A Cascaded Shallow Approach to Reference Resolution

Thierry Declerck and **Günter Neumann***
DFKI GmbH
Stuhlsatzenhausweg
D-66123 Saarbruecken, Germany
{declerck,neumann}@dfki.de

Abstract

In this paper we present the design of a reference resolution module (RRM) within an Information Extraction (IE) system for German free text. We show in some details how such a module can be distributed over sub-components of a shallow Natural Language (NL) processing chain, defining a modular approach to reference resolution.

1 Introduction

Information Extraction (IE) systems are identifying, collecting and normalizing relevant information from natural language (NL) text. The information of interest is typically pre-specified in form of uninstantiated domain and application specific frame-like structures called *templates*. The major task of an IE-system is thus the identification of the relevant parts of the (free) text that can be used to fill a template's slot. Filled templates can further be combined and/or merged in order to put the information in a coherent sequence of events (so for example in the case of management succession scenarios).

Reference resolution is necessary in order to ensure and improve both the template filling and the template combining/merging task. In the first case it is important, for example, to know when a pronoun refers to a domain-relevant entity, so that the pronoun (with a pointer to its related referential expression) can be selected for filling the appropriate template's slot. And in the second case, it is useful to know when distinct linguistic realizations of domain-relevant entities are coreferent, so that referential continuity can be established through a set of filled templates. It is also important to handle ellipsis resolution since otherwise, due to the lack of an overt syntactic realization, some domain-relevant entities, relations and events would not be detected. For all those reasons we have been defining and implementing a general approach to reference resolution, dealing with anaphora, coreference[1] and nominal ellipsis resolution[2]. Following a generally recommended strategy for handling co-referentiality in IE[3] we designed a separate module for reference resolution, which we call RRM. Since IE systems have to deal with large collection of unrestricted texts, the linguistic processing component of such systems has to be robust and efficient and most IE systems are therefore using shallow parsing techniques. It is in such a (cascaded) shallow processing architecture that we integrated RRM. But contrary to most of the IE systems known to us[4], RRM do not just apply to the result of the whole shallow NL processing chain but is interleaving with the subsystems of the NLP component and defines resolution rules at the various levels of the linguistic processing, proposing thus a modular approach to reference resolution.[5]

* The research underlying this paper was supported by grants from the German Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMB+F) to the DFKI project PARADIME (FKZ ITW 9704) and by grants from the EU to the project MUMIS (IST-1999-1056). Thanks also to the anonymous reviewers for their fruitful comments.

[1] We adopt the definition of "coreference" as being a symmetrical and transitive relation between referential NPs, and distinguish thus coreference from anaphora resolution, since the latter involves at least one non-referential element (the pronoun) and is not inherently symmetric. We stress this distinction here, since often in the literature the word "coreference" is used in the more general sense of two items sharing a reference.

[2] In our actual implementation, only "nominal" ellipses are handled: we are looking for missing heads in nominal sequences and generate (re-construct) the corresponding NP. (Lappin & Shih 96) is proposing an algorithm for ellipsis resolution taking into account verbal constructs, which is the common view on ellipsis

[3] For example (Appelt 99) notes that anaphoric elements are so constrained in the way they can match their possible antecedents, that a sole template unification algorithm could not solve this task in a satisfying way, and therefore stresses the need for of a separate coreference module coming into play between the output of the NLP components and the domain-specific analysis.

[4] See for example (MUC 95), (MUC 98) or the general description of IE technology in (Appelt 99).

