Exploiting logical document structure for anaphora resolution

Daniela Goecke, Andreas Witt

Universität Bielefeld
Fakultät für Linguistik und Literaturwissenschft
- Computerlinguistik und Texttechnologie -
Postfach 10 01 31, 33501 Bielefeld, Germany
{daniela.goecke, andreas.witt}@uni-bielefeld.de

Abstract

The aim of the paper is twofold. Firstly, an approach is presented how to select the correct antecedent for an anaphoric element according to the kind of text segments in which both of them occur. Basically, information on logical text structure (e.g. chapters, sections, paragraphs) is used in order to select the antecedent life span of a linguistic expression, i.e. some linguistic expressions are more likely to be chosen as an antecedent throughout the whole text than others. In addition, an appropriate search scope for an anaphora expressed by an expression can be defined according to the document structuring elements that include the linguistic expression. Corpus investigations give rise to the supposition that logical text structure influences the search scope of candidates for antecedents. Second, a solution is presented how to integrate the resources used for anaphora resolution. In this approach, multi-layered XML annotation is used in order to make a set of resources accessible for the anaphora resolution system.

1. Introduction

For anaphora resolution various types of information have to be taken into account (e.g. grammatical form, grammatical function, agreement constraints or collocation patterns). Additionally, information on the possible distance between antecedent and anaphora is of crucial importance. Distance can be measured as distance in words, sentences, paragraphs on the textual level or as distance in discourse entities on the discourse level (Strube and Müller, 2003; Xiaofeng et al., 2004; Poesio and Kabadjov, 2004). Mitkov (2002, p.17f) points out that information about the possible distance ”is not only interesting from the point of view of theoretical linguistics, but can be very important practically and computationally in that it can narrow down the search scope of candidates for antecedents.”

Corpus investigations show that the distance between anaphora and its antecedent varies according to the NP type of the anaphora (for an overview see Mitkov, 2002). Vieira & Poesio (2001) describe heuristics for the life span of antecedents for definite descriptions. The authors point out that due to the hierarchical organization of text segments some candidates for antecedents are accessible even if they are not in the defined window whereas others are not accessible although located within the defined window. Tetreault and Allen (2004) describe a pronoun resolution algorithm augmented with discourse segmentation information.

In addition to the information for anaphora resolution that have been mentioned above, the paper argues for an approach that includes information on the text segments in which a linguistic expression is located, too. According to the document structuring elements that include a linguistic expression, an appropriate search scope for antecedent can be defined.

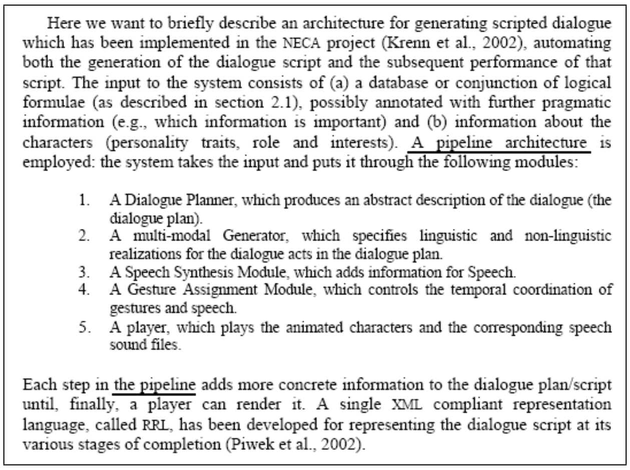
The remainder of the paper is stuctured as follows: Section 2 describes the corpus under investigation and the annotation schemata for anaphoric relations and logical document structure. Section 3 accounts for the benefit of integrating logical document structure for anaphora resolution and Section 4 presents the architecture for the integration of different levels of information. Section 5 derives a conclusion and gives clues for further development.

2. Corpus

The findings presented in this paper are based on a corpus of german scientific articles and from an additional small set of both english and german newspaper and scientific articles. The corpus has been chosen because it has been annotated in a partner project for several levels of information (logical document structure, thematic level, rhetorical level). The creation of the corpus is described in detail in Bayerl et al. (2003). In addition, for the purpose of anaphora resolution the corpus has been annotated for discourse entities (DEs) and anaphoric relations between the DEs. These annotation layers have been annotated according to the multi-layer annotation approach presented in Witt (2002).

