Salton und Wittgenstein in den Humanities: Über die Semantik in Philosophischen Texten

Marco Büchler, Gerhard Heyer
Natural Language Processing Group
Institute for Mathematics and Computer Science
Leipzig University, Germany
$[mbuechler|gheyer]@eaqua.net$

Abstract: In der Informatik wird die Semantik durch *diskriminierende Terme* beschrieben. Jedoch fehlen oftmals speziell in philosophischen Texten genau diese gewichtigen Terme. Ausgehend von der oft eingesetzten *diskriminierenden Semantik* wird am Problem der Sinn- und Weisheitssprüche eine *kontrastive Semantik* vorgestellt. Die eingeführte Methode stellt ein Lessons Learnt aus dem eAQUA-Projekt [BHG08, HBB+10] im Umgang mit antiken Texten dar.

1 Einführung

Dem Thema Semantik kann sich im Bereich der *Automatischen Sprachverarbeitung* auf verschiedenste Weise genähert werden. Aus der Sicht des *Information Retrieval* werden Suchmaschinen darauf optimiert, mittels möglichst weniger Eingabewörter ein relevantes Dokument zu finden (Semantik einer Textpassage). Dazu werden die Terme einer Textpassage gewichtet und repräsentieren somit den semantischen Raum. Auf der anderen Seite werden bspw. im *Text Mining* signifikante Assoziationen zwischen Wörtern berechnet (Semantik eines Wortes), die wiederum deren semantische Nutzung in einem Textkorpus wiedergeben.

Wird das Thema Semantik aus der Sicht des *Text Reuse* und *Knowledge Transfer* betrachtet, spielen beide Aspekte - Semantik einer Textpassage und Semantik eines Wortes - eine entscheidende Rolle. Im Kontext der eHumanities muss jedoch zwischen den *historischen* und *philosophischen Zitationsspuren* unterschieden werden. Während beim historischen Wissenstransfer oftmals eindeutig diskriminierende Terme wie *Orte*, *Personen* oder *Ereignisse* bestimmt werden können, ist das Vokabular der Philosophie sehr stark von Allgemeinsprache geprägt [Pie10], welches den Einsatz von semantischen *Reuse*-Verfahren deutlich erschwert. Dies kann am folgenden Spruch von William Shakespeare verdeutlicht werden.

To be, or not to be this is the question.
William Shakespeare in Hamlet

Auch wenn dieser Spruch von Shakespeare von vielen Menschen wiederverwendet wird,

ist es mit *Information-Retrieval*-Methoden sehr schwer eine semantische Repräsentation zu bestimmen, da er nahezu komplett aus Stoppwörtern besteht und dementsprechend keine oder nur schwach diskriminierende Terme enthält. Werden beispielsweise die Termgewichte nach dem $tf.idf$-Maß [SWY75] für diesen Spruch ausgerechnet, dann haben bis auf *question* alle Terme ein Gewicht von 0 (Stoppwörter). Des Weiteren liefern auch Verfahren wie die *Differenzanalyse* oder das *Log-Likelihood-Ratio* (beide siehe Abschnitt 2) im Vergleich zu einem Referenzkorpus keine ernsthafte semantische Repräsentation.

Ferner ist das Sprachvokabular oftmals so allgemein, dass sich philosophische Texte nur schwer vom sprachlichen Niveau eines Grundschülers unterscheiden. So entspricht das oben vorgestellte *Shakespeare*-Zitat nach dem *Dale Chall Readability Index* gerade einmal dem Sprachniveau eines Schülers der Klassenstufe 2 bis 3.

Da Methoden der *Semantik durch diskriminierende Terme* philosophische Texte nur sehr schwer beschreiben können, wird in diesem Papier ein einfaches Verfahren vorgestellt, welches *Semantik durch Kontrast* misst. Hierbei wird eine Textstelle nicht durch diskriminierende Terme beschrieben, sondern durch einen möglichst großen Kontrast mindestens zweier Wörter innerhalb dieser Textstelle.

