Saara: Anaphora Resolution on Free Text in Czech

Vašek Němčík

NLP Centre
Faculty of Informatics, Masaryk University
Brno, Czech Republic
xnemcik@fi.muni.cz

Abstract. Anaphora resolution is one of the key parts of modern NLP systems, and not addressing it usually means a notable performance drop. Despite the abundance of theoretical studies published in the previous decades, real systems for resolving anaphora are rather rare. In this article we present, to our knowledge, the first practical anaphora resolution system applicable to Czech free text. We describe the individual stages of the processing pipeline and sketch the data format used as an interface between individual modules.

Key words: anaphora resolution, Saara, Czech

1 Introduction

In this work, we present a natural language processing (NLP) application setting capable of anaphora resolution (AR) based on plain free text in Czech. This is accomplished by combining several NLP tools, described below, developed at the NLP Centre at the Masaryk University in Brno.

When analyzing texts, anaphoric expressions, especially pronouns, require special handling. On their own, they do not contain any semantic information, and therefore traditional morphological tagging or syntactic analysis as such do not make it possible to arrive at their full meaning. To obtain a complete representation of sentences containing pronouns, these need to be considered in context, namely, interpreted by an AR procedure. Failing to incorporate such a procedure into an NLP system means accepting only a partial text representation, and often a subsequent performance drop.

To our knowledge, there is only a limited number of stand-alone AR systems that work with plain text input, and we are not aware of any such system available for Czech.

In the next section, we mention similar anaphora resolution systems proposed so far. Section 3 describes the processing pipeline of Saara, and further, Section 4 presents performance figures. Finally, we sketch directions of our further research.

Aleš Horák, Pavel Rychlý (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2012, pp. 3–9, 2012. © Tribun EU 2012

2 Related Work

This section sums up existing systems relevant from our perspective, starting with complex AR systems for English, followed by proposals made for Czech.

A number of software tools for performing AR have been presented in the recent years. One of the prominent ones is MARS (Mitkov, Evans, and Orăsan, 2002), a system created at the Univeristy of Wolverhampton. The core of the underlying AR algorithm is a weighting scheme based on the so-called antecedent indicators. There are versions of MARS for various languages, such as English, French, Arabic, Polish, or Bulgarian.

A further notable AR system is BART (Versley et al., 2008), a product of inter-institute cooperation encouraged by the Johns Hopkins Summer Workshop in 2007. BART is a framework allowing straightforward experimenting with various machine learning models. It operates over XML data and allows easy visualisation of results in the MMAX tool (Müller and Strube, 2006).

For Czech, mainly theoretical work has been published. First theoretical models have emerged from the long tradition of research on the Functional Generative Description (FGD) of language. Several algorithms were proposed, for instance by Hajičová (1987), Hajičová, Hoskovec, and Sgall (1995), and Hajičová, Kuboň, and Kuboň (1990), providing only tentative evaluation, due to the lack of sufficiently large annotated data at that time.

The emergence of the Prague Dependency Treebank (PDT) (Hajič et al., 2005), a richly annotated Czech treebank containing annotation of pronominal anaphora, made it possible to experiment with AR systems and to evaluate them. Apart from our work, a notable AR system for Czech is AČA presented by Linh (2006). It comprises rule-based algorithms and also machine learning models for resolving individual pronoun types. Further, a noun phrase coreference resolution system based on maximum entropy and perceptron models was proposed by Novák and Žabokrtský (2011). These systems are respectable results in the field of Czech computational linguistics, however, are fitted to the dependency-based formalism of PDT and their applicability to data in other formalisms may be limited.

The next section gives more details about Saara, a stand-alone AR system for Czech.

3 Saara Pipeline

Saara is is a modular AR system, currently containing re-implementations and variants of selected salience-based algorithms. The architecture of the system was inspired by the principles suggested by Byron and Tetreault (1999), the key points being modularity and encapsulation. They suggest segmenting system modules into three layers: Themselves, they propose three layers:

- the translation layer for creating data structures,
- the AR layer containing functions addressing AR itself,
- the supervisor layer for controlling the previous layers.

```
<s id="sent1" class_id="cls1" type="sentence">
<markable id="m1" class_id="cls2" type="clause">
<markable id="m2" class_id="cls3" type="np" gram="subj">
Filip         Filip          k1gMnSc1
</markable>
políbil       políbit        k5eAaPmAgInSrD,k5eAaPmAgMnSrD
<markable id="m3" class_id="cls4" refconstr="m2" type="np" gram="obj">
Lucii         Lucie          k1gFnSc4
</markable>
</markable>
<g/>
.             .              kIx.
</s>
<s id="sent2" class_id="cls5" type="sentence">
<markable id="m4" class_id="cls6" type="clause">
<markable id="ms1" anaref="m2" class_id="cls3" type="pron_pers_zero" gram="subj">
_             on             k3p3gMnSc1,k3p3gInSc1,k3p3gFnSc1,k3p3gNnSc1
</markable>
Miluje milovat k5eAaImIp3nS
<markable id="m5" anaref="m3" class_id="cls4" refconstr="ms1" type="pron_pers_weak" gram="obj">
ji            on             k3p3gFnSc4xP
</markable>
</markable>
<g/>
.             .              kIx.
</s>
```