2.1. Annotation of cohesive means

Anaphoric relations have been annotated using the annotation schema described in Holler et al. (2004). The perspective adopted in our approach is that anaphoric relations do not hold between the linguistic forms but between the discourse entities that are realized by these linguistic forms. A relevant discourse entity is a linguistic form that introduces a discourse referent in the sense of Kamp and Reyle (1993) (q.v. Karttunen (1976)). As we focus on nominal anaphora, especially definite description anaphora, only discourse entities of nominal type are annotated. Discourse entities of different types (e.g. propositional) are not taken into account. Anaphoric relations hold between discourse entities and are annotated as a kind of standoff annotation. For each anaphoric relation a tupel `cospecLink` is described that defines the relation type (`relType`, e.g. identity or bridging) between the anaphoric element (`phorIDRef`) and its antecedent(s) (`antecedentIDRefs`). Both the anaphoric element and the antecedent are modelled as attributes of type `IDREF` that refer to discourse entities. Up to now, a subset of the corpus has been manually annotated for anaphoric relations, a sample annotation of the text given in Figure 1 is shown in Figure 2 (example text taken from Piwek et al. (2005)).

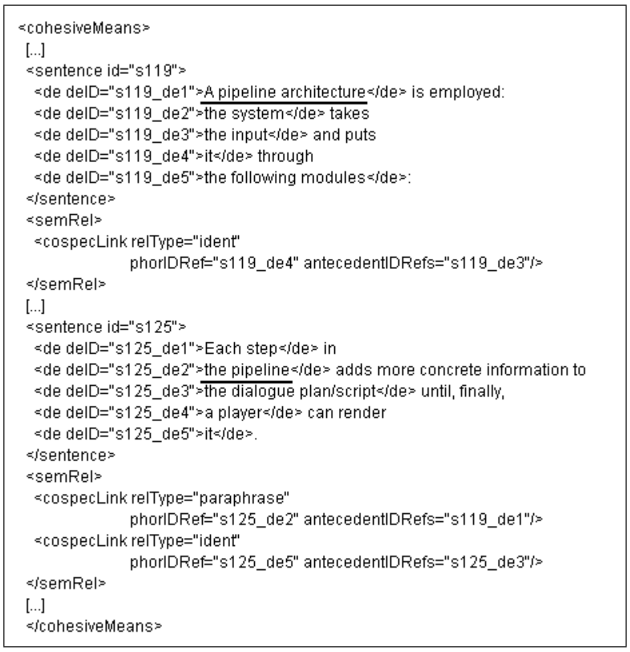


Here we want to briefly describe an architecture for generating scripted dialogue which has been implemented in the NECA project (Krenn et al., 2002), automating both the generation of the dialogue script and the subsequent performance of that script. The input to the system consists of (a) a database or conjunction of logical formulae (as described in section 2.1), possibly annotated with further pragmatic information (e.g., which information is important) and (b) information about the characters (personality traits, role and interests). A pipeline architecture is employed: the system takes the input and puts it through the following modules:

1. A Dialogue Planner, which produces an abstract description of the dialogue (the dialogue plan).
2. A multi-modal Generator, which specifies linguistic and non-linguistic realizations for the dialogue acts in the dialogue plan.
3. A Speech Synthesis Module, which adds information for Speech.
4. A Gesture Assignment Module, which controls the temporal coordination of gestures and speech.
5. A player, which plays the animated characters and the corresponding speech sound files.

Each step in the pipeline adds more concrete information to the dialogue plan/script until, finally, a player can render it. A single XML compliant representation language, called RRL, has been developed for representing the dialogue script at its various stages of completion (Piwek et al., 2002).