Gerade in philosophischen Texten ist dies von größtem Interesse, da in ihnen oftmals Lebensweisheiten bzw. gesellschaftliche und soziale Wertungen enthalten sind, die aktiv von Wissenschaftlern aus den Geisteswissenschaften gesammelt wurden und immer noch werden. Ein Spezialfall dieser philosophischen Texte stellen die punktierten *Sinn- und Weisheitssprüche* (Gnomologien) dar [Pie10, VOG10, Rou10]. Speziell hierbei werden Konzepte in Relation zueinander gestellt, die sowohl nicht erwartet als auch oftmals semantisch kontrastive Terme enthalten. Vielmehr bilden sie Wissensmuster ab, die alltägliche Konzepte in nicht-alltäglichen Zusammenhängen miteinander verbinden, die wiederum durch einen Lerneffekt gewonnen worden sind.

2 State Of The Art

Aus der Sicht des semantischen *Text Reuse* und *Knowledge Transfers* gibt es zwei Sichten auf das Modellieren von Semantik:

- *Semantische Repräsentation von Textpassage*: Hierbei wird eine Textpassage durch ihre diskriminierenden Terme beschrieben und dementsprechend repräsentiert bzw. gewichtet. Während Salton's $tf.idf$ [SWY75] auf einem einzelnen Korpus angewendet wird, können Methoden wie die *Differenzanalyse* [Wit04] bzw. eine entsprechende Modifikation des *Log-Likelihood-Ratios* [Wit04, Dun93] eingesetzt werden, um entsprechende Abweichungen bzgl. eines Referenzkorpus zu messen. Neben den *probabilistischen* gibt es auch *linguistische* und *vorwissensbasierte Verfahren*, auf die an dieser Stelle aber nicht im Detail eingegangen werden soll.
- *Semantischer Kontext eines Wortes*: Unabhängig von den einzelnen Textpassagen kann die semantische Umgebung eines Wortes bestimmt werden, um dessen Bedeutung innerhalb eines Korpus zu berechnen [Wit53]. Als gängige Methode hat

sich diesbezüglich die *Kookkurrenzanalyse* bewährt [HQW08, Büc08, Büc05]. Dabei wird die Assoziationsstärke zwischen zwei Wörtern gemessen. Die Menge aller assoziationsstarken Terme zu einem Wort repräsentieren dessen semantische Bedeutung innerhalb eines Korpus, die wiederum dazu genutzt werden kann, um ähnlich benutzte Wörter zu bestimmen [Bor07].

Unabhängig von der semantische Repräsentation einer Textpassage oder eines Wortes wurden *Readability*-Tests eingeführt, um Texte nach ihrem sprachlichen Niveau zu bewerten. So werden in den USA Scores wie der *Dale Chall Readability Index* [Cen10b], *Coleman Liau Readability Index* [Cen10a] und der *Automated Readability Index* [SS67] eingesetzt, um Textdaten altersgerecht und dem sprachlichen Niveau entsprechend einer Altersstufe zuzuordnen [Cen10b]. Methodisch messen solche Verfahren letztendlich immer zwei Merkmale: Einerseits spielt die *Satzlänge* eine wichtige Rolle. Andererseits die Menge *inhaltsbehafteter Wörter*. Je nach Maß wird dies über die Wortlänge, die Anzahl der Silben oder die Anzahl der Wörter, die nicht zu den 3000 häufigsten Wörtern zählen, gemessen. All diese *Readability*-Maße haben gemeinsam, dass sie den Score erhöhen, sobald sich diskriminierende Terme häufen und die Sätze länger werden.

3 Methodologie

Im eingangs erwähnten Spruch von William Shakespeare sind weder die im Abschnitt 2 genannten diskriminierenden Terme enthalten, noch scheint der Spruch inhaltlich schwierig zu sein. So kann nach dem *Dale Chall Readability Index* [Cen10b] für diesen Spruch ein Score von $DCI = 4.1821$ berechnet werden. Dies entspricht nach internationalen Standards und der in den USA aktiv eingesetzten Klassifikation dem Sprachniveau eines Grundschülers der 2. - 3. Klasse. Da jedoch dieser Spruch allgemein bekannt und oft zitiert ist, stellt sich die Frage, ob semantische Relevanz immer durch diskriminierende Terme gemessen werden kann.

