Comparison of Classification and Ranking Approaches to Pronominal Anaphora Resolution in Czech*

Nguỵ Giang Linh, Václav Novák, Zdeněk Žabokrtský
Charles University in Prague
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, CZ-11800
{linh,novak,zabokrtsky}.ufal.mff.cuni.cz

Abstract

In this paper we compare two Machine Learning approaches to the task of pronominal anaphora resolution: a conventional classification system based on C5.0 decision trees, and a novel perceptron-based ranker. We use coreference links annotated in the Prague Dependency Treebank 2.0 for training and evaluation purposes. The perceptron system achieves f-score 79.43% on recognizing coreference of personal and possessive pronouns, which clearly outperforms the classifier and which is the best result reported on this data set so far.

1 Introduction

Anaphora Resolution (AR) is a well established task in Natural Language Processing (Mitkov, 2002). Classification techniques (e.g., single candidate model aimed at answering: "Is there a coreference link between the anaphor and this antecedent candidate, or not?") are very often used for the task, e.g. in Mccarthy and Lehnert (1995) and Soon et al. (2001). However, as argued already in Yang et al. (2003), better results are achieved when the candidates can compete in a pairwise fashion. It can be explained by the fact that in this approach (called twin-candidate model), more information is available for the decision making. If we proceed further along this direction, we come to the ranking approach described in Denis and Baldridge (2007), in which the entire candidate set is considered at once and which leads to further significant shift in performance, more recently documented in Denis and Baldridge (2008).

In this paper we deal with supervised approaches to pronominal anaphora in Czech.[1] For training and evaluation purposes, we use coreferences links annotated in the Prague Dependency Treebank, (Jan Hajič, et al., 2006). We limit ourselves only to textual coreference (see Section 2) and to personal and possessive pronouns. We make use of a rich set of features available thanks to the complex annotation scenario of the treebank.

We experiment with two of the above mentioned techniques for AR: a classifier and a ranker. The former is based on a top-down induction of decision trees (Quinlan, 1993). The latter uses a simple scoring function whose optimal weight vector is estimated using perceptron learning inspired by Collins (2002). We try to provide both implementations with as similar input information as possible in order to be able to compare their performance for the given task.

Performance of the presented systems can be compared with several already published works, namely with a rule-based system described in Kučová and Žabokrtský (2005), some of the "classical" algorithms implemented in Němčík (2006), a system based on decision trees (Nguỵ, 2006), and a rule-based system evaluated in Nguỵ and Žabokrtský (2007). To illustrate the real complexity of the task, we also provide performance evaluation of a baseline solution.

The work on this project was supported by the grants MSM 0021620838, GAAV ČR 1ET101120503 and 1ET201120505, MŠMT ČR LC536, and GAUK 4383/2009

[1] Currently one can see a growing interest in unsupervised techniques, e.g. Charniak and Elsner (2009) and Ng (2008). However, we make only a very tiny step in this direction: we use a probabilistic feature based on collocation counts in large unannotated data (namely in the Czech National Corpus).

Proceedings of SIGDIAL 2009: the 10th Annual Meeting of the Special Interest Group in Discourse and Dialogue, pages 276–285,
Queen Mary University of London, September 2009.

The most important result claimed in this paper is that, to the best of our knowledge, the presented ranker system outperforms all the previously published systems evaluated on the PDT data. Moreover, the performance of our ranker (f-score 79.43%) for Czech data is not far from the performance of the state-of-the-art system for English described in Denis and Baldridge (2008) (f-score for 3rd person pronouns 82.2 %).[2]

A side product of this work lies in bringing empirical evidence – for a different language and different data set – for the claim of Denis and Baldridge (2007) that the ranking approach is more appropriate for the task of AR than the classification approach.

The paper is structured as follows. The data with manually annotated links we use are described in Section 2. Section 3 outlines preprocessing the data for training and evaluation purposes. The classifier-based and ranker-based systems are described in Section 4 and Section 5 respectively. Section 6 summarizes the achieved results by evaluating both approaches using the test data. Conclusions and final remarks follow in Section 7.

