The Saara Framework

Work in Progress

Vašek Němčík

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

Abstract. The determination of reference and referential links in discourse is one of the important challenges in natural language understanding. The first commonly adopted step towards this objective is to determine coreference classes over the set of referring expressions. We present a modular framework for automatic anaphora resolution which makes it possible to specify various anaphora resolution algorithms and to use them to build AR systems, in principle, for any natural language. The functionality of the system is shown on selected salience-based algorithms customized for Czech.

1 Introduction

In this work, we present Saara (System for Automatic Anaphora Resolution and Analysis), a framework for anaphora resolution (AR) which is modular in many ways. Modularity in the context of AR has many obvious advantages. It allows easy experimentation on various algorithms, their evaluation using various metrics, and easy use of already implemented algorithms with various languages and data formats. In our framework, this was achieved mainly by defining multiple abstraction levels and separate modules for individual phases of processing. The architecture is in accord with the principles formulated by Byron and Tetreault [1] for their own system.

Space and time constraints do not permit an investigation across the whole spectrum of AR algorithm types. We chose to focus on traditional algorithms based on the concept of salience. Salience-based algorithms serve as a sensible initial probe for a language and at the same time provide a good basis for further implementations exploiting more complex resources. We evaluate the performance of these approaches for Czech.

At the moment, we decided to disregard grammatical coreference and rather concentrate on textual anaphora. Whereas grammatical coreference phenomena are typically subject to language-specific constraints applicable to particular syntactic constructions, textual reference phenomena tend to follow similar principles across languages, and therefore it makes sense to experiment with cross-linguistic reuse of algorithms.

Petr Sojka, Aleš Horák (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2008, pp. 11–16, 2008. © Masaryk University, Brno 2008

In the next section, we briefly review other modular AR systems and AR systems for Czech proposed in existing literature. Section 3 describes the architecture of our framework, sketches the algorithms re-implemented and provides evaluation figures. Finally, the last section suggests directions for future work.

2 Related Work

At present, mechanisms for performing anaphora resolution are becoming integral parts of modern NLP systems. Also for Czech, several AR algorithms have been formulated (e.g. [2,3,4]), however, the efforts to implement them have been substantially discouraged by the lack of Czech data suitable for evaluating AR.

The situation has changed only recently through the emergence of the Prague Dependency TreeBank [5,6], which contains annotation of pronominal coreference on its tectogrammatical level of representation. The Prague Dependency TreeBank (PDT), in its version 2.0, contains tree-representations of about 50,000 sentences with approximately 45,000 manually annotated coreference links (over 23,000 grammatical ones, and over 22,000 textual ones).

The only other relevant AR system for Czech known to me at this moment, was presented in the master thesis of Linh [7]. Her system, called AČA, contains a rule-based algorithm for resolving pronominal anaphors and defines machine-learning algorithms for all types of anaphors annotated in PDT 2.0. In our opinion, the only rather minor flaw that can be pointed out is the lack of detail concerning the presented evaluation figures.

The figures given in [7] are rather difficult to compare with the figures for our system presented below. Firstly, in AČA (unlike in Saara), the annotation data is used to detect the anaphors. Secondly, it treats nodes in certain artificially generated constructions as textual anaphora, whereas in our system, we exclude them either as nodes of technical character, or beyond the range of the AR task at the analysis depth we aim at.

To avoid problems of this sort, we took inspiration from earlier modular AR systems and their advantages. Here we mention at least the two which we consider most notable.

The first one was developed at the University of Rochester by Byron and Tetreault [1]. The authors emphasize the advantages of modularity and encapsulation of the system modules into layers. For their system, they define three: the AR layer containing functions addressing AR itself, the translation layer for creating data structures, and the supervisor layer for controlling the previous layers.

Another modular system was produced by Cristea et al. [8] and defines layers from a different perspective. The representation of the discourse being processed is divided into the text layer, containing the representation of the individual referring expressions, the projection layer, consisting of feature structures with attribute values describing the individual referring expressions, and finally the semantic layer with representations of the individual discourse entities.

We agree with the authors of the above-mentioned frameworks that modularity is an invaluable characteristic in the context of AR, and we have put emphasis on this fact when laying the foundations for our own framework, described in the following section.