Speziell in philosophischen Texten können nur selten diskriminierende Terme aus dem *Iota*-Bereich der Wortverteilung (seltene Wörtern) genutzt werden. Vielmehr werden tendenziell häufige und allgemeine Terme der *Delta*- und *Zeta*-Bereiche (Details zu *Delta*-, *Zeta*- und *Iota*-Wörtern in [RE10]) beobachtet. *Burrow's Delta* [Bur07] wurde als Methode im Bereich des *Authorship Attribution* eingeführt. Hierbei werden speziell die Stoppwörter des *Delta*-Bereichs genutzt, um nach stilistischen Unterschieden zu suchen [Arg08]. Da der Schwerpunkt nicht auf stilistischen Merkmalen von Termen des *Delta*-Bereiches, sondern auf Semantik speziell des *Delta*- und *Iota*-Bereiches liegt, wird nachfolgend eine Adaption von *Burrow's Delta* genutzt, um nach möglichst großem semantischen Kontrast zweier Terme innerhalb einer Textstelle mit möglichst geringem Abstand im Text zu suchen.

Hierzu werden in einem ersten Schritt die semantischen Kookkurrenzen K zu jedem Wort bestimmt [Büc08, Büc05]. Als Signifikanz-Maß wurde das Log-Likelihood-Ratio sim_{lgl} mit einem Schwellwert von 6.63 und einer Mindestkookkurrenzfrequenz von 2 gewählt.

Basierend auf dem berechneten Kookkurrenzgraphen K werden anschließend die paar-

weisen Ähnlichkeiten zweier Wörter w_i und w_j mit dem Dice-Koeffizient berechnet.

$$sim_{dice}(w_i, w_j) = 2 * \frac{|K_{w_i} \cap K_{w_j}|}{|K_{w_i}| + |K_{w_j}|} \tag{1}$$

Hierbei entsprechen K_{w_i} und K_{w_j} den Kookkurrenzen der Wörter w_i und w_j. In diesem Schritt könnte auch ein Ähnlichkeitsmaß wie das *Cosinus*-Measure benutzt werden. Da dafür jedoch entsprechende Termgewichte nötig sind, ist im Rahmen dieses Papiers aus mehrfach genannten Gründen verzichtet worden.

Im Gegensatz zum Bestimmen von Wörtern mit ähnlichen Kontexten [Bor07] werden im zweiten Schritt genau die Wörter mit sehr ähnlichen Kontexten entfernt, um Kandidaten für Assoziationen zu bestimmen, die einen semantischen Kontrast repräsentieren.

$$contrast(w_i, w_j) = \begin{cases} 1 - sim_{dice}(w_i, w_j) & \text{if } sim_{dice}(w_i, w_j) \leq eps \\ 0 & \text{if } sim_{dice}(w_i, w_j) > eps \end{cases} \tag{2}$$

In der konkreten Anwendung hat sich im Altgriechischen ein $eps = [0.1, 0.15]$ als praktikabel herausgestellt.

Da die Menge der Assoziationen kontrastiver, unähnlicher bzw. unerwarteter Kookkurrenzprofile $C_{cand} = \cup_{i,j \in V} contrast(w_i, w_j)$ mit dem Vokabular V nicht zwangsweise auch im Text zusammen vorkommen, werden in einem dritten Schritt aus der Menge C_{Cand} diejenigen *kontrastiven Assoziationen* selektiert, die innerhalb eines Textfensters (hier Satz) auch zusammen auftreten. Dies entspricht dem Durchschnitt $C = K \cap C_{Cand}$ der beiden Menge K und C_{Cand} mit der zusätzlichen Bedingung.

$$dist(w_i, w_j) \leq eps_{dist} \text{ aus } (w_i, w_j) \in C \tag{3}$$

Hierbei entspricht die Restriktion $dist(w_i, w_j)$ dem Abstand der beiden Wörter im Text.

Visuell kann sich diese Methode wie folgt vorgestellt werden: Es sei angenommen, dass sich die Semantik eines Wortes durch Farben ausdrücken lässt. Dann bestimmt der Algorithmus in den beiden ersten Schritten einen möglichst großen farblichen Unterschied (Kontrast), der im dritten Schritt auf einen kleinen Raum beschränkt wird, um bspw. die Kante eines Objektes bzw. Gegenstandes auf einem Bild zu erkennen.

