Exposé zur Diplomarbeit

Word Sense Disambiguation

Torsten Schiemann*

17. Mai 2006
Betreuung: Prof. Dr. Ulf Leser und Jörg Hakenberg
Zeitraum: Mai - Oktober

1 Motivation

Das System Ali Baba [1] durchsucht Texte aus den Lebenswissenschaften nach biologischen Entitäten (Objekten aus den Klassen Proteine/Gene, Spezies, Krankheiten, usw.), stellt sie in einem Interaktionsnetzwerk zusammen und visualisiert dieses als Graphen. *Entitäten* und ihre Beziehungen werden bei der *Named Entity Recognition* (NER) mit Hilfe von klassenspezifischen Wortlisten erkannt. Eine bestimmte Bedeutung eines Wortes wird Instanz genannt und kann auch mehrere Synonyme und somit mehrere Vorkommen in der Wortliste haben. Einige Wörter, wie Homonyme oder Abkürzungen, können für sich genommen nicht eindeutig einer biologischen Klasse oder Instanz einer Klasse zugeordnet werden. Zum Beispiel kann "this" ein Organismus (Fliege) oder ein Protein (UniProt-ID: Q65L95_BACLD) sein. Es kann aber auch eine Instanz aus dem üblichen englischen Sprachgebrauch sein – this als Fürwort. Dies macht bei der automatischen Wissensextraktion aus Texten große Probleme. Oft werden Wörter einer falschen biologischen oder nicht-biologischen Wortklasse zugeordnet. *Word Sense Disambiguation* beschäftigt sich mit der Auflösung solcher Mehrdeutigkeiten.

Bei der in Ali Baba verwendeten NER werden Wörter auch dann erkannt, wenn ihre Schreibweise leicht von dem zugrunde liegenden Standard abweicht. Dementsprechend ist die Wahrscheinlichkeit höher, dass Wörter in mehreren Klassen zu finden sind. Erschwerend kommt hinzu, dass Wörter mehrere Bedeutungen innerhalb einer Klasse haben können, insbesondere dann, wenn Abkürzungen verwendet werden [3].

Die für die NER benötigten Daten werden von unterschiedlichen Institutionen erstellt und gepflegt. Zum Beispiel sind die 199.693 Wörter einer Protein-Wortliste aus der UniProt-Datenbank herausgefiltert worden [7]. Andere Datenbanken für z. B. Krankheiten oder Medikamente werden bei der National Library of Medicine [6] seit Jahrzehnten gepflegt. Einige Einträge dieser Listen können auch aus zusammengesetzte Wörter bestehen. Diese müssen nur selten disambiguiert werden, da sie die Instanz sehr gut beschreiben. Tabelle 1 stellt die Häufigkeit der doppelt belegten Wörter[1] im Vergleich von jeweils zwei Thesauri dar.

	Gewebe	Spezies	Medika.	Krankheiten	Zellen	Übliche W.[2]	#Einträge
Proteine	1	301	89	16	0	1423	199693
Gewebe	–	4	4	34	326	117	1582
Spezies	4	–	10	34	0	352	79073
Medikamente	4	10	–	6	2	200	12390
Krankheiten	34	34	6	–	8	322	12401
Zellen	326	0	2	8	–	88	708

Tabelle 1: Anzahl der kritischen Wörter bei Vergleich von zwei Wortklassen und Anzahl der Einträge einer biologischen Wortliste. Zum Beispiel gibt es 1423 englische Wörter, die ebenfalls Proteine/Gene bezeichnen können.

*schiemann@informatik.hu-berlin.de
[1] Jeder Eintrag mit einzelnen oder zusammengesetzten Wörtern wird verglichen.
[2] 10.000 häufigste englische Wörter

Die Disambiguierung eines Wortes gilt als schwierig und wird als *AI-complete* beschrieben [4]. Da die gesuchten Wortklassen bekannt sind und weitere Eigenschaften einer Instanz z. B. durch den Kontext bestimmbar sind, ist eine Disambiguierung trotzdem mit hoher Güte möglich. Dies gilt insbesondere dann, wenn ein im Dokument mehrmals vorkommendes Wort nur mit einer Bedeutung versehen ist („one sense per discourse“), was meistens zutrifft [2].

