Statistische Verfahren zur Ermittlung semantischer Nähe

André HALAMA
Sprachwissenschaftliches Institut, Ruhr-Universität Bochum
44780 Bochum, Germany
halama@linguistics.ruhr-uni-bochum.de

Abstract

Wir stellen ein Verfahren vor, das vermittels Methoden der analytischen Statistik Vorhersagen über die semantische Nähe eines substituierenden Worts zum Originalwort in Bigrammkontexten zulässt. Dabei soll berücksichtigt werden, ob es sich bei diesem Bigramm um ein Idiom, einen semi-kompositionellen oder einen frei kombinierbaren Ausdruck handelt.

Einleitung

Aus der Vielzahl der vorgeschlagenen Definitionen einer Kollokation hat sich in der Literatur diejenige durchgesetzt, die diesen Begriff mit dem des Idioms gleichsetzt. So findet man in Manning/Schütze (2000:184f) zur Klassifikation derartiger Sprachdaten die Kriterien, dass sich die Bedeutung einer Kollokation nicht aus der Verbindung ihrer Bestandteile ergibt (Freges Kompositionalitätsprinzip), dass man Bestandteile einer Kollokation nicht durch Wörter mit ähnlicher Bedeutung substituieren kann, und dass Kollokationen nicht durch zusätzliches lexikalisches Material oder grammatische Transformationen verändert werden können. Als solcher ist der Begriff der Kollokation auf die Definition grammatisch gebundener Elemente, die in einer bestimmten linearen Abfolge auftreten, beschränkt, und von allgemeineren Sprachphänomena wie Assoziationen zwischen Wörtern, die demselben Wortfeld angehören oder der reinen Kookurenz von Wörtern in einem gemeinsamen Kontext zu trennen.
Betrachtet man allerdings z.B. Sprachdaten des Deutschen wie die Kollokation des *reinen Weins* oder des *steifen Grogs*, kann die These der Nicht-Substituierbarkeit nicht uneingeschränkt aufrechterhalten werden. So lässt sich z.B. *rein* durch *klar* ersetzen[1] und *steif* durch *stark*, ohne dass sich die Bedeutung im Kontext z.B. der Redewendung *jemandem reinen Wein einschenken* verändern würde. Vollständig idiomatischen Charakter besitzt hingegen der *kalte Kaffee*, dessen Adjektiv sich nicht ohne Verlust der Bedeutung des Erfrischungsgetränks aus Cola und Limonade substituieren lässt, während das Attribut des *kalten Biers* durchaus gegen Synonyme ausgetauscht werden kann.
Dennoch sind die Substitutionsmöglichkeiten in den beiden erstgenannten Fällen stark eingegrenzt: *rein* lässt sich in einem Kontext mit Wein ebensowenig durch mögliche Synonyme wie *transparent*, *pur*, *einfach* oder *keusch* ersetzen, wie es Kandidaten für *steif*, wie z.B. *unbiegsam*, *plump* oder gar *erigiert*, in der Umgebung von *Grog* könnten.
Aus diesen Beobachtungen lässt sich folgende Fragestellung formulieren: Wenn Idiome starken syntagmatischen und paradigmatischen Kombinationsbeschränkungen unterliegen, zu welchem Grad können dann in semi-kompositionellen und frei kombinierbaren Ausdrücken Bestandteile alternieren? Oder anders ausgedrückt, in welchem Verhältnis stehen die substituierenden Elemente zu dem substituierten Wort und dem unmittelbaren Kontext, in dem es steht?

1 Verfahren

Zunächst definieren wir semantische Nähe als den Grad der Substituierbarkeit eines Wortes gegen ein anderes aus demselben Paradigma in einem bestimmten Kontext. Als Datenbasis für die Lösung des oben skizzierten Problems haben wir eine statistisch signifikante Menge von Nominalphrasen bestehend aus attributiven Adjek-

[1] Was auch in dem Deutschen eng verwandten Sprachen, wie dem Niederländischen oder dem Westfriesischen, möglich ist (*klare wijn (in)schenken/kleare wyn skinke*).

tiven und Nomina aus geparsten (bzw. getaggten) englisch- sowie deutschsprachigen Korpora extrahiert und deren Elemente durch Wörter semantischer Relationen ersetzt. Vermittels statistischer Verfahren haben wir die daraus resultierenden Kombinationen evaluiert.