4 Ergebnisse, Lessons Learnt und Scope

Das im Rahmen dieses Papiers vorgestellte Verfahren misst semantischen Kontrast. In Anlehnung an das Beispiel aus dem Bereich des Image Mining gibt es mehrere Ergebnis-Cluster, die im Wesentlichen von der Textsorte abhängig sind wie zum Beispiel

- *Philosohie*: beispielsweise Gnomologien,
- *Komödie*: Sarkasmus und Zynismus,

- *Historie/Geschichtsschreibung*: unerwartete historische Zusammenhänge,
- *Sentiment Analysis*: Künstliche Doppeldeutigkeit, wobei nur die Doppeldeutung und nicht deren positive oder negative Wertung erkannt werden.

Die vorgestellte Methode ist genau genommen in vielerlei Hinsicht gegenteilig zu existierenden Ansätze. Auf der einen Seite entspricht der Algorithmus einer neuen Klasse von Verfahren. In der Automatischen Sprachverarbeitung werden Sprachmodelle benutzt, um basierend auf Trainingsdaten Vorhersagen zu machen, was als wahrscheinlichste Assoziation gilt. Sei es auf der syntaktischen Ebene bei den Markov-Ketten oder auch bei den Kookkurrenzen auf dem semantischen Level. Herkömmliche Sprachmodelle messen immer das Offensichtliche bzw. Wissen, das als gesichert angesehen werden kann. In den Humanities jedoch, ist dieses Wissen bekannt und kann nach jahrhundertelanger Forschung als gegeben angesehen werden. In einer solchen geisteswissenschaftlichen Anwendungen werden *latente Sprachmodelle* benötigt, um einen Mehrwert zu generieren. Andererseits werden in der Informatik Graph-Partitionierungsalgorithmen angewandt, um semantische Cluster zu bilden. Der in Abschnitt 3 vorgestellte Algorithmus bewertet jedoch genau diejenigen Kanten, die ein solcher Partitionierungsalgorithmus entfernt.

In Anlehnung an die zugrunde liegenden philosophischen Texte kann eine signifikante Überlappung zwischen den *kontrastiven Relationen* aus diesem Papier und dem *Text Reuse* bzw. *Knowledge Transfer* ausgemacht werden. In über 90% einer kontrastiven Relation wird auch ein *Text Reuse* gemessen. Diese Beobachtung ist insofern interessant, als dass dadurch erstmals nicht die Frage nach dem *Wie wird Text Reuse gemessen* im Mittelpunkt steht, sondern *Warum wird Text wiederverwendet*. Des Weiteren kann beobachtet werden, dass aufgrund ihrer philosophischen Reife (philosophisch gut überlegten Formulierungen) solche Weisheitssprüche sehr stark am Original wiederverwendet werden. Das Kernproblem der syntaktischen Verfahren ist jedoch, dass nicht zwischen einem statistisch signifikanten und häufig benutzten N-Gramm wie *im Namen unseres Herren Jesus Christus* und einem Zitat unterschieden werden kann. Mittels der *kontrastiven Relationen* kann zwischen allgemeinen Phrasen und potentiellen Zitaten eine Unterscheidung gemacht werden.

Im konkreten Beispiel sei auf das *Korpus der arabischen und syrischen Gnomologien* [Pie10] verwiesen. Aus den deutschen Übersetzungen der Weisheitssprüche ist die Abbildung 1 für das Wort *Körper* visualisiert worden. Es gibt insgesamt 8 verschiedene semantische Cluster dieses Wortes. In 5 dieser Cluster, die jeweils für einen Weisheitsspruch stehen, kann ein offensichtlicher Kontrast durch paarweise Antonyme wie z.B. *lebend* und *tot*, *besitzen* und *verteilen* oder auch *Stärke* und *Schwäche* erkannt werden.