Fig. 1. An example of a structured vertical file

We adopt an anologous scheme of layers: the *technical layer* of scripts converting data from various formalisms into a general linear format containing structural tags, so-called markables and their attributes; the *markable layer* abstracting from formalism specifics, operating solely over the already known markables and their attributes, and focusing on the AR process as such; and finally the *supervisor layer* defining the application context, such as individual pre-processing steps and AR algorithm settings.

The interface between all modules is the so-called *structured vertical file*, a plain text format containing one line per token, with extra tags expressing higher-level units, such as sentences, clauses and referential expressions. A slightly abridged example of such a file is given in Figure 1.

The first phase of the processing is converting the input data into the vertical format and performing **morphological analysis**. For plain text input, this is performed by desamb (Šmerk, 2007), a Czech tagger assigning morphological tags to each token and disambiguating these tags based on a statistical model and a set of heuristic rules. For words that can not be disambiguated based on shallow linguistic information, such as pronouns, multiple morphological tags are restored by the Majka morphological analyzer (Šmerk, 2009). At the end of this phase, each token line contains a morphological tag and lemma.

Next, **syntactic analysis** is performed using either the SET (Kovář, Horák, and Jakubíček, 2011) or Synt parser (Jakubíček, Horák, and Kovář, 2009). We use the SET parser by default, as it is slightly more robust. It is based on a small set of rules detecting important patterns in Czech text. In the Saara pipeline,

Table 1. Performance of the system in MUC-6 and traditional measures

	MUC-6		IR-style	
	Precision	Recall	Precision	Recall
Plain Recency (baseline)	22.40	24.85	20.75	20.75
BFP Centering	42.36	44.85	38.98	37.47
Lappin and Leass' RAP	36.54	40.41	35.49	35.39

we use SET to extract phrases, which are subsequently incorporated into the vertical file as tags grouping tokens in question.

As the next step, necessary **syntactic post-processing** is carried out. This comprises assignment of coarse-grained grammatical roles, and based on that, detection of zero subjects, which are afterwards re-constructed as dummy tokens and makrables, including their morphological features.

The core phase of the computation is **anaphora resolution** as such. Modules implementing viriations of diverse AR algorithms, such as the BFP algorithm (Brennan, Friedman, and Pollard, 1987) or RAP (Lappin and Leass, 1994), are available. AR modules supplement markable tags representing individual discourse objects with information about their antecedents and coreference classes.

A web version of this application setting, accepting Czech free text, and with Saara configured to resolve personal pronouns, is freely available online at `http://nlp.fi.muni.cz/projekty/anaphora_resolution/saara/demo/`. We hope the availability of this demo will encourage experimenting with Saara and its extrinsic comparison with other systems.

4 Evaluation

Evaluation of AR systems (and complex NLP systems in general) is a rather complicated issue and gives rise to frequent misconceptions.

A number of sophisticated metrics have been proposed to assess the performance of AR systems with precise numbers, however, these numbers are often substantially biased by a broad range of factors not pointed out in the evaluation report. The figures largely depend on

- whether the evaluation is performed on manually corrected data or data susceptible to processing errors,
- whether errors propagated from the pre-processing (ie. tagging, markable detection) are counted,
- whether all errors are counted equally,
- the precise types of anaphora addressed,
- the size and genre of the text etc.

To evaluate Saara, we use the PDT data projected into structured verticals by the pdt2vert tool (Němčík, 2011), considering only personal pronouns, namely strong and weak personal pronouns, and zero subjects of finite verb groups (the total of 8648 anaphors). We are aware of the fact that the data may contain errors, for instance, due to imperfect detection of clause boundaries, however, we adopt the given structures as correct. Anaphors resolved to a different member of the same coreference chain are considered to be resolved correctly, and all errors have the same weight.

To compare the individual algorithm prototypes, their performance is revealed in Table 1. These results need to be considered as tentative, and are expected to improve with further parameter tuning and the contribution of anaphoric links of further types.

5 Future Work

We have described Saara as a part of a stand-alone NLP system accepting plain text as input, and performing syntax analysis supplemented by an interpretation of personal pronouns.