2 Zielsetzung

Das Ziel dieser Diplomarbeit ist es, ein Verfahren für die Bestimmung der korrekten Wortklasse mit Hilfe des Kontextes im Dokument zu entwickeln. Wie oben beschrieben, kommen vermehrt auch Wörter vor, die innerhalb einer Klasse unterschiedliche Bedeutungen haben. Auch hier soll, wann immer möglich, eine Disambiguierung dieser Wörter erfolgen. Diese Verfahren sollen in Ali Baba zur Anwendung kommen.

Zunächst sollen die Wörter ermittelt werden, die in Ali Baba kritische Fälle darstellen. Dies sind Wörter aus den Kategorien Proteine/Gene, Gewebe, Krankheiten, Medikamente, Zellen und Organismen. Für jedes dieser Wörter soll ein Klassifikator trainiert werden, der Klassenzugehörigkeiten bestimmt.

3 Vorgehen

Das Problem, dem wir uns zu Beginn stellen müssen, ist es, herauszufinden, welche Wörter überhaupt eine Disambiguierung notwendig machen. Sicher ist nicht für jedes Wort in einem Text eine Disambiguierung sinnvoll, sondern nur in relativ wenigen Fällen. Hierzu vergleichen wir alle Wörter aller Wortklassen, um Homonyme und Begriffe mit ähnlicher Schreibweise (Großschreibung, Singular/Plural) zu erhalten. Für jede Instanz eines kritischen Wortes müssen Trainingsdaten (Dokumente) gesucht werden. Z. B. können aus der UniProt-Datenbank Dokumente (Abstracts) zu einzelnen Proteinen geladen werden. Das soll soweit wie möglich automatisiert werden. Zudem können solche beschreibenden Elementen auch aus der jeweiligen Datenbank (MeSH: "ScapeNote", UniProt: "Function", "Submit", "Subcelluar Location", "Induction") gezogen werden.

Es muss ein geeignetes Klassifikationsmodell entwickelt werden. Dazu gehört die Auswahl der relevanten Merkmale (prägnante Wörter im Dokument) und die Zusammenstellung der Eigenschaften einer Instanz zu Merkmalsvektoren. Diese prägnanten Wörter sollen die Instanzen eines zu klassifizierenden Falls am besten von den anderen Instanzen abgrenzen. Dazu gibt es Verfahren, die insbesondere die Frequenz der Kontext-Wörter berücksichtigt.

Als Klassifikator soll die mittlerweile etablierte *Support Vector Machine* (SVM) mit dem *One-Against-All*-Ansatz, für die Klassifikation von mehr als 2 Klassen, zur Anwendung kommen [5]. Im Rahmen dieser Diplomarbeit soll auch ein Vergleich der Verfahren zur Auswahl der Wörter und eine Erläuterung der Möglichkeiten zur Abbildung der Eigenschaften auf Merkmalsvektoren vorgenommen werden.

Zum Schluss muss die Güte des Verfahrens ermittelt werden. Um eine möglichst genaue Annäherung an die wahren Evaluationsparameter zu erhalten, werden die Trainingsdaten mehrmals in Trainingsdaten und Testdaten aufgeteilt und die so erhaltenden Schätzer gemittelt (*n-Fold Cross Validation*).

Literatur

[1] Ali Baba, 2006. http://wbi.informatik.hu-berlin.de/alibaba/.

[2] W. Gale, K. Church, and D. Yarowsky. One sense per discourse. *In Proceedings of the DARPA Speech and Natural Language Workshop*, pages 233-237, 1992.

[3] S. Gaudan, H. Kirsch, and D. Rebholz-Schuhmann. Resolving abbreviations to their senses in Medline. *Bioinformatics*, 21(18):3658–3664, 2005.

[4] N. Ide and J. Veronis. Word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1):1–40, 1998.

[5] T. Joachims. A statistical learning learning model of text classification for support vector machines. In *SIGIR 2001: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 128–136, New York, NY, USA, 2001. ACM Press.

[6] National Library of Medicine, 2006. http://www.nlm.nih.gov/.

[7] UniProt, 2006. http://www.ebi.uniprot.org/.