Regelbasierte Koreferenz mit BART

Forschungsplan zum Softwareprojekt im Sommersemester 2014

Julian Baumann, Xenia Kühling, Sebastian Ruder

13. Mai 2014

1 Projektbeschreibung

BART, das "Beautiful Anaphora Resolution Toolkit", wurde beim Projekt "Exploiting Lexical and Encyclopedic Resources For Entity Disambiguation" am Johns Hopkins Summer Workshop 2007 erstellt. BART unternimmt automatische Koreferenzresolution mithilfe einer modularen Pipeline, die aus einer Vorverarbeitungsphase (Daten von MMAX2-Annotationsebenen werden aggregiert), der Extraktion der NP-Kandidaten, der Extraktion der NP-Merkmale und der Kandidatenpaare sowie aus einem Resolutionsmodell besteht. BART verwendet momentan einen auf einem Ansatz von Soon basierenden Resolutionsalgorithmus, der Kandidaten-NPs hinsichtlich ihrer Merkmale paarweise vergleicht. Statt diesem soll nun das Resolutionssystem der Stanford-NLP-Gruppe (im Folgenden Stanford-System) implementiert werden, das sich durch seine Sieb-Architektur auszeichnet. Obwohl es hauptsächlich auf Regeln basiert, konnte es dennoch das beste Ergebnis beim CoNLL-2011 shared task erzielen. Im Rahmen der Sieb-Architektur werden nacheinander - absteigend nach ihrer Präzision geordnet - eine Reihe von deterministischen Koreferenzmodellen angewendet, wobei jedes Modell auf den Output seines Vorgängers aufbaut. Besonders das Entität-zentrische Modell, in bei dem Merkmale über alle Vorkommen einer Entität geteilt werden, bietet einen deutlichen Wissensgewinn, der von Nutzen für BARTs Performanz sein wird.

2 Ziel

Das Ziel besteht darin, die deterministischen Koreferenzmodelle des Stanford-Systems in BART für das Deutsche zu implementieren. Falls nach Erreichen dieses Ziels noch Zeit bleibt, wäre eine Erweiterung auf das Englische und Italienische, sowie die Implementierung weiterer Regeln vorstellbar. Idealerweise soll ein (wie BART) grundsätzlich sprachunabhängiges System entwickelt werden, das durch sprachspezifische Spezifikationen modifiziert wird.

3 Lösungsansatz

Der Lösungsansatz basiert auf den Koreferenzmodellen (Sieben) des Stanford-Systems.

- **Speaker Identification:** Es werden Sprecher identifiziert und mit möglichen koreferenten Pronomen verbunden.
- **String Match:** Zwei Entitäten werden als koreferent angesehen, wenn sie denselben Text(umfang) haben, einschließlich ihrer Attribute und Artikel.
- **Relaxed String Match:** Zwei nominale Entitäten sind koreferent, wenn ihre Köpfe gleich sind.
- **Precise Constructs:** Zwei Entitäten sind koreferent, wenn sie gemeinsam in einer Appositions- oder Subjekt-Objekt-Konstruktion stehen. Wenn die Entität ein zum Kopf des Antezedens gehöriges Relativpronomen ist, ein Akronym oder ein Demonym ist.
- **Strict Head Match A, Strict Head Match B, Strict Head Match C, Proper Head Noun Match, Relaxed Head Match:** Diese Regeln bezeichnen Entitäten als koreferent, wenn sie denselben Kopf haben und bestimmte Bedingungen erfüllen.
- **Pronoun Match:** Pronominale Koreferenz besteht, wenn bestimmte Agreement-Bedingungen erfüllt sind. Z.B.: Numerus, Genus, Person, Belebtheit, Satzentfernung zwischen Pronomen und Antezedens ≤ 3

Zunächst konzentrieren wir uns dabei auf Stringmatching und Pronounmatching.

4 Tools

Das Projekt implementieren wir in das bereits bestehende BART , das bereits einige Tools enthält.

5 Daten

Folgende Datensätze werden zur Evaluation genutzt:

- Die TüBA-D/Z Baumbank für das Deutsche
- Die Penn Treebank für das Englische
- Der iCab Korpus für das Italienische

Es ist vorgesehen auf dem Deutschen basierend zu entwickeln und das System später für Englisch und Italienisch zu erweitern.

References

[1] S. Broscheit, M. Poesio, S. P. Ponzetto, K. J. Rodriguez, L. Romano, O. Uryupina, Y. Versley, and R. Zanoli. Bart: A multilingual anaphora resolution system. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 104–107, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[2] H. Lee, A. Chang, Y. Peirsman, N. Chambers, M. Surdeanu, and D. Jurafsky. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Comput. Linguist.*, 39(4):885–916, Dec. 2013.

[3] Y. Versley, S. P. Ponzetto, M. Poesio, V. Eidelman, A. Jern, J. Smith, X. Yang, and A. Moschitti. Bart: A modular toolkit for coreference resolution. In *Proceedings of the ACL-08: HLT Demo Session*, pages 9–12, Columbus, Ohio, June 2008. Association for Computational Linguistics.