In our future work, we mainly aim at enhancing the AR methods by accounting for further information relevant to the antecedent choice. The long-term goal is to incorporate as much semantic information as possible with respect to its availability and the reliability of the lower-level analysis. As a first step, a decent approximation can be obtained by considering word co-occurrence statistics.

Further, we plan on to account for further types of anaphoric expressions, for example, certain uses of demonstrative pronouns. Demonstrative pronouns are rather complex to resolve, as they allow reference to abstract entities and discourse segments of arbitrary size.

Acknowledgments

This work has been partly supported by the Ministry of Education of the Czech Republic project No. LM2010013 (Lindat–Clarin – Centre for Language Research Infrastructure in the Czech Republic).

References

1. Brennan, Susan E., Marilyn W. Friedman, and Carl J. Pollard. 1987. A centering approach to pronouns. In *Proceedings of the 25th Annual Meeting of the ACL*, pages 155–162, Standford.
2. Byron, Donna K. and Joel R. Tetreault. 1999. A flexible architecture for reference resolution. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL-99)*.

3. Hajič, Jan et al. 2005. The prague dependency treebank 2.0, `http://ufal.mff.cuni.cz/pdt2.0/`. Developed at the Institute of Formal and Applied Linguistics, Charles University in Prague. Released by Linguistic Data Consortium in 2006.
4. Hajičová, Eva. 1987. Focusing – a meeting point of linguistics and artificial intelligence. In P. Jorrand and V. Sgurev (eds.), *Artificial Intelligence vol II: Methodology, Systems, Applications*. Elsevier Science Publishers, Amsterdam, pp 311–321.
5. Hajičová, Eva, Tomáš Hoskovec, and Petr Sgall. 1995. Discourse modelling based on hierarchy of salience. *The Prague Bulletin of Mathematical Linguistics*, (64):5–24.
6. Hajičová, Eva, Petr Kuboň, and Vladislav Kuboň. 1990. Hierarchy of salience and discourse analysis and production. In *Proceedings of Coling'90*, Helsinki.
7. Jakubíček, Miloš, Aleš Horák, and Vojtěch Kovář. 2009. Mining phrases from syntactic analysis. In Václav Matoušek and Pavel Mautner, editors, *Text, Speech and Dialogue: 12th International Conference, TSD 2009, Pilsen, Czech Republic, September 13-17, 2009. Proceedings*, volume 5729 of *Lecture Notes in Computer Science*. Springer, Heidelberg, pages 124–130.
8. Kovář, Vojtěch, Aleš Horák, and Miloš Jakubíček. 2011. Syntactic analysis using finite patterns: A new parsing system for czech. In Zygmunt Vetulani, editor, *Human Language Technology. Challenges for Computer Science and Linguistics*, volume 6562 of *Lecture Notes in Computer Science*. Springer, Heidelberg, pages 161–171.
9. Lappin, Shalom and Herbert J. Leass. 1994. An algorithm for pronominal anaphora resolution. *Computatinal Linguistics*, 20(4):535–561.
10. Linh, Nguy Giang. 2006. Návrh souboru pravidel pro analýzu anafor v českém jazyce. Master's thesis, Charles University, Faculty of Mathematics and Physics, Prague.
11. Mitkov, Ruslan, Richard Evans, and Constantin Orăsan. 2002. A new, fully automatic version of Mitkov's knowledge-poor pronoun resolution method. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pages 168–186, Mexico City, Mexico, February, 17 – 23. Springer.
12. Müller, Christoph and Michael Strube. 2006. Multi-level annotation of linguistic data with MMAX2. In Sabine Braun, Kurt Kohn, and Joybrato Mukherjee, editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*. Peter Lang, Frankfurt a.M., Germany, pages 197–214.
13. Novák, Michal and Zdeněk Žabokrtský. 2011. Resolving noun phrase coreference in czech. *Lecture Notes in Computer Science*, 7099:24–34.
14. Němčík, Vašek. 2011. Extracting Phrases from PDT 2.0. In Aleš Horák and Pavel Rychlý, editors, *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2011*, pages 51–57, Brno. Tribun EU.
15. Šmerk, Pavel. 2007. Towards morphological disambiguation of czech. Ph.D. thesis proposal, Faculty of Informatics, Masaryk University.
16. Šmerk, Pavel. 2009. Fast morphological analysis of czech. In *Proceedings of the Raslan Workshop 2009*, pages 13–16, Brno. Masarykova univerzita.
17. Versley, Yannick, Simone P. Ponzetto, Massimo Poesio, Vladimir Eidelman, Alan Jern, Jason Smith, Xiaofeng Yang, and Alessandro Moschitti. 2008. BART: a modular toolkit for coreference resolution. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 962–965, Marrakech, Morocco. European Language Resources Association (ELRA).