Figure 1: Example text



```
<cohesiveMeans>
 [...]
 <sentence id="s119">
  <de deID="s119_de1">A pipeline architecture</de> is employed:
  <de deID="s119_de2">the system</de> takes
  <de deID="s119_de3">the input</de> and puts
  <de deID="s119_de4">it</de> through
  <de deID="s119_de5">the following modules</de>:
 </sentence>
 <semRel>
  <cospecLink relType="ident"
             phorIDRef="s119_de4" antecedentIDRefs="s119_de3"/>
 </semRel>
 [...]
 <sentence id="s125">
  <de deID="s125_de1">Each step</de> in
  <de deID="s125_de2">the pipeline</de> adds more concrete information to
  <de deID="s125_de3">the dialogue plan/script</de> until, finally,
  <de deID="s125_de4">a player</de> can render
  <de deID="s125_de5">it</de>.
 </sentence>
 <semRel>
  <cospecLink relType="paraphrase"
             phorIDRef="s125_de2" antecedentIDRefs="s119_de1"/>
  <cospecLink relType="ident"
             phorIDRef="s125_de5" antecedentIDRefs="s125_de3"/>
 </semRel>
 [...]
 </cohesiveMeans>
```

Figure 2: Annotation of anaphoric relations

2.2. Logical document structure

The logical document structure describes the organisation of the text document in terms of chapters, sections, paragraphs, and the like. Based on the logical document structure (e.g. DocBook, or LaTeX, HTML[1]), a layout-oriented presentation can be generated. This structure is application-independent, and especially for texts from e-publishing sources a set of logical document structure elements will be easily available which can be used to identify different text segments. For the corpus under investigation the DocBook DTD has been chosen, which is a standard originally developed for technical documentation (Walsh and Muellner, 1999), e.g. manuals, but has been recently also used in academic writing. For the annotation, a subset of the DocBook DTD extended by additional logical elements (e.g. <toc> for table of contents) has been developed (Bayerl et al., 2003). The approach presented in this paper describes a possibility to use these structuring elements that are most often available when creating a corpus. A sample annotation is given in the next section.

3. Benefit

The influence of the logical document structure on the choice of an antecedent might be either (a) a direct influence on the discourse entities (or antecedent life span) or (b) an influence on the search window (comparable to different window sizes according to the NP type of the anaphora). The first type is related to the fact that discourse entities ”only serve as antecedents for anaphoric expressions within pragmatically determined segments” (Vieira and Poesio, 2001, p.549). Regarding the document structure, corpus evidence shows that some discourse entities are more prominent troughout the whole document than others, e.g. discourse entities described in the abstract of a text might be accessible during the whole text whereas discourse entities that have been evoked in a footnote-structure are less likely an antecedent for anaphoric elements in the main text. The set of document structuring elements is ordered hierarchically, discourse entities described in hierarchically higher elements (e.g. `sect3`) are more likely to find their antecedents in structuring elements of the same hierarchical or higher levels (`sect1/sect2`) than in a preceding but hierarchically lower segment (`sect4/sect5`).

The influence on the search window may either enlarge the search window, i.e. the antecedent may be located outside the standard window (e.g. located in the whole paragraph or in a preceding one), or may narrow the search window, e.g. due to the start of a new chapter or section. We consider the first case to be more important as the provision for logical document structure helps to find an antecedent where otherwise (i.e. with a fixed window size) no antecedent could be found. Information on document structuring elements may help to enlarge the search window according to the context instead of enlarging the standard search window. In addition to information on paragraphs, chapters and the like, emphasized text spans may give focus information that could possibly be used in order to rank candidates of antecedents. These assumptions are described in detail in the next subsections.

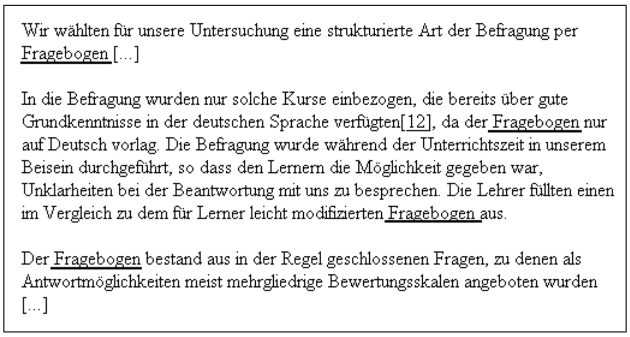
3.1. Example analysis

Figure 1 shows an example for the assumption that the search scope has to be enlarged in order to access the correct antecedent. In the example text an anaphoric relation holds between the discourse entities described by the linguistic forms *A pipeline architecture* and *the pipeline*.