2 Coreference links in the Prague Dependency Treebank 2.0

The Prague Dependency Treebank 2.0[3] (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 (FGD; introduced by Sgall (1967) and later elaborated, e.g. in by Sgall et al. (1986)). The PDT 2.0 data are Czech newspaper texts selected from the Czech National Corpus[4] (CNC).

The PDT 2.0 has a three-level structure. On the lowest *morphological* level, a lemma and a positional morphological tag are added to each token. The middle *analytical* level represents each sentence as a surface-syntactic dependency tree. On the highest *tectogrammatical* level, each sentence is represented as a complex deep-syntactic dependency tree, see Mikulová and others (2005) for details. This level includes also annotation of coreferential links.

The PDT 2.0 contains 3,168 newspaper texts (49,431 sentences) annotated on the tectogrammatical level. Coreference has been annotated manually in all this data. Following the FGD, there are two types of coreference distinguished: *grammatical* coreference and *textual* coreference (Panevová, 1991). The main difference between the two coreference types is that the antecedent in grammatical coreference can be identified using grammatical rules and sentence syntactic structure, whereas the antecedent in textual coreference can not.

The further division of grammatical and textual coreference is based on types of anaphors:

Grammatical anaphors: relative pronouns, reflexive pronouns, reciprocity pronouns, restored (surface-unexpressed) "subjects" of infinitive verbs below verbs of control,

Textual anaphors: personal and possessive pronouns, demonstrative pronouns.

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, while the evaluation test data should serve only for the very final evaluation of developed tools.

Table 1 shows the distribution of each anaphor type. The total number of coreference links in the PDT 2.0 data is 45,174.[5] Personal pronouns including those zero ones and possessive pronouns form 37.4% of all anaphors in the entire corpus (16,888 links).

An example tectogrammatical tree with depicted coreference links (arrows) is presented in Figure 1. For the sake of simplicity, only three node attributes are displayed below the nodes: tectogrammatical lemma, functor, and semantic part of speech (tectogrammatical nodes themselves are complex data structures and around twenty attributes might be stored with them).

Tectogrammatical lemma is a canonical word form or an artificial value of a newly created node

[2] However, it should be noted that exact comparison is not possible here, since the tasks are slightly different for the two languages, especially because of typological differences between Czech and English (frequent pro-drop in Czech) and different information available in the underlying data resource on the other hand (manually annotated morphological and syntactical information available for Czech).

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

[4] http://ucnk.ff.cuni.cz/

[5] In terms of the number of coreference links, PDT 2.0 is one of the largest existing manually annotated resources. Another comparably large resource is BBN Pronoun Coreference and Entity Type Corpus (Weischedel and Brunstein, 2005), which contains a stand-off annotation of coreference links in the Penn Treebank texts.

Type/Count	train	dtest	etest
Personal pron.	12,913	1,945	2,030
Relative pron.	6,957	948	1,034
Under-control pron.	6,598	874	907
Reflexive pron.	3,381	452	571
Demonstrative pron.	2,582	332	344
Reciprocity pron.	882	110	122
Other	320	35	42
Total	34,983	4,909	5,282

Table 1: Distribution of the different anaphor types in the PDT 2.0.

on the tectogrammatical level. E.g. 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).

Functor captures the deep-syntactic dependency relation between a node and its governor in the tectogrammatical tree. According to FGD, 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.).

Semantic parts of speech correspond to basic onomasiological categories (substance, feature, factor, event). The main semantic POS distinguished in PDT 2.0 are: semantic nouns, semantic adjectives, semantic adverbs and semantic verbs (for example, personal and possessive pronouns belong to semantic nouns).

3 Training data preparation

The training phase of both presented AR systems can be outlined as follows:

1. detect nodes which are anaphors (Section 3.1),
2. for each anaphor a_i, collect the set of antecedent candidates $\mathrm{Cand}(a_i)$ (Section 3.2),
3. for each anaphor a_i, divide the set of candidates into positive instances (true antecedents) and negative instances (Section 3.3),
4. for each pair of an anaphor a_i and an antecedent candidate $c_j \in \mathrm{Cand}(a_i)$, compute the feature vector $\Phi(c, a_i)$ (Section 3.4),
5. given the anaphors, their sets of antecedent candidates (with related feature vectors), and the division into positive and negative candidates, train the system for identifying the true antecedents among the candidates.

