Bestimmung von Text-Pattern für die Informationsextraktion

Exposee einer Diplomarbeit

Betreuer: Ulf Leser, Jörg Hakenberg
Bearbeiter: Conrad Plake

Oktober 2004 - März 2005

Motivation

Die Extraktion von Informationen über Entitäten und ihren Relationen aus einem natürlichsprachigen Text stellt viele Anforderungen an ein Textmining-System und gehört heute zu den schwierigen Aufgaben, die beim Wissensmanagement zu lösen sind. Im Bereich der Bioinformatik ist speziell die Extraktion von Beziehungen zwischen Proteinen (Protein-Protein-Interaktion) ein aktuell intensiv untersuchtes Gebiet.

Viele Extraktionsverfahren basieren auf Pattern-Matching (regelbasiert) oder komplexem Grammar-Parsing. Die Pattern werden manuell erstellt, können aber auch anhand von Trainingsbeispielen erlernt werden [4, 9]. Parser mit einer kontextfreien Grammatik arbeiten nicht-deterministisch, was dazu führt, dass viele mögliche syntaktische Ableitungen zu interpretieren sind [5]. Regelbasierte Systeme, die wenige Pattern verwenden, arbeiten sehr schnell, jedoch oft mit einer geringeren Genauigkeit. Beide Methoden erfordern eine Vorverarbeitung des Textes, um darin enthaltene Wörter zu annotieren (Part-Of-Speech (POS), Named-Entity-Recognition (NER)).

Die manuelle Bestimmung von Extraktionsregeln ist ein zeitaufwändiger Prozeß und viele Regeln sind nötig um eine akzeptable Genauigkeit zu erzielen. Typischerweise werden diese Regeln nach heuristischen Kriterien erstellt und sind kaum auf andere Domänen übertragbar.

Ziel

Für die textbasierte Informationsxtraktion sollen Pattern nach bestimmten Gütekriterien aus einem annotierten Korpus automatisch gewonnen werden. In der Arbeit werden zwei generische Verfahren implementiert und am Beispiel von Protein-Protein-Interaktionen (PPI) getestet.

Vorgehensweise

In dem ersten Verfahren betrachten wir Text-Pattern als reguläre Ausdrücke. Mittels eines Genetischen Algorithmus (GA) sollen DFAs gefunden werden, die nach Gütekriterien wie Precision, Recall oder F-measure optimiert sind. Die Interpretation der selektierten Phrasen (z. B. als PPI) wird durch den GA anhand des Trainigskorpus erlernt. Neben den Parametern eines DFA (Alphabet, Zustände, Transitionen) sind auch Strategien/Heuristiken für die Entwicklung möglichst weniger Automaten zur Erreichung einer gewählten Güte innerhalb des Trainingskorpus zu bestimmen.

Die zweite Methode basiert auf einem Clustering-Verfahren. Als Objekte dienen die POS-annotierten Wortsequenzen zwischen den relevanten Entitäten aus dem Trainingskorpus. Die Ähnlichkeit zwischen zwei Objekten soll von syntaktischen Eigenschaften abhängig sein und wird aus ihrem Alignment und einer Gewichtsfunktion berechnet. Durch eine Menge von Medoiden ist genau ein Clustering dadurch gegeben, dass jedes Objekt dem Medoid zugeordnet wird, zu dem es die größte Ähnlichkeit hat. Die Medoide und ihre Anzahl sind hinsichtlich oben genannter Gütekriterien aus der Menge aller Trainingsobjekte zu bestimmen. Neue Wortsequenzen werden dem ähnlichsten Medoid zugeordnet und können so mit zusätzlichen Informationen über die Objekte in einem Cluster (z.B. Typ und Richtung einer PPI) annotiert werden.

Nach [6] ist die Partitionierung des Textes in seine einzelnen Sätze ein vielversprechender Ansatz für die Erkennung von PPI und wird darum auch in dieser Arbeit verfolgt. Die Tokens in einem Satz sind zumindest mit den Entitätsklassen "Protein" und "Nicht-Protein" annotiert. Die Probleme der NER spielen bei der Informationsextraktion aus Freitexten eine wesentliche Rolle, sollen hier aber nicht im Vordergrund stehen.

Für die Evaluation stehen zwei Korpora zur Verfügung. Der BioCreative-Korpus [3] enthält 15000 Sätze, in denen alle Proteinbezeichner annotiert sind. Aus dieser Menge wurden 1003 Sätze zufällig herausgezogen und manuell auf PPI geprüft [7]. Der IEPA-Korpus [1, 6] umfasst 498 Sätze, in denen ebenfalls die Proteine und Interaktionen bekannt sind. Anhand der

gegebenen Annotationen und mit Hilfe eines Tools für die Erkennung von Proteinbezeichnern [2] kann die Evaluation mit und ohne NER-Problem für Proteine auf den Sätzen crossvalidiert durchgeführt werden. In einer Studienarbeit [8] wurden zudem Artikel gesammelt, die zusammen 297 zufällige Hefe-PPI aus der DIP [10] enthalten. Diese gilt es in 165 Abstracts und Volltexten zu finden.

Literatur

[1] D. Berleant, J. Ding, and A.W. Fulmer. Corpus Properties of Protein Interaction Descriptions in MEDLINE. Unpublished, 2003.

[2] S. Bickel, U. Brefeld, L. Faulstich, J. Hakenberg, U. Leser, C. Plake, and T. Scheffer. A Support Vector Classifier for Gene Name Recognition. In *EMBO Workshop: BioCreAtIvE*, 2004.

[3] BioLINK. BioCreative: Critical Assessment of Information Extraction Systems in Biology. *http://www.mitre.org/public/biocreative/*, 2003.

[4] C. Blaschke and A. Valencia. Can bibliographic pointers for known biological data be found automatically? Protein interactions as a case study. *Comparative and Functional Genomics*, pages 196–206, 2001.

[5] N. Daraselia, A. Yurgev, S. Egorov, S. Novichkova, A. Nikitin, and I. Mazo. Extracting human protein interactions from MEDLINE using a full-sentence parser. *Bioinformatics*, pages 604–611, 2004.

[6] J. Ding, D. Berleant, D. Nettleton, and E. Wurtele. Mining MEDLINE: Abstracts, Sentences or Phrases? *Pacific Symposium on Biocomputing 7 (PSB)*, pages 326–337, 2002.

[7] J. Hakenberg, C. Plake, and U. Leser. Optimizing Syntax Patterns for Discovering Protein-Protein Interactions. Submitted, 2004.

[8] C. Plake. Extraktion von Protein-Protein-Interaktionen aus Freitexten. Studienarbeit, HU Berlin, 2004.

[9] S. Soderland. Learning Information Extraction Rules for Semi-Structured and Free Text. *Machine Learning*, 34(1-3):233–272, 1999.

[10] I. Xenarios, D.W. Rice, L. Salwinski, M.K. Baron, E.M. Marcotte, and D. Eisenberg. DIP: the Database of Interacting Proteins. *Nucleic Acids Res*, 28:289–291, 2000.