Rule-based Approach to Pronominal Anaphora Resolution Applied on the Prague Dependency Treebank 2.0 Data

Nguy Giang Linh, Zdeněk Žabokrtský

Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague
Malostranské nám. 25, 118 00 Prague 1, Czech Republic
{linh, zabokrtsky}@ufal.mff.cuni.cz

Abstract

In this paper, we present a rule-based approach to resolution of anaphora links, as annotated in the Prague Dependency Treebank 2.0. The created system consists of handwritten rules developed and tested using the Treebank data, which contain more than 45,000 coreference links in almost 50,000 manually annotated Czech sentences. The F-measure of our system is 74.2%.

1. Introduction

Prague Dependency Treebank 2.0[1] (PDT 2.0, (Jan Hajič, et al., 2006)) is a large collection of linguistically annotated data and documentation, based on the theoretical framework of Functional Generative Description (Sgall et al., 1986). In PDT 2.0, Czech newspaper texts are annotated using a rich layered annotation scenario; the most abstract layer (called tectogrammatical layer) includes also annotation of coreferential links. Automatic detection of such links is the main aim of the presented work. In this paper we focus on resolving only a specific subset of the annotated coreference links, namely those which correspond to coreference of personal pronouns, coreference of possessive pronouns, and coreference of surface-deleted ('zero') pronouns. To our knowledge, the presented system outperforms previously published approaches evaluated on the same data (e.g. (Kučová and Žabokrtský, 2005)).

2. Layers of Annotation in PDT 2.0

The Prague Dependency Treebank 2.0 adds three layers of annotation to Czech texts selected from the Czech National Corpus (see Figure 1):

- morphological layer (m-layer), on which a lemma and a positional morphological tag are added to each token (word form or punctuation mark) in each sentence of the source texts,
- analytical layer (a-layer), where each sentence is represented as a surface-syntactic dependency tree, in which each node corresponds to one m-layer token; edges correspond either to dependency relations between tokens (such as subject, object, attribute), or to other relations of non-dependency nature (such as coordination),
- tectogrammatical layer (t-layer, see (Mikulová et al., 2005) for details), where each sentence is represented as a complex deep-syntactic dependency tree (tectogrammatical tree, t-tree), in which only autosemantic words have nodes of their own (functional words such as prepositions or auxiliary verbs are represented by other means); on the other hand, tectogrammatical trees contain also nodes having no counterparts in the surface shape of the sentences, for instance nodes corresponding to 'pro-dropped' subjects.

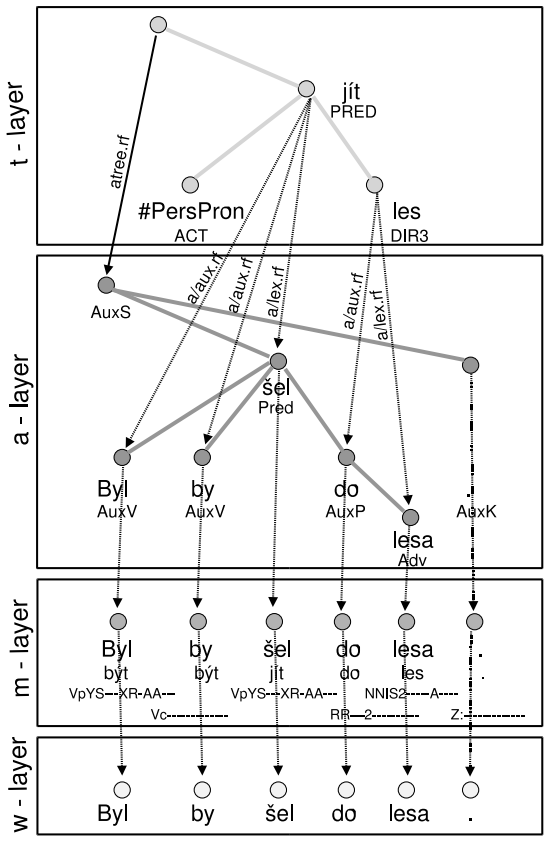


Figure 1: PDT 2.0 annotation layers (and the layer interlinking) illustrated (in a simplified fashion) on the sentence *Byl by šel do lesa.* ([He] would have gone into forest.)