2 Gewinnung semantischer Relationen

Zur Gewinnung semantischer Relationen haben wir uns für das Englische der lexikalischen Datenbank WordNet[2] in der Version 1.6 bedient. Aufsetzend auf dem in Rennie (2000) beschriebenen Perl-Modul konnten wir die für unsere Aufgabenstellung einschlägigen Beziehungen wie Synonymie, Antonymie, Hypero-/Hyponymie, Mero-/ Holonymie, sowie Ähnlichkeit, Partizip eines Verbs und Zugehörigkeit zu einem Nomen für Adjektive, extrahieren.[3]
Da für unsere Untersuchung keine ähnlich elaborierte, elektronisch verfügbare Ressource für das Deutsche zur Verfügung stand, haben wir uns auf Synonyme und (soweit vorhanden) Antonyme und "ähnliche Ausdrücke bzw. Wörter" aus dem Microsoft WinWord-Thesaurus beschränkt, die wir anhand eines Perl-Skripts, gewinnen konnten.

3 Statistische Verfahren

Zur Auswertung der erzeugten Bigramme kamen Methoden der analytischen Statistik zum Einsatz, anhand derer man von Verhältnissen von Stichproben auf Verhältnisse der Grundgesamtheit vermittels der Überprüfung von Hypothesen schließen kann. Dazu wandten wir den in Manning/Schütze (2000:166f) beschriebenen t-Test zur Differenzierung synonymer Kollokanteme und das von uns modifizierte log-likelihood Verfahren zur Auflösung von Anbindungsambiguitäten[4] an.

3.1 Differenzierender t-Test

Der t-Test verwendet den Mittelwert und die Varianz einer Datenstichprobe, wobei die Nullhypothese ist, dass die gewählte Stichprobe aus einer Verteilung mit dem Mittelwert μ stammt. Dabei wird die durch die Varianz der Daten skalierte Differenz zwischen den tatsächlich vorgekommenen und den erwarteten Mittelwerten berechnet, sodass man einen Wert erhält, der eine Aussage darüber erlaubt, ob die gewählte Stichprobe (unter der Annahme, dass sie aus einer Normalverteilung mit dem Mittelwert μ stammt) den erwarteten Mittelwert und die erwartete Varianz widerspiegelt, oder ob sie einem extremeren Wertebereich zuzurechnen ist:[5]

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}}$$

Dementsprechend kann der t-Test auch zur Abgrenzung synonymer Kollokanteme und zu deren Desambiguierung eingesetzt werden.[6] Dazu erweitert man den t-Test zum Vergleich der Mittelwerte zweier normaler Grundgesamtheiten folgendermaßen:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Als Nullhypothese formulieren wir, dass die durchschnittliche Differenz 0 ist. Daraus ergibt sich, dass

$$\bar{x} - \mu = \bar{x} = \frac{1}{N}(\sum x_{1_i} - x_{2_i}) = \overline{x_1} - \overline{x_2}\,.$$

Ist *w* das Kollokat und v^1 und v^2 die zu vergleichenden Kollokanteme, bekommen wir $\overline{x_1} = s_1^2 = P(v^1 w)$ und $\overline{x_2} = s_2^2 = P(v^2 w)$, sodass

$$t \approx \frac{P(v^1 w) - P(v^2 w)}{\sqrt{\frac{P(v^1 w) + P(v^2 w)}{N}}}$$

was sich zu

---

[2] Miller (1990)

[3] Da unser Ansatz keine semantische Desambiguierung durch den Kontext vorsieht, verwenden wir unterschiedslos alle im Thesaurus vorhandenen Bedeutungen.

[4] Manning/Schütze (2000:280-84)

[5] Manning/Schütze (2000:163); in der Gleichung bezeichnet $\bar{x}$ den Mittelwert, s^2 die Varianz, N die Größe der Stichprobe und μ den Durchschnitt der Verteilung.

[6] Manning/Schütze (2000:166f)

$$t \approx \frac{\frac{C(v^1w)}{N} - \frac{C(v^2w)}{N}}{\sqrt{\frac{C(v^1w)+C(v^2w)}{N}}} = \frac{C(v^1w)-C(v^2w)}{\sqrt{C(v^1w)+C(v^2w)}}$$

vereinfachen lässt.[7]