Both the anaphora and the antecedent are located in the main text, the list structure between them adds 5 sentences including 18 discourse entities. Apart from enlarging the search space in general, the context should be taken into

[1] Apart from logical markup, LaTeX and HTML allow layout markup, too, e.g. in Latex **bf** for text spans to be set in bold face.

account. In case of list structures, e.g. sentences preceding the list structure should be included in the search space.
Another evidence for enlarging the search space according to text structuring information is shown in the sample of a German scientific article in Figure 3.



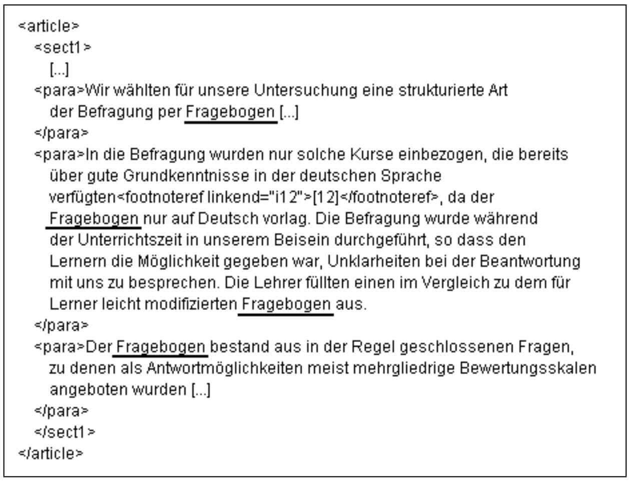
Wir wählten für unsere Untersuchung eine strukturierte Art der Befragung per Fragebogen [...]

In die Befragung wurden nur solche Kurse einbezogen, die bereits über gute Grundkenntnisse in der deutschen Sprache verfügten[12], da der Fragebogen nur auf Deutsch vorlag. Die Befragung wurde während der Unterrichtszeit in unserem Beisein durchgeführt, so dass den Lernern die Möglichkeit gegeben war, Unklarheiten bei der Beantwortung mit uns zu besprechen. Die Lehrer füllten einen im Vergleich zu dem für Lerner leicht modifizierten Fragebogen aus.

Der Fragebogen bestand aus in der Regel geschlossenen Fragen, zu denen als Antwortmöglichkeiten meist mehrgliedrige Bewertungsskalen angeboten wurden [...]

Figure 3: Example of scientific text

This example is taken from a text on *German as a Foreign Language* and describes the setting of a study on dialect. In this example the questionnaire (*Fragebogen*) used for the study is described. The first mention of *Fragebogen* introduces the questionnaire as the means of the study. The second mention describes the language of the questionnaire whereas the third mention describes that learners and teachers get different questionnaires and refers to the teachers' questionnaire. The fourth mention describes the questionnaire in general and not the teachers' questionnaire as a subset of the general one. Therefore, the anaphoric relation has to be established between the fourth and the second mention. For the correct choice of the antecedent it might be helpful to consider the document structuring elements given in Figure 4.



```
<article>
  <sect1>
    [...]
  <para>Wir wählten für unsere Untersuchung eine strukturierte Art
    der Befragung per Fragebogen [...]
  </para>
  <para>In die Befragung wurden nur solche Kurse einbezogen, die bereits
    über gute Grundkenntnisse in der deutschen Sprache
    verfügten<footnoteref linkend="i12">[12]</footnoteref>, da der
    Fragebogen nur auf Deutsch vorlag. Die Befragung wurde während
    der Unterrichtszeit in unserem Beisein durchgeführt, so dass den
    Lernern die Möglichkeit gegeben war, Unklarheiten bei der Beantwortung
    mit uns zu besprechen. Die Lehrer füllten einen im Vergleich zu dem für
    Lerner leicht modifizierten Fragebogen aus.
  </para>
  <para>Der Fragebogen bestand aus in der Regel geschlossenen Fragen,
    zu denen als Antwortmöglichkeiten meist mehrgliedrige Bewertungsskalen
    angeboten wurden [...]
  </para>
  </sect1>
</article>
```