[5] The design and the coverage of the reference resolution module has been directed by a manual corpus anal-

In the rest of the paper we briefly present the shallow text processing core engine in which the RRM is integrated and then describe the processing levels for which the reference resolution algorithm has been implemented already.[6]

2 The Shallow Core Text Processor

The shallow text processor has been designed specifically for efficient and robust processing of German text documents (see (Piskorski & Neumann 00)). Two main levels of processing are distinguished, the word level and the phrase level. Word level processing is subdivided into several components. First, a TEXT TOKENIZER maps sequences of characters into greater units, usually called tokens. Each token, which is identified as a potential word form is then lexically processed by a subsequent LEXICAL PROCESSOR. Lexical processing includes morphological analysis, on-line recognition of compounds and hyphen coordination. Afterwards the control is passed to the final step of processing on the word level, a POS FILTER, which performs word-based disambiguation. The overall task of the phrase level processing is to construct a hierarchical structure over the words of a sentence. This level in STP is divided into three modules. First a NAMED ENTITY FINDER identifies specialized expressions for time, date and several proper name expressions. Then a PHRASE RECOGNIZER identifies general nominal phrases, prepositional phrases and verb groups. In the final step, a CLAUSE RECOGNIZER analyzes the dependency-based structure of the fragments of each sentence following a divide-and-conquer strategy as described in (Neumann *et al.* 00). All those modules are based on finite-state devices. The result of each component as well of the whole processing is available in form of XML annotated text.

It is important to stress here STP's high modularity: each component can be used in isolation. This capability was also one the reasons for implementing a distributed algorithm for reference resolution: adding available co-reference information to an NLP component helps on the one hand in improving the quality of the applications building on this component, and is on the other hand supporting the subsequent processing components.

ysis and is therefore text oriented, but various theoretical insights of distinct binding theories (GB and HPSG) or discourse theories (Discourse Representation Theory (DRT) and Dynamic Predicate Logic (DPL)) have been taken into account. We could also benefit from some works dealing with shallow approaches to anaphora resolution for English (for example (Kennedy & Boguraev 96)) and (Mitkov 98) (the latter also proposing solutions for Arabic and Polish).

[6] An overall description of the methodology and of the algorithms used will be given in a forthcoming technical report.

3 The Reference Resolution Module

3.1 A Modular Approach

RRM implements a general reference resolution strategy, dealing with (nominal) ellipsis, pronoun and coreference resolution tasks at various processing levels of the generic (cascaded) linguistic components of the IE system[7], defining thus a multi-layered and modular strategy for reference resolution. This strategy also allows to verify for German some of the results achieved by studies on anaphora resolution at lower levels of NL processing, at a pre-parser level (Kennedy & Boguraev 96) or even at a level involving still less linguistic knowledge (Mitkov 98). So our approach can also be seen in the context of works looking for robust reference resolution. But the algorithms implemented in RRM will not always force a decision in a case of multiple candidates for a resolution, since in the cascaded chunck approach some decisions can be left open for a further processing step[8].

3.2 The Processing Levels

RRM applies for the time being at three distinct level of the NL processing:

3.2.1 Compound Analysis

Already at the level of morphological compound analysis some reference resolution is proposed. The traditional compound analysis has been extended to the so-called "hyphen coordination" of words, where some common substring of the conjuncts is only realized by the last conjunct of the coordination, whereas in the other conjuncts a hyphen is marking the elision: "Obst-, Gemüse- und Milcherzeugnispreise sind..." (*The price of fruits, vegetables and milk products*). The

[7] Excluding for the time being the CLAUSE RECOGNIZER module.

[8] The last processing step is not necessarily the parser (with GFs detection). In IE systems one can delay some decisions till the template filling/merging task (including domain analysis). Thus delaying decisions involving semantics or domain-knowledge to the level where this information is fully available and can be processed by the correpsonding techniques. In (Gaizauskas & Humphreys 96) interesting strategies are described, solving references at the level of (template-based) discourse modeling.

successful compound analysis of the last conjunct is extended to the other conjuncts, substituting the hyphen mark with the detected compound suffix: "Obstpreise, Gemüsepreise und Milcherzeugnispreise sind...". The whole (resolved) sequence is marked by the algorithm as being a "hyphen coordination".