5 Further Work

Für die Informatik hat sich zwischen den *kontrastiven Relationen* und dem *Text Reuse* ein unerwarteter Zusammenhang ergeben. Im Rahmen der weiteren Arbeiten im Bereich des *Text Reuse* und *Knowledge Transfers* wird die vorgestellte Methode verbessert. So kann

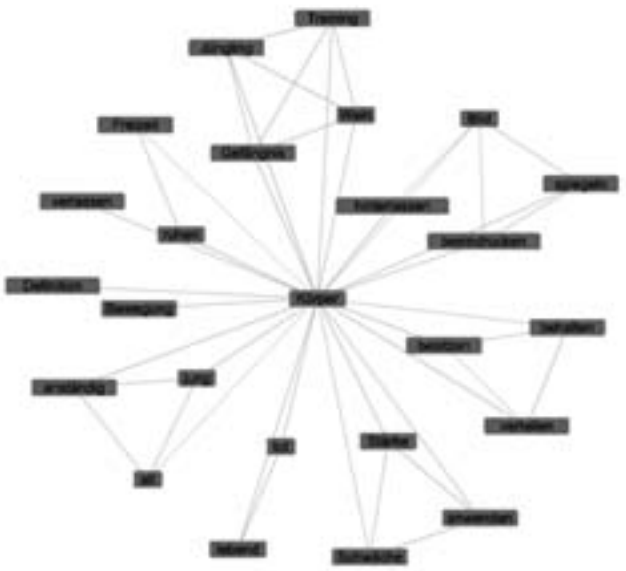

Abbildung 1: Die semantischen Cluster des Wortes *Körper* in den deutschen Übersetzungen aus dem *Korpus der arabischen und syrischen Gnomologien* [Pie10]. Jedes Cluster entspricht einem Weisheitsspruch. 5 der 8 Cluster beinhalten kontrastive Relationen wie *jung* und *alt*.

auf geshuffelten Texten gezeigt werden, dass das vorgestellte Verfahren durch einen Verzicht auf jeglichen probabilistischen Ansatz Schwächen hat. So kann der obere Grenzwert auch zufällig wie auf geshuffelten Texten sein. Daher wird bereits an einem Verfahren gearbeitet, welches einerseits die Kantengewichte nach dem *Log-Likelihood-Ratio* sowie die Topologie und der damit verbundene Dichte eines Wortes berücksichtigt. Hierbei werden diejenigen kontrastiven Assoziationen bevorzugt, die aus zwei sehr stabilen und gesicherten sowie möglichst unterschiedlichen Kontexten kommen.

Bei den *Readability*-Tests soll ein Text anhand der nötigen kognitiven Leistungsfähigkeit eines Menschen klassifiziert werden. Hierbei gibt es im Wesentlichen die Parameter der Satzlänge und die Menge an längeren Wörtern. Basierend auf diesen Kennzahlen werden die Texte klassifiziert. Jedoch muss angenommen werden, dass im Kontext eines Sprachmodells immer erwartbare bzw. leicht verständliche semantischen Dependenzen im Text vorkommen. So wird das eingangs erwähnte Zitat von Shakespeare auf das sprachliche Niveau eines Grundschülers eingestuft, der dieses Zitat sicher auch lesen aber jedoch wohl eher nicht verstehen kann. In diesem Sinne wird außerhalb der Arbeiten zum *Text Reuse* diese Methode zu einem *philosophischen* bzw. *semantischen Readability-Test* weiterentwickelt. Als konkrete Anwendung hierzu stehen aktive Forschungsarbeiten in eAQUA an, die sich mit der Frage nach dem *warum sind bestimmte Werke wichtig* beschäftigen und damit heutzutage noch erhalten, während andere nur noch in Fragmenten vorliegen.

6 Zusammenfassung

In diesem Papier wird eine neue Methode im Umgang mit Semantik beschrieben. Während Semantik bisher immer mit *diskriminierende Semantik* durch stark inhaltsbezogene Features beschrieben wird, ist ein Verfahren vorgestellt worden, das *kontrastive Semantik* durch semantische Differenz formuliert. Hierbei liegt der Arbeitsschwerpunkt auf philosophischen Texten mit der Spezialisierung auf den Sinn- und Weisheitssprüchen, die durch *diskriminierende Semantik* aufgrund eines eher allgemein bekannten Vokabulars eher mäßig beschrieben werden können. Ferner wurde skizziert, dass die kontrastiven Relationen gute Indikatoren für den Arbeitsbereich des *Text Reuse* sind, da sie nicht das *Wie*, sondern das *Warum* messen. Hierbei wird der Mehrwert durch semantischen Kontrast gemessen.