Coreference annotation is considered as one of the components of the t-layer annotation scheme. In FGD, the distinction between grammatical and textual coreference is drawn (Panevová, 1991). One of the differences is that (individ-

[1] http://ufal.mff.cuni.cz/pdt2.0/

ual subtypes of) grammatical coreference can occur only if certain local configuration requirements are fulfilled in the dependency tree (such as: if there is a relative pronoun node in a relative clause and the verbal head of the clause is governed by a nominal node, then the pronoun node and nominal node are coreferential), whereas textual coreference between two nodes (e.g. a personal pronoun node and its antecedent) does not imply any syntactic relation between the nodes in question or any other constraint on the shape of the dependency tree. Thus textual coreference easily crosses sentence boundaries.

3. Coreference Data in PDT 2.0

PDT 2.0 contains 3,168 newspaper texts annotated at the tectogrammatical level. Altogether, they consist of 49,431 sentences. Coreference has been annotated manually in all this data. There are 45,631 coreference links (counting both textual and grammatical ones).

In the PDT 2.0 following grammatical and textual coreference are annotated (see their percent occurrence in Table 1):

- grammatical coreference - verbs of control, reflexive pronouns, verbal complements, reciprocity and relative pronouns
- textual coreference - personal and possessive pronouns, demonstrative pronouns, pleonastic *it* (noun phrase anaphora and bridging (indirect) anafora are in process of manual annotation)

Anaphors	Percentage
Personal Pronouns	35%
Relative Pronouns	20%
Verbs of Control	19%
Reflexive Pronouns	9%
Demonstrative Pronouns	8%
Possessive Pronouns	4%
Verbal Complements	3%
Reciprocity Pronouns	2%

Table 1: The percent occurrence of anaphors in the PDT 2.0

Figure 2 (a) shows a sample t-tree sample in which coreference links are depicted. They form a coreferential chain corresponding to surface tokens *Novotná – své – jí* [Novotná – her (reflexive pronoun) – her (possessive pronoun)].

As the tectogrammatical structures are highly complex, there can be more than twenty attribute-value pairs associated with the individual nodes. The tree in the figure is displayed in a simplified fashion: the nodes are labeled only with tectogrammatical lemmas, functors, and semantic parts of speech. We will give only a brief explanation of these attributes in the following paragraphs.

The first attribute is tectogrammatical lemma, which stands either for the canonical word form of the word present in the surface sentence form or for the artificial value of a newly created node in the tectogrammatical layer. The (artificial) tectogrammatical lemma #PersPron stands for personal (and possessive) pronouns, be they expressed on the surface (i.e., present in the original sentence) or restored during the annotation of the tectogrammatical tree structure (zero pronouns).

The second attribute is functor, which describes the type of the edge leading from the node to its governor; the edge may represent dependency relation (mostly of semantic nature), or other technical phenomena. Following FGD, the dependency functors are divided into actants (ACT - actor, PAT - patient, ADDR - addressee, etc.) and free modifiers (LOC - location, BEN - benefactor, RHEM - rhematizer, TWHEN - temporal modifier, APP - appurtenance, etc.).

The third attribute displayed below the nodes is semantic part-of-speech, representing categories of the tectogrammatical layer corresponding to basic onomaziologic categories (substance, feature, factor, event) and are not identical with the 'traditional' parts of speech. The main semantic parts of speech distinguished in PDT 2.0 are: semantic nouns, semantic adjectives, semantic adverbs and semantic verbs. These basic sets are further subdivided. In the following list we present those subtypes of semantic nouns which most frequently appear as antecedent nodes (clearly, the value of `sempos` is helpful for selecting antecedent 'candidates'):

n.denot – denotative semantic noun,

n.denot.neg – denotative semantic noun with separately represented negation feature,

n.pron.def.demon – demonstrative definite pronominal semantic noun,

n.pron.def.pers – pronominal definite personal semantic noun,

n.pron.indef – indefinite pronominal semantic noun,

n.quant.def – quantification definite semantic noun.

