Verknüpfung heterogener texttechnologischer Ressourcen

Daniela Goecke, Dieter Metzing, Andreas Witt

Fakultät für Linguistik und Literaturwissenschaft
Universität Bielefeld
Postfach 10 01 31
33501 Bielefeld
{daniela.goecke, dieter.metzing, andreas.witt}@uni-bielefeld.de

Abstract: Gegenstand des Workshop-Beitrags ist die Verknüpfung heterogener linguistischer Ressourcen. Eine bedeutende Teilmenge von Ressourcen in der gegenwärtigen linguistischen Forschung und Anwendung besteht zum einen aus XML-annotierten Textdokumenten und zum anderen aus externen Ressourcen wie Grammatiken, Lexika oder Ontologien. Es wird eine Architektur vorgestellt, die eine Integration heterogener Ressourcen erlaubt, wobei die Methoden zur Integration unabhängig von der jeweiligen Anwendung sind und somit verschiedene Verknüpfungen ermöglichen. Eine exemplarische Anwendung der Methodologie ist die Analyse anaphorischer Beziehungen[1].

1 Einleitung

Die Verknüpfung heterogener, XML-basierter Datenquellen erlaubt es, existierende linguistische Ressourcen für neue Anwendungsdomänen verwenden zu können. Als Verbindungsglied wird eine abstrakte Repräsentation von Primärdaten und Auszeichnungen genutzt. Bestehende Ansätze zur Verbindung heterogener Ressourcen unterliegen verschiedenen Einschränkungen, für die im vorliegenden Beitrag Lösungsmöglichkeiten vorgestellt werden. Mit der Definition einer abstrakten Repräsentation von Primärdaten und Auszeichnungen sollen diese Einschränkungen aufgehoben werden. (1) Nicht nur Dokumente und Dokumentgrammatiken, sondern jegliche XML-basierte Ressourcen können eingespeist werden. (2) Durch eine multiple Annotation identischer Primärdaten lässt sich sowohl das dokumentbezogene Modellierungsinventar als auch das Modellierungsinventar externer, z.B. ontologischer oder lexikalischer Ressourcen nutzen. (3) Durch das abstrakte Repräsentationsformat lässt sich der informationelle Gehalt sämtlicher Ressourcen erhalten.

In Abschnitt 2 wird zunächst der Begriff der Heterogenität dargestellt. In Abschnitt 3 folgt die Darstellung der Architektur zur Verknüpfung heterogener Ressourcen sowie die

[1] Der Beitrag baut auf Arbeiten des Projekts "Sekimo" der DFG-Forschergruppe 437 "Texttechnologische Informationsmodellierung" auf.

Beschreibung der zugrunde liegenden Methoden. Der Nutzen dieser Methodologie wird in Abschnitt 4 für den Bereich anaphorischer Beziehungen exemplifiziert. Der Beitrag schließt mit einem Ausblick in Abschnitt 5.

2 Heterogenität linguistischer Ressourcen

In zunehmendem Maße werden linguistische, heteroge Ressourcen erstellt und verwendet. Die „Heterogenität von Ressourcen" bezieht sich zum einen auf das Datenformat, in dem die Ressourcen repräsentiert sind. Zum anderen unterscheiden die Ressourcen sich in Bezug auf die Funktion, welche sie in der linguistisch-empirischen Modellierung spielen. Der Begriff „Heterogenität" wird in diesem Beitrag im Sinne von [Si04] (S. 25) verwendet:

> „Machine-readable structured linguistic documents (comparative word lists, lexicons, annotated texts, audio and audio-video recordings aligned with transcriptions (possibly annotated), grammatical descriptions, etc.) are being made available in a wide variety of formats on the Web. Until recently, the linguistics community has not been particularly concerned about the ease with which those structures can be accessed by other users, nor about the comparability of the structures that can be accessed. Now that community is beginning to realize that XML encoding provides relatively straightforward access to the intended structures. […]"

Die Kodierung linguistischer Ressourcen in einem standardisierten Datenformat stellt jedoch eine wichtige Voraussetzung für ihre Wiederverwertbarkeit dar. Die Auszeichnungssprache XML erlaubt es, linguistische Ressourcen zu strukturieren und in Form von *Dokumentinstanzen* zu repräsentieren. Am Beispiel der Anaphernauflösung soll gezeigt werden, wie bestehende heterogene, XML-basierte Ressourcen verwendet und miteinander verknüpft werden können.

