

Die Repräsentation und Auflösung von ambigen Wortbedeutungen in der Computerlinguistik

SALTO

PD Dr. Sabine Schulte im Walde

Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart

27. März 2010

- Werkzeug für die manuelle Annotation in einer graphischen Umgebung
- Annotation von mehreren Ebenen linguistischer Information
- Korpus-Management
- Qualitätskontrolle

- Eingabe-Korpora sind annotiert oder pseudo-annotiert.
- Input-Format: TIGER XML oder SALSA/TIGER XML
- Transformation zu TIGER XML durch *TIGERRegistry*
- Format-Voraussetzung: einer oder mehrere Bäume
- Konzeptualisierung: gerichteter Graph
- Referenz auf beliebige Knoten in der Struktur
- Repräsentation von Konstituenten, Dependenz-Struktur, diskontinuierlichen Konstituenten

SALTO: besondere Eigenschaften

- diskontinuierliche Annotation: eine Markierung kann sich auf mehr als einen Knoten beziehen
- Kontext der Annotation:
 - beliebig großes Kontext-Fenster kann angeschaut werden
 - Annotation kann sich auf den Kontext (außerhalb des aktuellen Satzes) beziehen
- Unterspezifikation: demselben Knoten können mehrere Annotationen zugewiesen und in einer unterspezifizierten Menge zusammengefasst werden
- Das Tagset kann im Voraus oder während der Annotation definiert werden.

- ➊ Auswahl der Sätze zum Annotieren
- ➋ Verteilung der Annotations-Daten an Annotatoren
- ➌ Einsammeln der annotierten Daten
- ➍ Manuelle Prüfung und Korrektur von Uneinigkeiten

SALTO und Korpora aufrufen

- 1 Definieren Sie die Umgebungsvariable \$SALTO:
`export SALTO="/proj/courses/ambig-ws0910/SALTO"`
- 2 Gehen Sie in das SALTO-Verzeichnis: `cd $SALTO`.
- 3 Rufen Sie SALTO auf: `./SaLsa.sh`.
- 4 Loggen Sie sich mit Ihrem Benutzernamen (ambigX) ein.
Registrieren Sie sich NICHT als Administrator.
- 5 Befinden sich Dateien oder Verzeichnisse unter `Master`, haben Sie sich aus Versehen als Administrator eingeloggt. Bitte loggen Sie sich wieder aus.
- 6 Machen Sie mit der linken Maustaste einen Doppelklick auf das Verzeichnis `in` unter `User`. Sie sehen nun drei Korpora: *bestehen*, *glatt*, *leiter*.
- 7 Klicken Sie mit der rechten Maustaste auf das Korpus, das Sie annotieren möchten und wählen Sie `Move file to working directory`.
Das Korpus befindet sich nun im Verzeichnis `work`. Sie sehen es, wenn Sie mit der linken Maustaste einen Doppelklick auf das Verzeichnis `work` machen.
- 8 Machen Sie mit der linken Maustaste einen Doppelklick auf das Korpus im Verzeichnis `work`, das Sie annotieren möchten. Ignorieren Sie die Fehlermeldung, indem Sie mit `OK` bestätigen.

Sie haben nun das Korpus mit seiner syntaktischen Annotation geöffnet und können Ihre Annotation hinzufügen.

Mit SALTO annotieren

- 1 Machen Sie mit der rechten Maustaste einen einfachen Klick auf das Wort (*bestehen, glatt, Leiter*), das Sie annotieren möchten. Wählen Sie `Invoke frame` und dann das Wort, das Sie annotieren. Sie erhalten ein grünes Kästchen oberhalb des Wortes.
- 2 Machen Sie mit der rechten Maustaste einen einfachen Klick auf das grüne Kästchen, wählen Sie `Add element` und dann die Wortbedeutung (1-5) des Wortes in diesem Kontext, gemäß Ihrer Definition.
- 3 Falls Sie eine Annotation korrigieren möchten, machen Sie mit der rechten Maustaste einen einfachen Klick auf das grüne Kästchen und wählen Sie `Delete`.
- 4 Sie können sich anhand der Satznummern oder Pfeiltasten unterhalb des Satzes zwischen den Sätzen bewegen.
- 5 Sichern Sie Ihre Annotation, wenn Sie das Korpus schließen:
`File` → `Close corpus` → `Save`.
- 6 Wenn Sie die Annotation eines Korpus beendet haben, schieben Sie das Korpus in das Verzeichnis `out`: Machen Sie mit der rechten Maustaste einen einfachen Klick auf das Korpus im Verzeichnis `work` und wählen Sie `Move file to 'out' directory`. Das Korpus befindet sich nun im Verzeichnis `out`.



Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, and Sebastian Padó.

SALTO – a Versatile Multi-Level Annotation Tool.

In Proceedings of the 5th Conference on Language Resources and Evaluation, Genoa, Italy, 2006.



Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal.

The SALSA Corpus: a German Corpus Resource for Lexical Semantics.

In Proceedings of the 5th Conference on Language Resources and Evaluation, Genoa, Italy, 2006.



Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith.

The TIGER Treebank.

In Proceedings of the Workshop on Treebanks and Linguistic Theories, Sozopol, Bulgaria, 2002.



Wolfgang Lezius.

TIGERSearch – Ein Suchwerkzeug für Baumbanken.

In Proceedings der 6. Konferenz zur Verarbeitung natürlicher Sprache, Saarbrücken, Germany, 2002.



Wolfgang Lezius.

Ein Suchwerkzeug für syntaktisch annotierte Textkorpora.

PhD thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 2002.

Published as AIMS Report 8(4).