Coreference links are displayed as arrows in the figure, pointing from an anaphor to its antecedent. In the tree editor `tred`[2] used for PDT 2.0 annotation, different arrow colors are used for distinguishing textual and grammatical coreference.

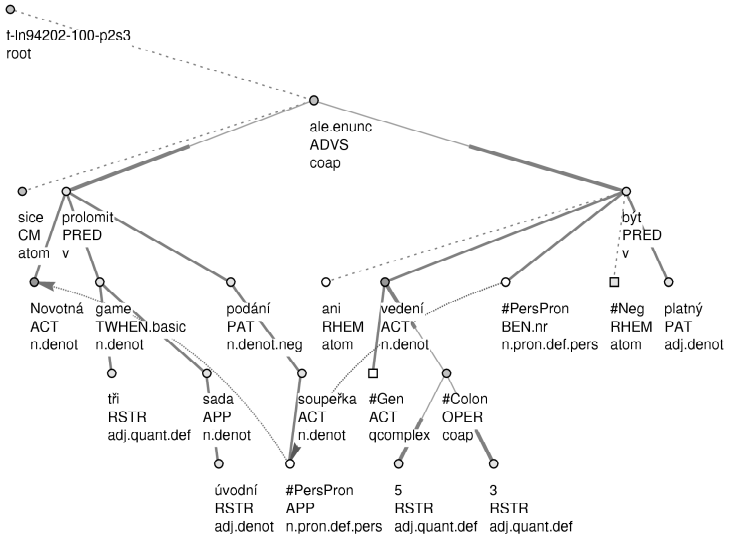
In the PDT 2.0 the data representation for coreferential chains differs from these described in (Kučová et al., 2003) and (Kučová and Hajičová, 2004). Three completely new attributes are established for each anaphor:

coref_gram.rf – identifier or a list of identifiers of the antecedent(s) related via grammatical coreference

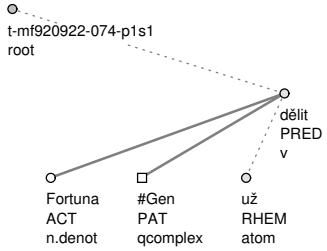
coref_text.rf – identifier or a list of identifiers of the antecedent(s) related via textual coreference

coref_special – values `segm` (segment) and `exoph` (exophora) standing for special types of textual coreference.

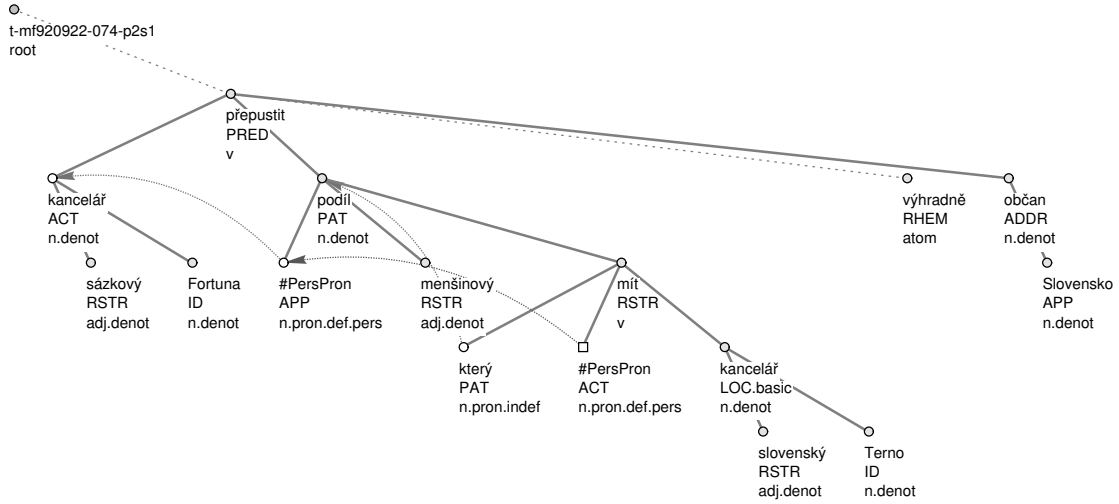
[2] http://ufal.mff.cuni.cz/~pajas/tred/



(a) Simplified t-tree representing the sentence *Novotná sice prolomila ve třetím gamu úvodní sady podání své soupeřky, ale ani vedení 5:3 jí nebylo platné.* (Lit.: Novotná indeed broke through in the third game of the initial set serve of her opponent, but not the lead 5:3 her was efficient.)