Steps 1-4 can be seen as training data preprocessing, and are very similar for both systems. System-specific details are described in Section 4 and Section 5 respectively.

3.1 Anaphor selection

In the presented work, only third person personal (and possessive) pronouns are considered,[6] be they expressed on the surface or reconstructed. We treat as anaphors all tectogrammatical nodes with lemma #PersPron and third person stored in the gram/person grammateme. More than 98 % of such nodes have their antecedents (in the sense of textual coreference) marked in the training data. Therefore we decided to rely only on this highly precise rule when detecting anaphors.[7]

In our example tree, the node #PersPron representing his on the surface and the node #PersPron representing the zero personal pronoun he will be recognized as anaphors.

3.2 Candidate selection

In both systems, the predicted antecedent of a given anaphor a_i is selected from an easy-to-compute set of antecedent candidates denoted as $\mathrm{Cand}(a_i)$. We limit the set of candidates to semantic nouns which are located either in the same sentence before the anaphor, or in the preceding sentence. Table 2 shows that if we disregard cataphoric and longer anaphoric links, we loose a chance for correct answer with only 6 % of anaphors.

[6] The reason is that antecedents of most other types of anaphors annotated in PDT 2.0 can be detected – given the tree topology and basic node attributes – with precision higher than 90 %, as it was shown already in Kučová and Žabokrtský (2005). For instance, antecedents of reflexive pronouns are tree-nearest clause subjects in most cases, while antecedents of relative pronouns are typically parents of the relative clause heads.

[7] It is not surprising that no discourse status model (as used e.g. in Denis and Baldridge (2008)) is practically needed here, since we limit ourselves to personal pronouns, which are almost always "discourse-old".

Antecedent location	Percnt.
Previous sentence	37 %
Same sentence, preceding the anaphor	57 %
Same sentence, following the anaphor	5 %
Other	1 %

Table 2: Location of antecedents with respect to anaphors in the training section of PDT 2.0.

3.3 Generating positive and negative instances

If the true antecedent of a_i is not present in $\mathrm{Cand(a_i)}$, no training instance is generated. If it is present, the sets of negative and positive instances are generated based on the anaphor. This preprocessing step differs for the two systems, because the classifier can be easily provided with more than one positive instance per anaphor, whereas the ranker can not.

In the classification-based system, all candidates belonging to the coreferential chain are marked as positive instances in the training data. The remaining candidates are marked as negative instances.

In the ranking-based system, the coreferential chain is followed from the anaphor to the nearest antecedent which itself is not an anaphor in grammatical coreference.[8] The first such node is put on the top of the training rank list, as it should be predicted as the winner (E.g., the nearest antecedent of the zero personal pronoun he in the example tree is the relative pronoun who, however, it is a grammatical anaphor, so its antecedent Brien is chosen as the winner instead). All remaining (negative) candidates are added to the list, without any special ordering.

3.4 Feature extraction

Our model makes use of a wide range of features that are obtained not only from all three levels of the PDT 2.0 but also from the Czech National Corpus and the EuroWordNet. Each training or testing instance is represented by a feature vector. The features describe the anaphor, its antecedent candidate and their relationship, as well as their contexts. All features are listed in Table 4 in the Appendix.

When designing the feature set on personal pronouns, we take into account the fact that Czech personal pronouns stand for persons, animals and things, therefore they agree with their antecedents in many attributes and functions. Further we use the knowledge from the Lappin and Leass's algorithm (Lappin and Leass, 1994), the Mitkov's robust, knowledge-poor approach (Mitkov, 2002), and the theory of topic-focus articulation (Kučová et al., 2005). We want to take utmost advantage of information from the antecedent's and anaphor's node on all three levels as well.

Distance: Numeric features capturing the distance between the anaphor and the candidate, measured by the number of sentences, clauses, tree nodes and candidates between them.

