Ruprecht-Karls-Universität Heidelberg
Institut für Informatik
Sommersemester 2010
Seminar: Text Mining
Dozenten: Jannik Strötgen
Prof. Dr. Michael Gertz

Seminararbeit

Coreference of Named Entities

Name:	Philipp Schäfer, Daniel Kruck
Matrikelnummer:	2612579 (Philipp), 2440234 (Daniel)
Studiengang:	Angewandte Informatik (6. Semester)
Email:	trashzopf@googlemail.com (Philipp), daniel.kruck@gmx.net (Daniel)
Datum der Abgabe:	25. Juli 2010

Hiermit versichere ich **Philipp Schäfer, Daniel Kruck**, dass ich die Hausarbeit mit dem Titel **Coreference of Named Entities** im Seminar **Text Mining** im **Sommersemester 2010** bei **Jannik Strötgen** und **Prof. Dr. Michael Gertz** selbstständig und nur mit den in der Arbeit angegebenen Hilfsmitteln verfasst habe. Zitate sowie der Gebrauch fremder Quellen, Texte und Hilfsmittel habe ich nach den Regeln wissenschaftlicher Praxis eindeutig gekennzeichnet. Mir ist bewusst, dass ich fremde Texte und Textpassagen nicht als meine eigenen ausgeben darf und dass ein Verstoß gegen diese Grundregel des wissenschaftlichen Arbeitens als Täuschungs- und Betrugsversuch gilt, der entsprechende Konsequenzen nach sich zieht. Diese bestehen in der Bewertung der Prüfungsleistung mit "nicht ausreichend"(5,0) sowie ggf. weiteren Maßnahmen.
Außerdem bestätige ich, dass diese Arbeit in gleicher oder ähnlicher Form noch in keinem anderen Seminar vorgelegt wurde.

Heidelberg, den 25. Juli 2010 _____

Inhaltsverzeichnis

1 Einleitung

In letzter Zeit hat sich das Sortier- und Suchverhalten der Menschheit geändert. Sortierte man früher noch seine Dokumente in Ordner, ist man heute glücklich, wenn man mit leistungsstarken Suchalgorithmen schnell und präzise das gewünschte Dokument findet.
Dabei beschränken sich Suchanfragen nicht nur auf lokale Daten, sondern werden sogar großteils ans Web gestellt. Eine häufige Anfrageform an Suchmaschinen sind hierbei die *named entity queries*[1].

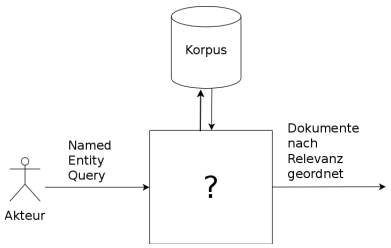

Tabelle 1: Häufige Suchanfrage an Suchmaschinen

Ein Problem bei solchen Suchanfragen ist es, die Häufigkeit der **named entity** in Dokumenten richtig zu erfassen[3]. Denn die Referenzierung des Objektes, das sich hinter einem Eigennamen verbirgt, wird sowohl mit dem Eigennamen selbst, als auch mit kompakteren Ausdrücken vorgenommen.
So wird beispielsweise in einem Text, der von Peter Chen handelt, die Person *Peter Chen* auch mit *„er“* referenziert.

> „**Chen** has recieved serveral awards in the fields of Information Technology. **He** received the Data Resource Management Technology Award [. . .]“ [2]

In der folgenden Hausarbeit wird ein statistisches Verfahren erklärt, welches die Häufigkeit von Eigennamen in Dokumenten schätzen soll. Grundlage für diese Ausarbeitung ist ein Paper der Herren Na und Ng über „A 2-Poisson Model for Probabilistic Coreference of Named Entities for Improved Text Retrieval“[3].

2 CEEF

2.1 Plausible Annahme über die Verteilung von Koreferenzen

Die Häufigkeit von Eigennamen in Dokumenten ist schwer zu erfassen. Zählt man nur die *named entites* selbst, erhält man die raw entity frequency:

$$tf\,(e;d) = \sharp(\text{named entity})$$

Die raw entity frequency beachtet noch keine zusätzlichen Koreferenzen auf das Objekt hinter dem gesuchten Eigennamen. Um später ein besseres Ranking ausführen zu können, möchten wir möglichst alle Referenzen erfassen. Also addieren wir die Häufigkeit der Koreferenzen zu unserer raw entity frequency:

$$tf\,(e;d) \leq tf_{true}\,(e;d) = tf\,(e;d) + \underbrace{atf(e_Q;A,d)}_{Korefernzen}$$

Theoretisch sind wir nun am Ziel. Praktisch ist es aber nicht so einfach, die Häufigkeit der Koreferenzen zu bestimmen. Mit Programmen, die Koreferenzen vollständig auflösen, geht ein überdimensionaler Berechnungsaufwand einher. Zudem sind diese Programme noch nicht besonders Präzise und ordnen weniger als 70% der Koreferenzen richtig zu.[3]
Seung-Hoon Na und Hwee Tou Ng untersuchten deswegen einen statistischen Ansatz zur Schätzung der Koreferenzen. Dafür nehmen sie folgendes an:

> "Our key assumption is that the frequency of anaphoric expressions is distributed over named entities in a document according to the probabilities of whether the document is elite for the named entities."[3]

2.2 Formalisierung

Bevor wir zur eigentlichen Statistik kommen, legen wir noch die mathematische Schreibweise für diese Aufgabe fest. Zunächst sei Q unsere query.

- Q = query (Suche)
- e_Q = query entity (gesuchte Entität)
- e_N = non-query entity

- A = Menge plausibler anaphorischer Ausdrücke
- $tf(A; d)$ = Anzahl von A in Dokument d
- $\varepsilon(A; d)$ = Menge plausibler Entitäten in Dokument d
- $tf(e; d)$ = raw entity frequency
- $atf(e; A, d)$ = anaphoric entity frequency

2.3 Elitness

2.4 Poisson

Literatur

[1] Guha, R. und A. Garg. Disambiguating people in search. *In TAP: Building the Semantic Web*, 2003.

[2] Modified: 21:02, 21 May 2010 Nineball. Peter chen, 2010.

[3] Seung-Hoon Na, Hwee Tou Ng. A 2-poisson model for probabilistic coreference of named entities for improved text retrieval. *Annual ACM Conference on Research and Development in Information Retrieval*, 2009.