3 Saara and the Algorithms Re-implemented

In this section, we briefly describe the main features of our modular AR framework, sketch the algorithms used, and provide evaluation figures for them.

The main aspects of modularity reside in encapsulation of the AR algorithms. The encapsulation is achieved by defining two processing levels: markable[1] level and sentence structure level. All AR algorithms are formulated on the markable level, and thus abstract away from the actual formalism and data format used. Markable features and relations among them are accessible only through an interface that "translates" the concept in question to the actual sentence representation.

In other words, for each new data format, it is necessary to define methods determining how to recognize referential expressions, anaphors and important relationships between them (e.g. morphological agreement, embededness, the syntactic role in the current clause). Then, in principle, any AR algorithm implemented can be used with this data.

As already mentioned, we investigated classic (mainly salience-based) AR algorithms dealing with textual pronominal anaphora and compared their performance on Czech – using the annotation in PDT 2.0 as golden standard. The following algorithms have been re-implemented:

Plain Recency As a baseline, we consider an algorithm based on plain recency, which links each anaphor to the closest antecedent candidate agreeing in morphology.

The Hobbs' Syntactic Search [9] is one of the earliest yet still popular AR approaches. Unlike all other approaches mentioned here, it does not build any consecutive discourse model. It is formulated procedurally, as a search for the antecedent by traversing the corresponding syntactic tree(s). The traversal is specified by a number of straightforward rules motivated by research in transformational grammar. In spite of the fact that the underlying ideas are quite simple, the algorithm accounts for numerous common instances and its performance on English is even today regarded as respectable.

The BFP Algorithm [10] employs the principles of centering theory, a more complex theory for modeling local coherence of discourse. One of its main claims is that each utterance has a single center of attention. Further it postulates certain rules and preferences on how centers can be realized, referred

[1] Markable is a collection of sentence representation tokens that correspond to phrases that are either themselves anaphors, or have the potential to be their antecedents.

Table 1. Performance of the system in traditional measures

	Recency	Haj87	HHS95	Hobbs	BFP	L&L
Classic						
Precision	34.21	33.91	33.98	26.76	53.36	43.12
Recall	33.70	33.41	33.48	26.30	39.90	42.18
F-measure	33.95	33.66	33.72	26.53	45.66	42.64
Success rate	36.79	36.47	36.55	28.71	43.56	46.05
MUC-6						
Precision	41.78	41.33	41.33	38.87	52.26	49.86
Recall	37.28	36.81	36.80	33.91	39.20	46.28

to etc. These rules account for numerous phenomena concerning anaphors, such as certain garden-path effects. The BFP algorithm considers all possible referential linking combinations between two neighbouring utterances and applies the claims of centering theory to rule out the implausible ones and subsequently to select the most preferred one among those left.

Activation models considering TFA[2] were originally formulated in the Praguian framework of Functional Generative Description. It stipulates that the hearer and speaker co-operate during communication to build a common structure, the so-called Stock of Shared Knowledge (SSK), which among other things, reflects the fact that some entities previously mentioned in the discourse are more activated, i.e. closer to the attention of the hearer, than others. Hajičová [2] presented a simple model of SSK and a set of rules for updating it based on the current utterance and its topic-focus articulation (TFA). These rules are applied iteratively to preserve the correctness of the model at each point of the discourse. An anaphor is linked to the most activated item of the SSK agreeing in morphology. Hajičová, Hoskovec, and Sgall [4] extended the previous model by defining a more fine-grained activation scale and certain referential constraints.

The method of combining salience factors is inspired by the RAP system presented by Lappin and Leass [11]. Its main idea is that the salience of a discourse object is influenced by a variety of factors. Each of them contributes to its salience in an uniform way and can be attributed to certain well-defined features of the corresponding referring expression and its context. For example, one factor "rewards" entities mentioned in the subject position. With each new sentence, the salience values of all previously considered entities are cut by half to account for salience fading through time. The antecedent of an anaphor is identified as the most salient object according to the adopted factors.