3 Projektarchitektur

Voraussetzung für eine flexible Verwendung heterogener Daten ist eine Verwendung in ihrem ursprünglichen Format. Zwar wäre es möglich, die Ressourcen bei Bedarf jeweils zu konvertieren, allerdings stellt dies nicht nur einen hohen Aufwand dar, sondern kann auch zu Informationsverlust führen. Externe Daten und Wissensquellen werden deshalb in dem hier vorstellten Ansatz durch entsprechende Prozesse direkt in ein abstraktes Datenformat gespeist. Ausgabe soll ein einzelnes XML-Dokument sein, welches aus den textuellen Rohdaten sowie der abstrakten Repräsentation der verschiedenen Informationseinheiten aufgebaut wird. Das Ausgabedokument kann sowohl die gesamten Primärdaten (z.B. Informationsanreicherung von XML-Dokumenten) oder auch Textabschnitte (z.B. Informationsextraktion für Lexika) enthalten. Eine Darstellung der Architektur für das Beispiel der Anaphernauflösung findet sich in Abbildung 1. Die

zugrunde liegenden Methoden zur Verknüpfung heterogener Ressourcen werden in den nachfolgenden Abschnitten kurz vorgestellt.

Verschiedene Ressourcen unterschiedlicher Herkunft werden in das abstrakte Repräsentationsformat eingespeist. Die Primärdaten bilden hierbei das verknüpfende Element. Primärdaten sind die zugrunde liegenden nicht-annotierten Rohdaten. Die externen Wissensquellen, wie beispielsweise lexikalische Netze (z.B. WordNet), Grammatiken etc. werden verwendet, um verschiedene Prozesse zu unterstützen: syntaktisches Parsing, Generierung weiterer Auszeichnungen, Transformation oder Verfeinerung von Auszeichnungen, etc.

Die abstrakte Repräsentation ist konzeptueller Natur, die praktische Umsetzung erfolgte bisher in zwei konkreten Formaten. Eines der Formate basiert auf Vorarbeiten von [Ba03]) und wurde in Prolog umgesetzt. Eine weitere Umsetzung integriert die im Rahmen des NITE-Projekts entwickelten Repräsentation von multi-rooted trees (vgl. [Ca03]).

Durch das gemeinsame Format können nun Prozesse auf die vormals heterogenen Daten zugreifen und es können neue Informationsebenen generiert werden. Dabei sollen bestehende Auszeichnungen nicht verändert werden, sondern es werden weitere Auszeichnungsebenen hinzugefügt. Eine Teilmenge der Ressourcen ist dann in Form von Auszeichnungen repräsentiert. Es existiert somit eine Trennung in textuelle Daten und Auszeichnungen, die jede Form von externen Ressourcen darstellen können. Um aus dem Repräsentationsformat Ausgabedokumente zu generieren, werden unterschiedliche Prozesse verwendet, z.B. die im Projektverbund entwickelte Markup-Unifikation, die verwendet wird, um aus dem abstrakten Repräsentationsformat ein Ausgabedokument zu generieren. (vgl. [Wi05]).