(b) Simplified t-tree representing the sentence *Fortuna už "dělila"* (Lit.: Fortuna already divided)



(c) Simplified t-tree representing the sentence *Sázková kancelář Fortuna přepustila svůj menšinový podíl, který Ø měla ve slovenské kanceláři Terno, výhradně občanům Slovenska.* (Lit.: Betting agency Fortuna rendered its minority share, which Ø had in the Slovak agency Terno, entirely citizen of Slovakia.)

Figure 2: Sample t-trees

4. Rule-based Approach

Our pronoun coreference resolution is mainly inspired by the Lappin and Leass's algorithm (Lappin and Leass, 1994) and the Mitkov's robust, knowledge-poor approach (Mitkov, 2002). However, certain adaptations are necessary, e.g. because of the fact that none of the mentioned algorithms were developed for dependency trees.

In our system, the following procedure is used for each anaphor candidate (t-tree node having tectogrammatical lemma equal to #PersPron):[3]

All semantic nouns from the previous and current sentence are checked for gender and number agreement with the anaphor. Nouns not preceding the anaphor but occurring in the same sentence are included in the procedure too, because the tectogrammatical word order is different from the surface word order. Newly created nodes representing zero pronouns sometimes precede its antecedent, but they are not cataphors.

Then the remaining candidates are assigned a positive or negative score. Positive scores can be related to preferences; negative scores to constraints.

The scores are:[4]

- Subject: Score +1 is added to the subject of a clause.
- Subject in main clause: Score +1 is added to the subject of a main clause.
- Frequent noun: Score +1 is added to nouns occurring in the current text more than once.
- Frequent functor: Score +1 is added to nouns dependent on a verb and having one of the following most frequent antecedent functors: ACT, ADDR, PAT, APP (see table 2). The other nouns are assigned the score of -1.
- Collocation: Score +2 is added to nouns having identical collocation pattern with the anaphor. The set of collocation patterns are created from the current text using verbs and its denotative semantic nouns, which occur in the text as its actant.
- Distance: Score +2 is added to those nouns, which are found in the same sentence as the anaphor and precede it; score +1 is added to the nouns occurring in the previous sentence (see table 3).

The noun with the highest accumulated score is proposed as antecedent; in the rare event of a tie, priority is given to the most recent candidate preceding the anaphor. If all remaining candidates occur after the anaphor, the closest one is chosen as antecedent.

The algorithm is illustrated in the following example:
Fortuna$_i$ už "dělila" (Figure 2 (b))

[3] As it was already mentioned, we limit ourselves only to this subset of coreference types, because in (Kučová et al., 2003) and (Linh, 2006) the resolution for grammatical coreference is shown as quite clear one.

[4] If information about article titles and paragraph dividing was included in the Treebank, we could use it as another scores for antecedent candidates.

Functors	Percentage
ACT	59%
PAT	22%
APP	7%
ADDR	5%
Other	7%

Table 2: The percent occurrence of antecedent functors in the PDT 2.0

Antecedent Location	Percnt.
Previous Sentence	37%
Current Sentence and Preceding the Anaphor	57%
Current Sentence and Following the Anaphor	5%
Other	1%

Table 3: The percent occurrence of antecedent - anaphor distances in the PDT 2.0

Sázková kancelář$_i$ Fortuna$_j$ přepustila svůj menšinový podíl, který Ø měla ve slovenské kanceláři$_j$ Terno, výhradně občanům Slovenska. (Figure 2 (c))

Table 4 below shows scores added to the set of candidates, which match with the anaphor Ø in gender and number.