Betrachten wir dazu ein Beispiel: In einem Kontext, in dem *prekäre* das Kollokat zum Kollokantem *Situation* bildet, stellen Systemwörter wie *Lage* oder *Sachlage* potentielle Substitutionskandidaten zum Nomen dar. Zur Berechnung der t-Test-Differenz benötigen wir lediglich die Bigramm-Häufigkeiten:[8]

$$t = \frac{C(\textit{prekäre}, \textit{Situation}) - C(\textit{prekäre}, \textit{Lage})}{\sqrt{C(\textit{prekäre}, \textit{Situation}) + C(\textit{prekäre}, \textit{Lage})}}$$

$$= \frac{2470-1880}{\sqrt{2470+1880}} \approx 8.95$$

$$t = \frac{C(\textit{prekäre}, \textit{Situation}) - C(\textit{prekäre}, \textit{Sachlage})}{\sqrt{C(\textit{prekäre}, \textit{Situation}) + C(\textit{prekäre}, \textit{Sachlage})}}$$

$$= \frac{2470-2}{\sqrt{2470+2}} \approx 49.64$$

Durch den Vergleich der beiden gewonnenen Werte ergibt sich also, dass *Situation* in einer Umgebung mit *prekäre* eher durch *Lage* ersetzt werden kann als durch *Sachlage*.

3.2 log-likelihood

Ebenso wie für den t-Test lässt sich auch für log-likelihood ein Verfahren konstruieren, mit dem sich graduelle Präferenzen für verschiedene Kollokanteme erfassen lassen. Wir entlehnen diesen Ansatz der Auflösung von Anbindungsambiguitäten (z.B. der Frage, ob eine Präposition an ein Verb oder ein Nomen gebunden wird) von Hindle und Rooth (1993),[9] um zu entscheiden, welches der zur Auswahl stehenden Systemwörter dem im Ausgangsbigramm stehenden Kollokantem semantisch am nähesten steht.

Bestimmt man wiederum im Bigramm *prekäre* Situation das Nomen zum Kollokantem und betrachtet dazu die möglichen Substitute *Lage* oder *Sachlage*, lässt sich folgende Berechnung durchführen: Zunächst ermitteln wir die einfachen Ausgangswahrscheinlichkeiten aus der Häufigkeit des Bigramms und der Häufigkeit des gewählten Systemworts:

$$P(\Pr ox_{Situation} = 1 \mid \textit{prekäre})$$

$$= \frac{C(\textit{prekäre}, \textit{Situation})}{C(\textit{Situation})} = \frac{2470}{12100000} \approx 0.0002$$

$$P(\Pr ox_{Lage} = 1 \mid \textit{prekäre})$$

$$= \frac{C(\textit{prekäre}, \textit{Lage})}{C(\textit{Lage})} = \frac{1880}{788000} \approx 0.0024$$

Berechnen wir nun die Wahrscheinlichkeit dafür, dass *Lage* nicht in der Nähe von *prekäre* vorkommt:

$$P(\Pr ox_{Lage} = 0 \mid \textit{prekäre})$$

$$= 1 - P(\Pr ox_{Lage} = 1 \mid \textit{prekäre}) \approx 0.9998$$

Aus diesen Werten lässt sich nun die likelihood-ratio errechnen:

$$\lambda(\textit{prekäre}, \textit{Situation}/\textit{Lage})$$

$$\approx \log \frac{0.0024 \times 0.9998}{0.0002} \approx 0.1684$$

Führen wir nun die gleiche Berechnung für Sachlage durch,

$$P(\Pr ox_{Situation} = 1 \mid \textit{prekäre})$$

$$= \frac{C(\textit{prekäre}, \textit{Situation})}{C(\textit{Situation})} = \frac{2470}{12100000} \approx 0.0002$$

$$P(\Pr ox_{Sachlage} = 1 \mid \textit{prekäre})$$

$$= \frac{C(\textit{prekäre}, \textit{Sachlage})}{C(\textit{Sachlage})} = \frac{2}{31500} \approx 0.00006$$

$$P(\Pr ox_{Sachlage} = 0 \mid \textit{prekäre})$$

$$= 1 - P(\Pr ox_{Sachlage} = 1 \mid \textit{prekäre}) \approx 0.99994$$

$$\lambda(\textit{prekäre}, \textit{Situation}/\textit{Sachlage})$$

$$\approx \log \frac{0.00006 \times 0.99994}{0.0002} \approx 10.2701$$

ergibt sich aus dem Vergleich der beiden ermittelten Werte, dass das Systemwort *Lage* dem betrachteten Kollokantem *Situation* in einem Kontext mit *prekäre* wesentlich näher steht als *Sachlage*, was unserer Intuition über diese Sprachdaten entspricht.

---

[7] Manning/Schütze (2000:168); C bezeichnet die Häufigkeit des Wortes im Korpus.

[8] Die Synonyme haben wir gemäß dem in Abschnitt 2 erläuterten Verfahren aus dem WinWord-Thesaurus gewonnen, während die Frequenzen einer Anfrage per Perl-Skript an die WWW-Suchmaschine Google entstammen.