Morphological agreement: Categorial features created from the values of tectogrammatical gender and number[9] and from selected morphological categories from the positional tag[10] of the anaphor and of the candidate. In addition, there are features indicating the strict agreement between these pairs and features formed by concatenating the pair of values of the given attribute in the two nodes (e.g., masc_neut).

Agreement in dependency functions: Categorial features created from the values of tectogrammatical functor and analytical functor (with surface-syntactic values such as Sb, Pred, Obj) of the anaphor and of the candidate, their agreement and joint feature. There are two more features indicating whether the candidate/anaphor is an actant and whether the candidate/anaphor is a subject on the tectogrammatical level.[11]

Context: Categorial features describing the context of the anaphor and of the candidate:

- parent – tectogrammatical functor and the semantic POS of the effective parent[12] of the

[8] Grammatical anaphors are skipped because they usually do not provide sufficient information (e.g., reflexive pronouns provide almost no cues at all). The classification approach does not require such adaptation – it is more robust against such lack of information as it treats the whole chain as positive instances.

[9] Sometimes gender and number are unknown, but we can identify the gender and number of e.g. relative or reflexive pronouns on the tectogrammatical level thanks to their antecedent.

[10] A positional tag from the morphological level is a string of 15 characters. Every positions encodes one morphological category using one character.

[11] A subject on the tectogrammatical level can be a node with the analytical functor Sb or with the tectogrammatical functor Actor in a clause without a subject.

[12] The "true governor" in terms of dependency relations.

anaphor and the candidate, their agreement and joint feature; a feature indicating the agreement of both parents' tectogrammatical lemma and their joint feature; a joint feature of the pair of the tectogrammatical lemma of the candidate and the effective parent's lemma of the anaphor; and a feature indicating whether the candidate and the anaphor are siblings.[13]

- coordination – a feature that indicates whether the candidate is a member of a coordination and a feature indicating whether the anaphor is a possessive pronoun and is in the coordination with the candidate

- collocation – a feature indicating whether the candidate has appeared in the same collocation as the anaphor within the text[14] and a feature that indicates the collocation assumed from the Czech National Corpus.[15]

- boundness – features assigned on the basis of contextual boundness (available in the tectogrammatical trees) {contextually bound, contrastively contextually bound, or contextually non-bound}[16] for the anaphor and the candidate; their agreement and joint feature.

- frequency – 1 if the candidate is a denotative semantic noun and occurs more than once within the text; otherwise 0.

Semantics: Semantically oriented feature that indicates whether the candidate is a person name for the present and a set of 63 binary ontological attributes obtained from the EuroWordNet.[17] These attributes determine the positive or negative relation between the candidate's lemma and the semantic concepts.

4 Classifier-based system

Our classification approach uses C5.0, a successor of C4.5 (Quinlan, 1993), which is probably the most widely used program for inducing decision trees. Decision trees are used in many AR systems such as Aone and Bennett (1995), Mccarthy and Lehnert (1995), Soon et al. (2001), and Ng and Cardie (2002).[18]

Our classifier-based system takes as input a set of feature vectors as described in Section 3.4 and their classifications (1 – true antecedent, 0 – non-antecedent) and produces a decision tree that is further used for classifying new pairs of candidate and anaphor.

The classifier antecedent selection algorithm works as follows. For each anaphor a_i, feature vectors $\Phi(c, a_i)$ are computed for all candidates $c \in \mathrm{Cand}(a_i)$ and passed to the trained decision tree. The candidate classified as positive is returned as the predicted antecedent. If there are more candidates classified as positive, the nearest one is chosen.

If no candidate is classified as positive, a system of handwritten fallback rules can be used. The fallback rules are the same rules as those used in the baseline system in Section 6.2.

5 Ranker-based system

In the ranker-based AR system, every training example is a pair (a_i, y_i), where a_i is the anaphoric expression and y_i is the true antecedent. Using the candidate extraction function Cand, we aim to rank the candidates so that the true antecedent would always be the first candidate on the list. The ranking is modeled by a linear model of the features described in Section 3.4. According to the model, the antecedent $\hat{y}_i$ for an anaphoric expression a_i is found as:

$$\hat{y}_i = \underset{c \in \mathrm{Cand}(a_i)}{\mathrm{argmax}} \; \Phi(c, a_i) \cdot \overrightarrow{w}$$

The weights $\overrightarrow{w}$ of the linear model are trained using a modification of the averaged perceptron al-

[13] Both have the same effective parent.