Literatur

[Arg08] Shlomo Argamon. Interpreting Burrows's Delta: Geometric and Probabilistic Foundations. *Lit Linguist Computing*, Seite fqn003, 2008.

[BHG08] M. Büchler, G. Heyer und S. Gründer. Bringing Modern Text Mining Approaches to Two Thousand Years Old Ancient Texts. In *e-Humanities – an emerging discipline: Workshop in the 4th IEEE International Conference on e-Science*, 2008.

[Bor07] S. Bordag. *Elements of Knowledge-free and Unsupervised Lexical Acquisition*. Dissertation, Universität Leipzig, 2007.

[Büc05] M. Büchler. Medusa Release Homepage, 2005. URL: http://mbuechler.eaqua.net/medusa/ last accessed Feb. 14th, 2010.

[Büc08] M. Büchler. *Medusa: Performante Textstatistiken auf großen Textmengen: Kookkurrenzanalyse in Theorie und Anwendung*. Vdm Verlag Dr. Müller, 2008.

[Bur07] John Burrows. All the Way Through: Testing for Authorship in Different Frequency Strata. *Lit Linguist Computing*, 22(1):27–47, 2007.

[Cen10a] RFP Evaluation Centers. Coleman-Liau Grade Level Readability Score, reading scores, 2010. URL: http://rfptemplates.technologyevaluation.com/readability-scores/coleman-liau-readability-score.html last accessed Jul. 21th, 2010.

[Cen10b] RFP Evaluation Centers. Dale-Chall 3000 Simple Word List, Readability Grade Score, 2010. URL: http://rfptemplates.technologyevaluation.com/dale-chall-list-of-3000-simple-words.html last accessed Jul. 21th, 2010.

[Dun93] T.E. Dunning. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1):61–74, 1993.

[HBB+10] G. Heyer, M. Büchler, V. Boehlke, C. Utschig-Utschig und C. Schubert. Aspects of an Infrastructure for eHumanities. In *Journal of Computing and Cultural Heritage, http://jocch.acm.org/)*, 2010.

[HQW08] Gerhard Heyer, Uwe Quasthoff und Thomas Wittig. *Text Mining: Wissensrohstoff Text – Konzepte, Algorithmen, Ergebnisse*. W3L-Verlag, 2008.

[Pie10] Ute Pietruschka. Corpus der arabischen und syrischen Gnomologien, 2010.

[RE10] Jan Rybicki und Maciej Eder. Deeper Delta Across Genres and Languages: Do We Really Need the Most Frequent Words? *Digital Humanities 2010 - Conference Abstracts. Centre for Computing in the Humanities, King's College London*, 2010. URL: http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/pdf/book-final.pdf last accessed Jul. 21th, 2010.

[Rou10] Charlotte Roueché. Sharing Ancient Wisdoms, 2010.

[SS67] E. A. Smith und R. J. Senter. *Automated Readability Index (ARI)*. Wright-Patterson AFB, OH: Aerospace Medical Division. AMRL-TR, 66–22, 1967.

[SWY75] G. Salton, A. Wong und C.S. Yang. A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[VOG10] Marie-Christine Bornes Varol, Marie-Sol Ortola und Jean-Daniel Gronoff. Aliento project - Intercultural Analysis of Sapiential statements and Transmission, 2010. URL: http://www.aliento.eu/en/node/63 last accessed Jul. 21th, 2010.

[Wit53] Ludwig Wittgenstein. *Philosophical Investigations*. Blackwell, Oxford, 1953. Translated by G.E.M. Anscombe.

[Wit04] F. Witschel. Text, Wörter, Morpheme - Möglichkeiten einer Terminologie-Extraktion. Diplomarbeit, Universität Leipzig, 2004.