Figure 4: Structural annotation

Concerning the logical document structure, corpus evidence shows that anaphoric elements that are located in the middle or at the end of a paragraph tend to find their antecedents within the same paragraph, whereas anaphoric elements at the begin of the paragraph tend to have a larger scope. E.g. for a German newspaper article, the majority of anaphoric elements find their antecedents within the same paragraph, most of those that find their antecedents across a paragraph boundary are located in the first part of the paragraph. Without considering the paragraph structure it would be likely to choose the directly preceding antecedent (in our example the third mention of *Fragebogen*), whereas the paragraph structure indicates the antecedent to be in one of the preceding paragraphs.
Taking these findings into account we propose to include information on document structure elements into an anaphora resolution system. In Section 4 we describe our approach for the integration of different levels of annotation.

4. Integrating logical document structure

Basis for the resolution of anaphoric relations is the annotation of discourse entities. Both the discourse entities and the logical document structure are realised as separate XML annotation layers, i.e. single XML files. This multiple annotated corpus now serves as the input to the anaphora resolution using heterogeneous linguistic resources. The architecture is shown in Figure 5. The different informational levels can be merged according to the XML-based multilayer annotation developed in previous work by the authors (Witt et al., 2005).

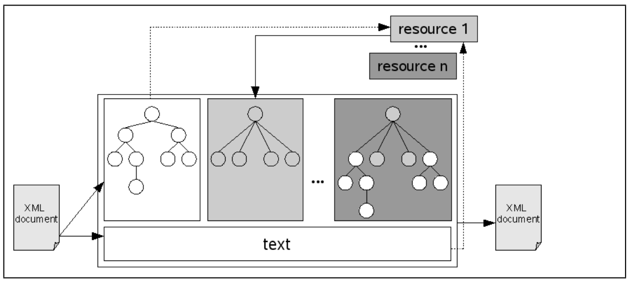


Figure 5: Integration of XML annotated data

The central idea of this architecture is to split up all annotations into their common underlying textual data (*primary data*) and different structure trees that describe the annotations. When combining the different annotations, each annotation is split from the primary data. Thus, a set of structures is spanned on the same primary data. In the current implementation, these structures are realized as a Prolog fact base[2]. Prolog has been chosen as logical programming language due to its simplicity regarding the implementation of inferences.
Each element or attribute of the XML annotation is stored as a Prolog fact describing the annotation level, the textual position within the primary data and the element name or attribute/value-pair respectively. On the basis of the Prolog facts, two XML layers can be merged, i.e. a combined XML structure is created[3]. In Figure 6, the merging process (or *markup unification*) is shown. First, all XML layers are

[2] An example of another possible representation format is the NITE object model developed by Carletta et al. (2003)

[3] Overlapping structure that cannot be encoded directly within one XML structure are encoded as milestones or fragments according to the TEI Guidelines (Sperberg-McQueen and Burnard, 2004)

converted into Prolog facts. For all markup a new hierarchical structure is created on the basis of their textual position according to the primary data. The result of the markup unification are new Prolog facts of the combined markup structure. These Prolog facts are reconverted into a single wellformed XML document. The markup unification process is described in detail in Witt et al. (2005).