[14] If the anaphor's effective parent is a verb and the candidate is a denotative semantic noun and has appeared as a child of the same verb and has had the same functor as the anaphor.

[15] The probability of the candidate being a subject preceding the verb, which is the effective parent of the anaphor.

[16] Contextual boundness is a property of an expression (be it expressed or absent in the surface structure of the sentence) which determines whether the speaker (author) uses the expression as given (for the recipient), i.e. uniquely determined by the context.

[17] The Top Ontology used in EuroWordNet (EWN) contains the (structured) set of 63 basic semantic concepts like Place, Time, Human, Group, Living, etc. For the majority of English synsets (set of synonyms, the basic unit of EWN), the appropriate subset of these concepts are listed. Using the Inter Lingual Index that links the synsets of different languages, the set of relevant concepts can be found also for Czech lemmas.

[18] Besides C5.0, we plan to use also other classifiers in the future (especially Support Vector Machine, which is often employed in AR experiments, e.g. by Ng (2005) and Yang et al. (2006)) in order to study how the classifier choice influences the AR system performance on our data and feature sets.

gorithm (Collins, 2002). This is averaged perceptron learning with a modified loss function adapted to the ranking scenario. The loss function is tailored to the task of correctly ranking the true antecedent, the ranking of other candidates is irrelevant. The algorithm (without averaging the parameters) is listed as Algorithm 1. Note that the training instances where $y_i \notin \mathrm{Cand}(a_i)$ were excluded from the training.

input : N training examples (a_i, y_i),
number of iterations T
init : $\overrightarrow{w} \leftarrow \overrightarrow{0}$;
for $t \leftarrow 1$ **to** T, $i \leftarrow 1$ **to** N **do**
$\hat{y}_i \leftarrow \mathrm{argmax}_{c \in \mathrm{Cand}(a_i)} \Phi(c, a_i) \cdot \overrightarrow{w}$;
if $\hat{y}_i \neq y_i$ **then**
$\overrightarrow{w} = \overrightarrow{w} + \Phi(y_i, a_i) - \Phi(\hat{y}_i, a_i)$;
end
end
output: weights $\overrightarrow{w}$

Algorithm 1: Modified perceptron algorithm for ranking. Φ is the feature extraction function, a_i is the anaphoric expression, y_i is the true antecedent.

Antecedent selection algorithm using a ranker: For each third person pronoun create a feature vector from the pronoun and the semantic noun preceding the pronoun and is in the same sentence or in the previous sentence. Use the trained ranking features weight model to get out the candidate's total weight. The candidate with the highest features weight is identified as the antecedent.

6 Experiments and evaluation

6.1 Evaluation metrics

For the evaluation we use the standard metrics:[19]

$$\text{Precision} = \frac{\text{number of correctly predicted anaphoric third person pronouns}}{\text{number of all predicted third person pronouns}}$$

$$\text{Recall} = \frac{\text{number of correctly predicted anaphoric third person pronouns}}{\text{number of all anaphoric third person pronouns}}$$

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

We consider an anaphoric third person pronoun to be correctly predicted when we can successfully indicate its antecedent, which can be any antecedent from the same coreferential chain as the anaphor.

Both the AR systems were developed and tested on PDT 2.0 training and development test data. Finally they were tested on evaluation test data for the final scoring, summarized in Section 6.3.

6.2 Baseline system

We have made some baseline rules for the task of AR and tested them on the PDT 2.0 evaluation test data. Their results are reported in Table 3. Baseline rules are following: For each third person pronoun, consider all semantic nouns which precede the pronoun and are not further than the previous sentence, and:

- select the nearest one as its antecedent (BASE 1),
- select the nearest one which is a clause subject (BASE 2),
- select the nearest one which agrees in gender and number (BASE 3),
- select the nearest one which agrees in gender and number; if there is no such noun, choose the nearest clause subject; if no clause subject was found, choose the nearest noun (BASE 3+2+1).