4 Prozedurale Analyse anaphorischer Beziehungen

Ausgehend von der Auszeichnung und Modellierung koreferentieller Phänomene bzw. anaphorischer Beziehungen (vgl. [Sa02]) wird die im vorigen Abschnitt beschriebene Architektur für die Auflösung anaphorischer Beziehungen angewendet. Abbildung 1 zeigt die integrierte Nutzung verschiedener Ressourcen zur Anaphernauflösung. Verfahren zur Auflösung von anaphorischen Beziehungen arbeiten in zunehmendem Maße unter Zuhilfenahme von Korpusdaten. Dabei werden verschiedene, zu den ausgezeichneten Daten externe Ressourcen genutzt: manuell erstellte Regeln über referentielle Beziehungen (vgl. [PK04]), unter Einsatz stochastischer, maschineller Verfahren automatisch erlernte Regeln oder eine Kombinationen dieser Ansätze.

Für die Integration der Ressourcen werden drei Prozesse definiert: Eingabe ist ein Text, der mit Paragraphsegmentierungen, d.h. <p> Elementen ausgezeichnet ist. Im Prozess 1 werden Informationen über Wortarten und Nominalphrasen erzeugt. Prozess 2 verwendet die externe Ressource WordNet sowie die durch Prozess 1 generierten Auszeichnungen von Nominalphrasen. Für jedes Nomen in einer Nominalphrase wird die eindeutig identifizierbare Position in der Hyperonymie-Hierarchie von WordNet

durch einen Attribut-Wert Paar beschrieben, z.B. position="hyp-1.2.5" für das Nomen Animal. Diese Informationen bilden die Eingabe für Prozess 3. In diesem Prozess wird die Hyperonymie-Beziehung zwischen Animal und Fox bzw. Animal und Hens aus der extern vorliegenden WordNet-Datenbank inferiert.

Diese Beziehung bildet in Prozess 3 die Grundlage der Generierung des <bridging> Elements für Animal. Die <p> Elemente, die bereits vor der Ausführung der Prozesskette vorlagen, werden in Prozess 3 ebenfalls genutzt. Sie beschränken den Suchraum der Anaphernauflösung auf den jeweils aktuellen Paragraphen. Die Anwendung der Markup-Unifikation führt schließlich zu einem Ausgabedokument, das die ursprünglich gegebenen Paragraphenauszeichnungen mit den generierten Bridging-Relationen kombiniert.