Candidates	Sb	MC	N	F	Cl	Ds	Sum
Fortuna$_i$	+1	+1	+1	+1		+1	5
kancelář$_i$	**+1**	**+1**	**+1**	**+1**		**+2**	**6**
Fortuna$_j$			+1	-1		+2	2
kancelář$_j$			+1	-1			0

Table 4: Antecedent selection for the anaphor Ø (Sb: subject, MC: subject in main clause, N: frequent noun, F: frequent functor, Cl: collocation match, Ds: distance, Sum: sum)

5. Evaluation

The data in the PDT 2.0 are divided into three groups: training set (80%), development test set (10%), and evaluation test set (10%). The training and development test set can be freely exploited and tested by users. But the evaluation test data should be never looked into, they are intended for evaluation and reporting purposes only.

For the evaluation purposes, we have used the standard metrics:

$$\text{Precision} = \frac{\text{number of correctly predicted coreference links}}{\text{number of all predicted links}}$$

$$\text{Recall} = \frac{\text{number of correctly predicted coreference links}}{\text{number of links to predict}}$$

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Our approach was developed and tested on training and development test data. Finally it was tested on evaluation test

data for the final scoring and gave the following results: Precision 73.9%, Recall 74.5%, F-measure 74.2%.

6. Final Remarks

To our knowledge, the presented system outperforms two previously published systems evaluated on the same data. In (Kučová and Žabokrtský, 2005) a set of filters for personal pronominal anaphora resolution was proposed. The list of candidates was built from the preceding and the same sentence as the personal pronoun. After applying each filter the improbably candidates were cut off. If there was more than one candidate left at the end, the nearest one to the anaphor was chosen as its antecedent. The final success rate was 60.4%. In (Němčík, 2006), various algorithms for anaphora resolution have been implemented, but the presented results are also significantly than the results presented in this paper. Some experiments with using C4.5 top-bottom decision trees for Czech anaphora resolution are described in (Linh, 2006), but surprisingly, this machine learning approach was not more successful than our rule-based approach.

In the future we would like to continue on improving Czech anaphora resolution with various statistical methods. Our success on it will be helpful for other projects on natural language processing in the PDT 2.0. One of them is eg. building the fourth layer – the logical layer. We are also going to focus on resolution of bridging anaphora and noun phrase anaphora, the pilot data sets for which have been already annotated for Czech.

7. Acknowledgements

The research reported on in this paper has been carried out under the projects GAAV ČR 1ET101120503, MSM0021620838, and MŠMT ČR LC536.

8. References

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

Lucie Kučová and Eva Hajičová. 2004. Coreferential Relations in the Prague Dependency Treebank. In *5th Discourse Anaphora and Anaphor Resolution Colloquium*. Edições Colibri.

Lucie Kučová and Zdeněk Žabokrtský. 2005. Anaphora in Czech: Large Data and Experiments with Automatic Anaphora. In *LNCS/Lecture Notes in Artificial Intelligence/Proceedings of Text, Speech and Dialogue*. Springer Verlag Heidelberg.

Lucie Kučová, Veronika Kolářová, Zdeněk Žabokrtský, Petr Pajas, and Oliver Čulo. 2003. Anotování koreference v Pražském závislostním korpusu. Technical Report TR-2003-19, ÚFAL MFF UK, Prague, Prague.

Shalom Lappin and Herbert J. Leass. 1994. "an algorithm for pronominal anaphora resolution". *Computational Linguistics*, 20(4):535–561.

Nguy Giang Linh. 2006. Proposal of a Set of Rules for Anaphora Resolution in Czech. Master's thesis, Faculty of Mathematics and Physics, Charles University.

Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolářová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský, and Lucie Kučová. 2005. Anotace na tektogramatické rovině Pražského závislostního korpusu. Anotátorská příručka (t-layer annotation guidelines). Technical Report TR-2005-28, ÚFAL MFF UK, Prague, Prague.

Ruslan Mitkov. 2002. *Anaphora Resolution*. Longman, London.

Václav Němčík. 2006. Anaphora Resolution. Master's thesis, Faculty of Informatics, Masaryk University.

Jarmila Panevová. 1991. Koreference gramatická nebo textová? In *Etudes de linguistique romane et slave*. Krakow.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. D. Reidel Publishing Company, Dordrecht.