6.3 Experimental results and discussion

Scores for all three systems (baseline, clasifier with and without fallback, ranker) are given in Table 3. Our baseline system based on the combination of three rules (BASE 3+2+1) reports results superior to the ones of the rule-based system described in Kučová and Žabokrtský (2005). Kučová and Žabokrtský proposed a set of filters for personal pronominal anaphora resolution. The list of candidates was built from the preceding and the same sentence as the personal pronoun. After applying each filter, improbable 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 reported final success rate was 60.4 % (counted simply as the number of correctly predicted links divided by the number of pronoun anaphors in the test data section).

An interesting point of the classifier-based system lies in the comparison with the rule-based

[19] Using simple accuracy would not be adequate, as there can be no link (or more than one) leading from an anaphor in the annotated data. In other words, finding whether a pronoun has an antecedent or not is a part of the task. A deeper discussion about coreference resolution metrics can be found in Luo (2005).

Rule	P	R	F
BASE 1	17.82%	18.00%	17.90%
BASE 2	41.69%	42.06%	41.88%
BASE 3	59.00%	59.50%	59.24%
BASE 3+2+1	62.55%	63.03%	62.79%
CLASS	69.9%	70.44%	70.17%
CLASS+3+2+1	76.02%	76.60%	76.30%
RANK	79.13%	79.74%	**79.43%**

Table 3: Precision (P), Recall (R) and F-measure (F) results for the presented AR systems.

system of Nguy and Žabokrtský (2007). Without the rule-based fallback (CLASS), the classifier falls behind the Nguy and Žabokrtský's system (74.2%), while including the fallback (CLASS+3+2+1) it gives better results.

Overall, the ranker-based system (RANK) significantly outperforms all other AR systems for Czech with the f-score of 79.43%. Comparing with the model for third person pronouns of Denis and Baldridge (2008), which reports the f-score of 82.2%, our ranker is not so far behind. It is important to say that our system relies on manually annotated information[20] and we solve the task of anaphora resolution for third person pronouns on the tectogrammatical level of the PDT 2.0. That means these pronouns are not only those expressed on the surface, but also artificially added (reconstructed) into the structure according to the principles of FGD.

7 Conclusions

In this paper we report two systems for AR in Czech: the classifier-based system and the ranker-based system. The latter system reaches f-score 79.43% on the Prague Dependency Treebank test data and significantly outperforms all previously published results. Our results support the hypothesis that ranking approaches are more appropriate for the AR task than classification approaches.

References

Chinatsu Aone and Scott William Bennett. 1995. Evaluating automated and manual acquisition of anaphora resolution strategies. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 122–129, Morristown, NJ, USA. Association for Computational Linguistics.

António Branco, Tony McEnery, Ruslan Mitkov, and Fátima Silva, editors. 2007. *Proceedings of the 6th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2007)*, Lagos (Algarve), Portugal. CLUP-Center for Linguistics of the University of Oporto.

Eugene Charniak and Micha Elsner. 2009. EM works for pronoun anaphora resolution. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 148–156, Athens, Greece, March. Association for Computational Linguistics.

Michael Collins. 2002. Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. In *Proceedings of EMNLP*, volume 10, pages 1–8.

Pascal Denis and Jason Baldridge. 2007. A ranking approach to pronoun resolution. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI2007)*, pages 1588–1593, Hyderabad, India, January 6–12.

Pascal Denis and Jason Baldridge. 2008. Specialized models and ranking for coreference resolution. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP2008)*, pages 660–669, Honolulu, Hawaii, USA, October 25–27.

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

Lucie Kučová and Zdeněk Žabokrtský. 2005. Anaphora in Czech: Large Data and Experiments with Automatic Anaphora. *LNCS/Lecture Notes in Artificial Intelligence/Proceedings of Text, Speech and Dialogue*, 3658:93–98.

