Universität zu Köln
Sprachliche Informationsverarbeitung
Hauptseminar: Angewandte linguistische Datenverarbeitung
Prof. Dr. Jürgen Rolshoven
Hausarbeit

Semantische Spezifität im Word Space Model

Von C. Friedrich

(Vorgelegt am 30. August 2016)

1 Einleitung

Semantische Spezifität

Jones (1972, 11) Beschreibt die Spezifität eines Begriffes so:

> Specificity ... is a semantic property of index terms: a term is more or less specific as its meaning is more or less detailed and precise.

Semantische Spezifität und Anglizismen

2 Ein Maß für die semantische Spezifität

2.1 Textgrundlage

2.2 Document Frequency

Bereits Jones (1972) schlug ein statistisches Maß für die semantische Spezifität eines Wortes vor. Es ist die simple Frequenz, mit der ein Wort im Korpus auftaucht, das ein Indiz für die Spezifität darstellen soll. Caraballo und Charniak (1999) konnten die *document frequency* als Eigenschaft von Worten dazu nutzen, für beliebige Wortpaare festzustellen, welches Wort spezifischer oder genereller ist. Überprüft wurde das mit Beispielwortpaaren, die in einer Hyperonym- bzw. Hyponymrelation zueinander stehen: Das Wort *Getränk* ist ein Oberbegriff zum Wort *Cola*. Für diese Art von Relation gilt: Wenn ein Wort ein Oberbegriff eines anderen ist, so ist der Unterbegriff semantisch spezifischer als der Oberbegriff. Die klar unterschiedene Spezifität ist also eine notwendige Bedingung für die Hyperonym- bzw. Hyponymrelation. Das macht solche Wortpaare zu natürlichen Kandidaten, um Maße für semantische Spezifität zu testen.

Insoweit es die Textgrundlage hergibt, also in Dokumenten geordnet ist, verwende ich die *document frequency* als erste Annäherung an ein brauchbares Maß für die semantische Spezifität. Die Berechnung ist simpel und performant. Entscheidend wird die Frage sein, ob es den komplexeren Modellen gelingt, eine höhere Erfolgsrate zu erzielen. Daher verwende ich die *document frequency* als Benchmark für die anderen Modelle.

Weil das spätere Ziel ist, Worte in unterschiedlichen Korpora miteinander zu vergleichen, muss die *document frequency* noch normiert werden, wodurch das Maß etwas unabhängiger vom verwendeten Korpus wird.

Definition. *Sei N die Gesamtzahl aller untersuchen Dokumente und df die Anzahl der Dokumente, in denen das Fokuswort auftritt. Dann ist die normierte document frequency df_n*

$$df_n = \frac{df}{N} \tag{1}$$

2.3 Satzkontexte vs. Kontextfenster

Grundlage dieser Arbeit ist das *Word Space Model* (WSM) oder auch Termvektormodell, das unter anderem sehr ausführlich in Sahlgren (2006) beschrieben wird. Das WSM

Paradigmatischer vs. Syntagmatischer Kontext

Zur Berechnung der verschiedenen Maße muss zunächst eine Word-Word Kookkurrenzmatrix erstellt werden

2.4 Größe des paradigmatischen Kontexts

2.5 Semantischen Nähe des Kontextes

2.6 Frequenz vs. Chi Squared vs. Dice Coefficient

2.7 Getestete Maße

2.8 Wordpaare

2.9 Resultate

3 Ein Modell für Anglizismen

3.1 Textgrundlage

3.2 Verwendete Technologien

3.3 Resultate

4 Konklusion

Literatur

Caraballo, Sharon A. und Charniak, Eugene: Determining the Specificity of Nouns From Text. In: In Proceedings SIGDAT-99. 1999, 63–70

Han, Jiawei, Kamber, Micheline und Pei, Jian: Data Mining: Concepts and Techniques. Morgan Kaufmann, 2011

Heyer, Gerhard, Quasthoff, Uwe und Wittig, Thomas: Text Mining: Wissensrohstoff Text – Konzepte, Algorithmen, Ergebnisse. 1. Auflage. W3L-Verlag, 2008

Jones, Karen Sparck: A Statistical Interpretation of Term Specificity and its Application in Retrieval. In: Journal of Documentation, 28 01 1972, Nr. 1, 11–21, ISSN 0022–0418

Manning, Christopher D. und Schütze, Hinrich: Foundations of Statistical Natural Language Processing. MIT Press, 1999

Sahlgren, Magnus: The Word-space model. Dissertation, University of Stockholm (Sweden), 2006