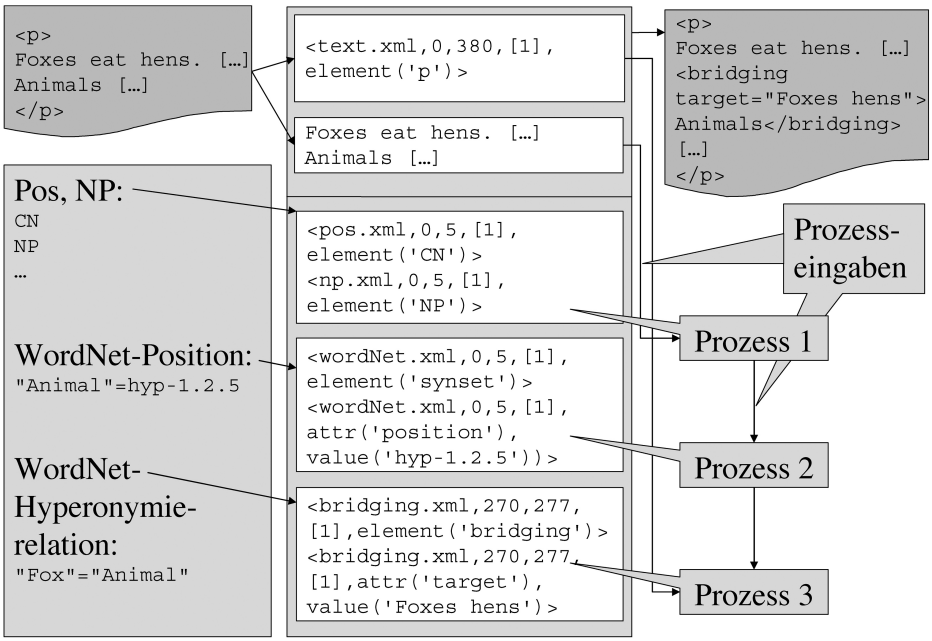

Abbildung 1: Auflösung von Bridging-Beziehungen

5 Zusammenfassung und Ausblick

Textuelle Auszeichnungen verschiedener linguistischer Beschreibungsebenen einerseits und heterogene (externe) Ressourcen andererseits sind ein bedeutender Bestandteil gegenwärtiger linguistischer Forschungen und Anwendungen. Sie stellen somit eine zentrale Variante von Heterogenität linguistischer Ressourcen dar. Dennoch gibt es bisher keine Verfahren, die eine nachhaltige, für variierende Anwendungsszenarien nutzbare Verknüpfung derartiger Ressourcen erlauben. Der Beitrag stellt als

Lösungsansatz die Verwendung des beschriebenen abstrakten Repräsentationsformates für textuelle Auszeichnungen vor. Dieses Format dient als Verknüpfungsglied für heterogene, in XML repräsentierte Ressourcen. Durch automatische Verarbeitungsprozesse werden externe Ressourcen als Auszeichnungen in die abstrakte Repräsentation eingebracht. Diese Prozesse müssen nur einmalig ausgeführt werden, und die generierten Auszeichnungen sind unabhängig von anwendungsspezifischen Prozessketten. Auf diese Weise soll ermöglicht werden, für variierende Anwendungs- und Forschungsszenarien verschiedenste Kombinationen der gleichen Ressourcen zu erzeugen.

Literaturverzeichnis

[Ba03] Bayerl, P. S., H. Lüngen, D. Goecke, A. Witt und D. Naber (2003) Methods for the semantic analysis of document markup. In: Roisin, Cécile, Ethan Munson und Christine Vanoirbeek (Hrsg.): Proceedings of DocEng 2003, INRIA Rhône-Alpes, Grenoble.

[Ca03] Carletta, J., J. Kilgour, J., T. O'Donnell S. Evert, Stefan und 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.

[Cl77] Clark, H. H. (1977) Bridging. In P.N. Johnson-Laird und P.C. Wason (Hrsg.) Thinking: Readings in Cognitive Science. London New York. Cambridge University Press.

[PK04] Poesio, M. und M. A. Kabdajov (2004) A General-Purpose, off-the-shelf Anaphora Resolution Module: Implementation and Preliminary Evalutation. In: Proceedings of LREC 2004, Lissabon.

[Sa02] Sasaki, F., C. Wegener, A. Witt, D. Metzing und J. Pönninghaus (2002) Co-reference annotation and resources: a multilingual corpus of typologically diverse languages. In: Proceedings of LREC 2002, Las Palmas.

[SW04] Sasaki, F. und A. Witt (2004) Co-reference in Japanese task-oriented dialogues: A contribution to the development of language-specific and general annotation schemes and resources. In: Proceedings of LREC 2004, Lissabon.

[Si04] Simons, G., W. Lewis, S. Farrar, T. Langendoen, B. Fitzsimons und H. Gonzalez (2004) The Semantics of Markup: Mapping Legacy Markup Schemas to a Common Semantics. In: Proceedings of the ACL 2004 Workshop on RDF/RDFS and OWL in Language Technology (NLPXML-2004), Barcelona.

[SHR00] Sperberg-McQueen, C. M., C. Huitfeldt und A. Renear (2000) Meaning and interpretation of Markup. Markup Languages 2.3, 215-234.

[VP00] Vieira, R. und M. Poesio (2000) An Empirically-based System for Processing Definite Descriptions. Computational Linguistics 26:4, S. 525-579.

[Wi05] Witt, A., Goecke, D., Sasaki, F., Lüngen, H. (2005). Unification of XML Documents with Concurrent Markup. Literary and Linguistic Computing 2005 20(1):103-116