Lucie Kučová, Kateřina Veselá, Eva Hajičová, and Jiří Havelka. 2005. Topic-focus articulation and anaphoric relations: A corpus based probe. In Klaus Heusinger and Carla Umbach, editors, *Proceedings of Discourse Domains and Information Structure workshop*, pages 37–46, Edinburgh, Scotland, UK, Aug. 8-12.

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

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 25–32, Morristown, NJ, USA. Association for Computational Linguistics.

[20] In the near future, we plan to re-run the experiments using sentence analyses created by automatic tools (all needed tools are available in the TectoMT software framework (Žabokrtský et al., 2008)) instead of manually created analyses, in order to examine the sensitivity of the AR system to annotation quality.

J Mccarthy and Wendy G. Lehnert. 1995. Using decision trees for coreference resolution. In *In Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 1050–1055.

Marie Mikulová et al. 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.

Vincent Ng and Claire Cardie. 2002. Improving machine learning approaches to coreference resolution. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 104–111, Morristown, NJ, USA. Association for Computational Linguistics.

Vincent Ng. 2005. Supervised ranking for pronoun resolution: Some recent improvements. In Manuela M. Veloso and Subbarao Kambhampati, editors, *AAAI*, pages 1081–1086. AAAI Press / The MIT Press.

Vincent Ng. 2008. Unsupervised models for coreference resolution. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP2008)*, pages 640–649, Honolulu, Hawaii, USA.

Giang Linh Nguy and Zdeněk Žabokrtský. 2007. Rule-based approach to pronominal anaphora resolution applied on the prague dependency treebank 2.0 data. In Branco et al. (Branco et al., 2007), pages 77–81.

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

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

J. Ross Quinlan. 1993. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

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

Petr Sgall. 1967. *Generativní popis jazyka a česká deklinace*. Academia, Prague, Czech Republic.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Comput. Linguist.*, 27(4):521–544.

Zdeněk Žabokrtský, Jan Ptáček, and Petr Pajas. 2008. TectoMT: Highly Modular MT System with Tectogrammatics Used as Transfer Layer. In *Proceedings of the 3rd Workshop on Statistical Machine Translation, ACL*.

Ralph Weischedel and Ada Brunstein. 2005. BBN Pronoun Coreference and Entity Type Corpus. CD-ROM, Linguistic Data Consortium, LDC Catalog No.: LDC2005T33, Philadelphia.

Xiaofeng Yang, Guodong Zhou, Jian Su, and Chew Lim Tan. 2003. Coreference resolution using competition learning approach. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 176–183, Morristown, NJ, USA. Association for Computational Linguistics.

Xiaofeng Yang, Jian Su, and Chew Lim Tan. 2006. Kernel-based pronoun resolution with structured syntactic knowledge. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL2006)*, pages 41–48, Sydney, Australia, July 17–21.