The evaluation figures for the above-mentioned AR algorithms within Saara are given in Table 1. The presented results offer an exceptional opportunity to compare the individual algorithms. It is always difficult to compare results given by different authors, or obtained through different systems, as it is

[2] TFA stands for Topic-focus articulation; similar ideas are also known as information structure, or functional sentence perspective.

usually not clear what exactly has been counted and taken into account, and whether the choice of the data set used for evaluation did play a role. In contrast, the figures provided here offer a basis for a straightforward comparison, as they were acquired using the same pre-processing facilities, the same linguistic assumptions and the same data.

4 Future Work

The main goal of our work is to arrive at a modular, state-of-the-art AR system for Czech that would be readily used as a module in bigger NLP applications. With regard to the figures presented in the previous section, it is obvious that the precision of the algorithms needs to be improved. We would like to pursue several means of achieving that.

Firstly, as error analysis revealed this to be an interesting point, we plan to investigate the referential properties of Czech anaphors across clauses of complex sentences. Next, we are interested in the role of TFA in anaphorical relations. In spite of the abundant theoretical studies, practical experiments haven't confirmed any relevant theoretical claims based on TFA values. Finally, we aim at improving the quality of the AR process by exploiting linguistic resources available for Czech, especially the Verbalex valency lexicon [12] and Czech WordNet [13].

Acknowledgments. This work has been partly supported by the Ministry of Education of CR within the Center of basic research LC536 and in the National Research Programme II project 2C06009.

References

1. Byron, D.K., Tetreault, J.R.: A flexible architecture for reference resolution. In: Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL-'99). (1999).
2. Hajičová, E.: Focusing – a meeting point of linguistics and artificial intelligence. In: Jorrand, P., Sgurev, V. (Eds.): Artificial Intelligence Vol. II: Methodology, Systems, Applications. Elsevier Science Publishers, Amsterdam (1987), pp. 311–321.
3. Hajičová, E., Kuboň, P., Kuboň, V.: Hierarchy of salience and discourse analysis and production. In: Proceedings of Coling '90, Helsinki (1990).
4. Hajičová, E., Hoskovec, T., Sgall, P.: Discourse modelling based on hierarchy of salience. The Prague Bulletin of Mathematical Linguistics (64) (1995), pp. 5–24.
5. Hajič, J., et al.: The Prague Dependency Treebank 2.0. Developed at the Institute of Formal and Applied Linguistics, Charles University in Prague. (2005) `http://ufal.mff.cuni.cz/pdt2.0/`.
6. Kučová, L., Kolářová, V., Žabokrtský, Z., Pajas, P., Čulo, O.: Anotování koreference v pražském závislostním korpusu. Technical report, Charles University, Prague (2003).
7. Linh, N.G.: Návrh souboru pravidel pro analýzu anafor v českém jazyce. Master's thesis, Charles University, Faculty of Mathematics and Physics, Prague (2006).

8. Cristea, D., Postolache, O.D., Dima, G.E., Barbu, C.: AR-engine – a framework for unrestricted co-reference resolution. In: Proceedings of The Third International Conference on Language Resources and Evaluation, LREC 2002, Las Palmas, Spain (2002).
9. Hobbs, J.R.: Resolving pronoun references. In: Grosz, B.J., Spärck-Jones, K., Webber, B.L. (Eds.): Readings in Natural Language Processing. Morgan Kaufmann Publishers, Los Altos (1978), pp. 339–352.
10. Brennan, S.E., Friedman, M.W., rd, C.J.P.: A centering approach to pronouns. In: Proceedings of the 25th Annual Meeting of the ACL, Stanford (1987), pp. 155–162.
11. Lappin, S., Leass, H.J.: An algorithm for pronominal anaphora resolution. Computational Linguistics **20**(4) (1994), pp. 535–561.
12. Hlaváčková, D., Horák, A.: VerbaLex – New Comprehensive Lexicon of Verb Valencies for Czech. In: Computer Treatment of Slavic and East European Languages, Bratislava, Slovakia, Slovenský národný korpus (2006), pp. 107–115.
13. Pala, K., Smrž, P.: Building Czech Wordnet. Romanian Journal of Information Science and Technology **7**(2–3) (2004), pp. 79–88.