connected with the selected word via a coreferential chain are highlighted in such colors that indicate whether the last connecting relation in the chain was textual or grammatical. Moreover, all words that are connected via a bridging anaphora with any word of this coreferential chain, are highlighted in a specific color.

4 Application and Evaluation

The annotation of the extended textual coreference and the bridging anaphora started in November 2008. Two annotators work on different texts (each document is annotated only by one annotator), except for a small overlap used for measuring the inter-annotator agreement.

As of April 2009, about one fifth of PDT 2.0 data has been annotated. The detailed numbers are summed in Table 1:

number of annotated documents	611
total number of sentences	9,425
total number of words	157,817
total number of tectogrammatical nodes (excl. the technical root)	127,954
number of newly annotated co-referring nodes (bridging relations and textual coreference)	16,874
number of co-referring nodes including the textual coreference originally annotated in PDT 2.0	20,532
% of co-referring nodes	16 %

Table 1. Annotation statistics

Figure 1 presents the proportion of types of coreferential and bridging relations in the currently annotated part of PDT[1]. TK_0 is used for textual coreference of specific NPs, TK_NR for textual coreference of non-specific NPs, other abbreviations are believed to be self-explaining.

[1] Including the originally annotated textual coreference in PDT 2.0.

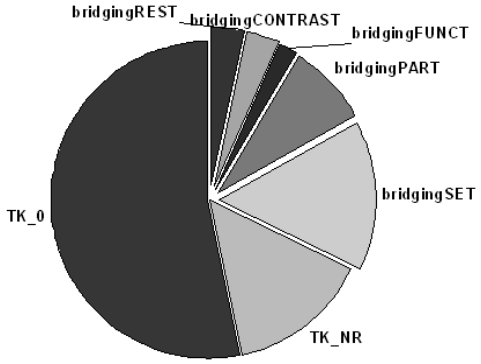

Figure 1. Types of relations

Inter-annotator agreement: For the purposes of checking and improving the annotation guidelines, we require a more strict inter-annotator agreement than agreement on sets (for coreference), often used in other projects (e.g. Passoneau 2004). For both the extended textual coreference and the bridging anaphora, we use F_1-measure for the agreement on the antecedent, and Cohen's κ (Cohen 1960) for the agreement on the type of the link. In Table 2, the results of the three performed measurements of the inter-annotator agreement are presented:

	arrows TC (F_1)	arrows TC + types (F_1)	TC types only (κ)	arrows bridging (F_1)	arrows bridging + types (F_1)	bridging types only (κ)
1st measurement (40 sent.)	0.76	0.67	0.54	0.49	0.42	0.79
2nd measurement (40 sent.)	0.64	0.41	0.33	0.52	0.52	1
3rdmeasurement (100 sent.)	0.80	0.68	0.67	0.59	0.57	0.88

Table 2. Evaluation of the inter-annotator agreement

5 Conclusion

We have presented the annotation scheme and principles for the extended textual coreference and the bridging anaphora in PDT 2.0.

Pre-annotation and features of the annotation tool that help the annotators have been described in detail. We have presented basic statistics about the annotation completed so far and results of first measurements of the inter-annotator agreement (which are difficult to compare to other approaches, as we do not use "markables").

Improvement of the inter-annotator agreement is in our focus for the upcoming stage of the project. The experience shows that the agreement is greatly affected by parameters of the text as a whole. Short texts are generally far less demanding for their interpretation than longer ones, texts with many abstract and general notions allow more possibilities of interpretation and so on. Frequent problems causing inter-annotator disagreement are of two types - different understanding of the content and inaccuracy of the coding scheme. The first case is hardly to be solved entirely. The problems of the second type are being worked on: we prepare the detailed classification of the inter-annotator disagreement and regularly specify the annotation guidelines.

Acknowledgment

We gratefully acknowledge the support of the Czech Ministry of Education (grant MSM-0021620838), the Czech Grant Agency (grant 405/09/0729), the European Union (project Companions – FP6-IST-5-034434), and the Grant Agency of the Academy of Sciences of the Czech Republic (project 1ET101120503).

References

Clark, H. 1977. Bridging. In Johnson-Laird and Wason, editors, *Thinking: Readings in Cognitive Science.* Cambridge. 411-420.

Cohen, J. 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37-46.

Hajič, J. et al. 2006. Prague Dependency Treebank 2.0.CD-ROM, Linguistic Data Consortium, LDC Catalog No.: LDC2006T01, Philadelphia.

Hirschman, L. 1997. MUC-7 coreference task definition. Version 3.0.

Krasavina, O. and Ch. Chiarcos. 2007. PoCoS – Potsdam Coreference Scheme. Proc. of ACL 2007, Prague, Czech Republic

Kučová L. and E. Hajičová. 2004. Coreferential Relations in the Prague Dependency Treebank. In 5th Discourse Anaphora and Anaphor Resolution Colloquium. Ediçôes Colibri.

Pajas, P. and J. Štěpánek 2008. Recent advances in a feature-rich framework for treebank annotation. In The 22nd Interntional Conference on Computational Linguistics – Proceedings of the Conference. Manchester, pp. 673-680.

Passonneau, R. 2004. Computing Reliability for Coreference. In Proceedings of LREC, vol. 4, Lisbon, pp. 1503-1506.

Poesio, M. 2004 The MATE/GNOME Proposals for Anaphoric Annotation, Revisited. Proc. of SIGDIAL.