A Appendix

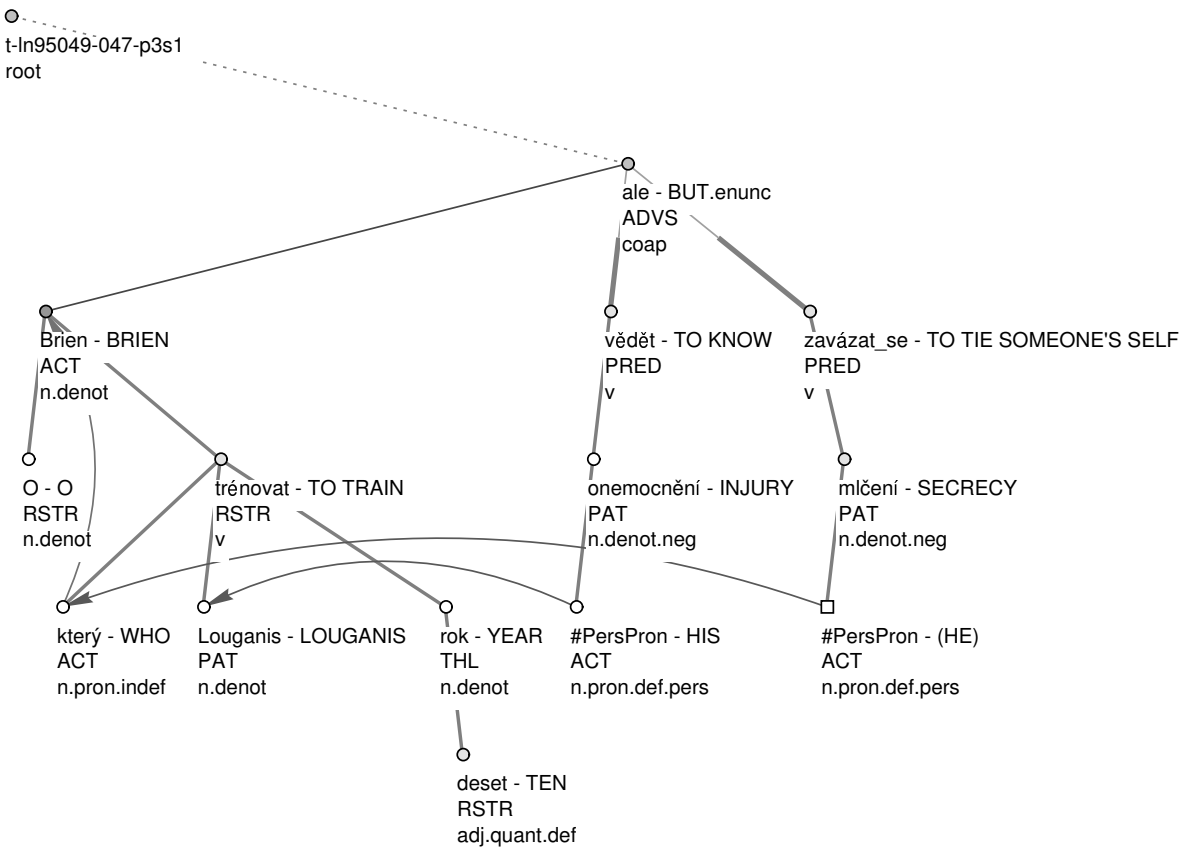

Figure 1: Simplified tectogrammatical tree representing the sentence *O'Brien, který Louganise trénoval deset let, o jeho onemocnění věděl, ale zavázal se mlčením.* (Lit.: O'Brien, who Louganis trained for ten years, about his injury knew, but (he) tied himself to secrecy.) Note two coreferential chains {Brien, who, (he)} and {Louganis, his}.

Distance	
sent_dist	sentence distance between c and a_i
clause_dist	clause distance between c and a_i
node_dist	tree node distance between c and a_i
cand_ord	mention distance between c and a_i
Morphological Agreement	
gender	t-gender of c and a_i, agreement, joint
number	t-number of c and a_i, agreement, joint
apos	m-POS of c and a_i, agreement, joint
asubpos	detailed POS of c and a_i, agreement, joint
agen	m-gender of c and a_i, agreement, joint
anum	m-number of c and a_i, agreement, joint
acase	m-case of c and a_i, agreement, joint
apossgen	m-possessor's gender of c and a_i, agreement, joint
apossnum	m-possessor's number of c and a_i, agreement, joint
apers	m-person of c and a_i, agreement, joint
Functional Agreement	
afun	a-functor of c and a_i, agreement, joint
fun	t-functor of c and a_i, agreement, joint
act	c/a_i is an actant, agreement
subj	c/a_i is a subject, agreement
Context	
par_fun	t-functor of the parent of c and a_i, agreement, joint
par_pos	t-POS of the parent of c and a_i, agreement, joint
par_lemma	agreement between the parent's lemma of c and a_i, joint
clem_aparlem	joint between the lemma of c and the parent's lemma of a_i
c_coord	c is a member of a coordination
app_coord	c and a_i are in coordination & a_i is a possessive pronoun
sibl	c and a_i are siblings
coll	c and a_i have the same collocation
cnk_coll	c and a_i have the same CNC collocation
tfa	contextual boundness of c and a_i, agreement, joint
c_freq	c is a frequent word
Semantics	
cand_pers	c is a person name
cand_ewn	semantic position of c's lemma within the EuroWordNet Top Ontology

Table 4: Features used by the perceptron-based model