3.2.2 Named Entities Recognition

At this level we look at all kinds of coreference involving named entitites. So for example in the case of the word sequence "Martin Marietta Corp." occuring in our corpus, the whole string is detected as a company name (because of the designator "Corp.") and if in the continuation of the text the string "Martin Marietta" (or "Marietta" alone) is occuring, then this string will be corefered to the first (complete) string introducing this named entity. This feature is also used in order to establish a coreference relation between all occurences of the substrings of the detected NE. If one of the substrings occuring alone is also encoded as a normal word in the generic lexicon[9] ambiguity will be generated (to be solved at subsequent processing levels) and the expression is considered to act a as candidate NE.

3.2.3 Linguistic Fragment Recognition

At this level we have handled the non-local (nominal) ellipsis resolution, since we try to find in the preceeding text larger structures giving some information for the syntactic reconstruction of the elided structure, and for this we need structural information delivered by the fragment recognizer.

We also solve all kinds of pronominal resolutions at this level[10], using agreement information as a filter for the detection of antecedent candidates. We prefer to solve this task at this level over just looking at the items tagged as substantives by the POS filter since the computation of the agreement information of a phrase helps in excluding a lot of morpholgical ambiguities attached to the lexical item and thus avoid considering certain candidates at the word level (this remark being specially valid for rich morphological languages, like German). Recognized NPs also carry de facto information about definitness of a nominal expression, an important information for determining the binding scope of NPs (as has been shown by various semantic discourse theories). And a further advantage of proposing the anaphoric resolution at this level of processing concerns the fact that attaching a pronoun to a whole fragment allows to bind the pronoun also to the modifications of the head noun, in the case the fragment is an NP[11].

In the case of multiple candidates for the antecedent relation, we have established certain preferences in dependency of the distance and the type of pronouns involved, but we take no final decisions at this level. The preference rules defined at this level of the analysis can be overruled by some stronger rules defined at some subsequent level of processing.

[9] A famous example being given by the former chancellor of Germany, Helmut Kohl, "Kohl" meaning as well "cabbage".

[10] At least RRM establishes a list of candidates and propose some preferences, since in some cases the final decision cannot be taken at this level. So for example when agreement and distance do not suffice in uniquely determining an antecedent, the detection of GFs can often help in taking a decision.

[11] Which is mostly the case but sometimes pronouns can also refer to whole event, which are normally expressed by a complete clause.

References

(Appelt 99) Doug E. Appelt. An introduction to information extraction. *AI Communications*, 12, 1999.

(Gaizauskas & Humphreys 96) Robert Gaizauskas and K. Humphreys. Quantitative evaluation of coreference algorithms in an information extraction system. In *Proceedings of Discourse Anaphora and Anaphor Resolution Colloquium, DAARC-96)*, 1996.

(Kennedy & Boguraev 96) Christopher Kennedy and B. Boguraev. Anaphora for everyone: Pronominal anaphora resolution without a parser. In *Proceedings of the 16th International Conference on Computational Linguistics, COLING-96*, pages 113–118, Copenhagen, 1996.

(Lappin & Shih 96) Shalom Lappin and H-H. Shih. A generalized algorithm for ellipsis resolution. In *Proceedings of the 16th International Conference on Computational Linguistics, COLING-96*, pages 687–692, Copenhagen, 1996.

(Mitkov 98) Ruslan Mitkov. Robust pronoun resolution with limited knowledge. In *Proceedings of the 17th International Conference on Computational Linguistics, COLING-98*, pages 869–875, Montreal, 1998.

(MUC 95) MUC, editor. *Sixth Message Understanding Conference (MUC-6)*. Morgan Kaufmann, 1995.

(MUC 98) MUC, editor. *Seventh Message Understanding Conference (MUC-7)*, http://www.muc.saic.com/, 1998. SAIC Information Extraction.

(Neumann *et al.* 00) Guenter Neumann, C. Braun, and J. Piskorski. A divide-and-conquer strategy for shallow parsing of german free texts. In *Proceedings of the 6th Conference on Applied Natural Language Processing, ANLP-00*, 2000.

(Piskorski & Neumann 00) Jakub Piskorski and G. Neumann. An intelligent text extraction and navigation system. In *Proceedings of the 6th Conference on Recherche d'Information Assistée par Ordinateur, RIAO-2000*, 2000.