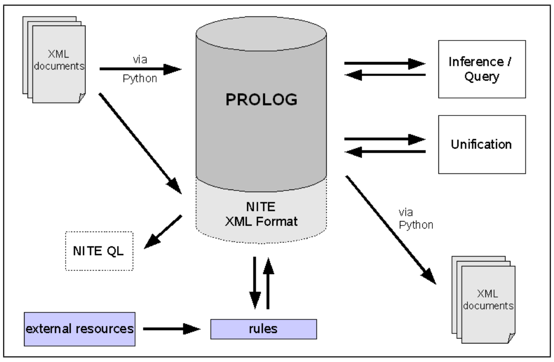


Figure 6: Integration of XML annotated data

5. Outlook

We have presented results of a corpus study on how to integrate logical document structure for the benefit of anaphora resolution. Document structuring elements might influence the choice of an anaphora's antecedent either via the antecedent life span or the antecedent search window.

Further work should extend the annotation of anaphoric relations and thus the details of the corpus study. An extension of the annotation schema will distinguish between cospecification links and bridging links. The findings of the complete corpus study will be included into the anaphora resolution system. Additional work is planned regarding the representation of the multi-layer annotation. Instead of the Prolog representation and query engine, an XML database representation and querying via XQuery is evaluated.

6. Acknowledgements

The work presented in this paper is part of the project A2 Sekimo of the Research Group ”Text-technological Modelling of Information” funded by the German Research Foundation. The annotation schema for logical document structure has been developed in the projects C1 SemDok and B1 HyTex, the corpus has been created in the C1 projekt.

7. References

P. S. Bayerl, D. Goecke, H. Lüngen, and A. Witt. 2003. Methods for the semantic analysis of document markup. In Ccile Roisin, Ethan Munson, and Christine Vanoirbeek, editors, *Proceedings of the 3rd ACM Symposium on Document Engineering (DocEng)*, pages 161–170, Grenoble.

J. Carletta, J. Kilgour, T. O'Donnell, Stefan S. Evert, and H. Voormann. 2003. The nite object model library for handling structured linguistic annotation on multimodal data sets. In *3rd Workshop on NLP and XML*, Budapest, Ungarn.

A. Holler, J.-F. Maas, and A. Storrer. 2004. Exploiting coreference annotations for text-to-hypertext conversion. In *Proceeding of LREC*, volume II, pages 651–654, Lisbon, Portugal.

H. Kamp and U. Reyle. 1993. *From Discourse to Logic*. Kluwer, Dordrecht.

Lauri Karttunen. 1976. Discourse referents. *Syntax and Semantics: Notes from the Linguistic Underground*, 7:363–385.

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

P. Piwek, R. Power, D. Scott, and K. van Deemter. 2005. Generating multimedia presentations: from plain text to screenplay. In O. Stock and M. Zancanara, editors, *Intelligent Multimodal Information Presentation.*, volume 27 of *Text, Speech and Language Technology*, pages 203–225. Springer, Dordrecht.

M. Poesio and M. A. Kabadjov. 2004. A general-purpose, off-the-shelf anaphora resolution module: Implementation and preliminary evaluation. In *Proceeding of LREC*, pages 663–666, Lisbon, Portugal.

C. M. Sperberg-McQueen and Lou Burnard, editors. 2004. *Guidelines for Text Encoding and Interchange*. published for the TEI Consortium by Humanities Computing Unit, University of Oxford.

M. Strube and C. Müller. 2003. A machine learning approach to pronoun resolution in spoken dialogue. ACL 03.

J. Tetreault and J. Allen. 2004. Dialogue structure and pronoun resolution. In *Proceedings of the 5th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC04)*, Lisbon, Portugal.

R. Vieira and M. Poesio. 2001. An empirically based system for processing definite descriptions. *Computational Linguistics*, 26(4):539–593.

N. Walsh and L. Muellner. 1999. *DocBook: The Definitive Guide*. OReilly.

A. Witt, D. Goecke, F. Sasaki, and H. Lüngen. 2005. Unification of xml documents with concurrent markup. *Literary and Linguistic Computing*, 20(1):103–116.

Andreas Witt. 2002. Meaning and interpretation of concurrent markup. In *Joint Conference of the ALLC and ACH (ALLCACH2002)*, Tübingen, Germany.

Y. Xiaofeng, J. Su, G. Zhou, and C. L. Tan. 2004. Improving pronoun resolution by incorporating coreferential information of candidates. In *Proceedings of ACL*.