[9] Manning/Schütze (2000:280-84)

4 Auswertung

Entgegen der in Manning/Schütze (166ff, 280-84) vorgeschlagenen Interpretation der Werte, lässt sich aus unseren Ergebnissen folgende Heuristik ableiten: Je kleiner der aus dem t-Test oder dem log-likelihood-Verfahren gewonnene Wert, umso eher ist das betrachtete Wort im Kontext ein geeignetes Substitut.

Betrachten wir dazu zunächst wiederum das Bigramm *prekäre Situation* als Beispiel für eine freie Alternation: Der WinWord-Thesaurus liefert als Substitute für *Situation* die Einträge *Sache*, *Lage*, *Konstellation*, *Zustand* und *Sachlage*. Aus den t-Werten ergibt sich eindeutig, dass *Lage* mit seinem niedrigen Wert mit ca. 9 das beste Substitut ist, während die anderen Werte relativ homogen um 48 und 49 liegen, woraus sich keine gute Abstufung erkennen lässt. Anders dagegen im Fall der log-likelihood: Hier besitzt *Lage* wiederum den niedrigsten Wert, doch kann man deutlich aus den anderen Werten ablesen, dass sich *Zustand* und *Sache* als nächstmögliche Substitute mit ungefähr gleichem Abstand anbieten, während *Konstellation* und *Sachlage* aufgrund ihrer Werte in diesem Kontext als nicht akzeptabel gelten dürften.

Anders stellt sich die Situation dar, wenn man z.B. die Substitute für *kalt* des statischen Idioms des *kalten Kaffees* aus der Einleitung betrachtet: Zunächst lässt sich beobachten, dass von den 44 Resultaten, die WinWord geliefert hat, im betrachteten Korpus nur sieben Adjektive ein Bigramm mit *Kaffee* bilden. Darüber hinaus ist auffällig, dass nur *frischer* signifikant häufig in diesem Bigramm auftritt, während die restlichen Kandidaten - mit Ausnahme von *bitterer* (46) - nicht mehr als 15 mal in einem Kontext mit *Kaffee* zu finden sind. Aus unseren statistischen Verfahren ergeben sich daher folgende (gerundete) Werte:

	t-Test	log-likelihood
lauer Kaffee	57.66	6.60
frischer Kaffee	47.03	0.05
kühler Kaffee	57.69	6.93
eiskalter Kaffee	57.40	5.02
harter Kaffee	57.77	8.93
bitterer Kaffee	56.74	3.56
grausamer K.	57.77	8.93

Zwar lassen sich durch eine solche Methode die statistisch signifikantesten Substitute ermitteln, doch stellt sich im Falle eines Idioms dadurch ein unerwünschtes Ergebnis als Vorhersage ein. Dementsprechend sollte diesem Verfahren eine zuverlässige Methode zur Klassifikation von Kollokationen als Idiom vorgeschaltet sein.

Ähnliches gilt für das Englische: Für das Bigramm *careful consideration* lieferte WordNet 21 Kandidaten als Substitute für *careful*, von denen nur zwei nicht in Bigrammen mit *consideration* vorkamen:[10]

	t-Test	log-likelihood
studious consideration	439.21	11.31
measured consideration	437.96	7.67
mindful consideration	439.18	10.98
close consideration	425.39	4.29
provident consideration	439.31	16.31
diligent consideration	438.00	7.71
cautious consideration	438.23	7.99
conscientious consideration	438.21	7.96
prudent consideration	437.16	7.00
unhurried consideration	439.01	9.80
detailed consideration	350.40	1.50
deliberate consideration	432.70	5.37
protective consideration	439.27	12.61
minute consideration	437.89	7.59
troubled consideration	439.31	14.72
careless consideration	439.07	10.14

[10] Aus den ermittelten Zahlen ergibt sich, dass die aus dem t-Test resultierenden Wertebereiche stark fluktuieren, sodass sich einerseits Schwierigkeiten bei der Interpretation der Werte ergeben können, andererseits ihre einheitliche Visualisierung problematisch ist.

certain consideration	436.72	6.73
particular consideration	388.73	2.37
thorough consideration	404.88	2.95

Die besten Substitute finden sich in den ersten drei Bedeutungen von *careful* (*detailed*, *particular*, *thorough*, *close*), während die schlechtesten Kandidaten aus den letzten beiden Einträgen (*troubled*, *provident*). Aus den Glossen zu diesen lässt sich ablesen, dass diese Bedeutungen nicht sehr gebräuchlich sind: "*(archaic) full of cares or anxiety*" (*troubled*), "*mindful of the future in spending money*" (*provident*).

Als Idiom des Englischen haben wir *blue chips* ausgewählt: Aus den 43 Kandidaten für *blue* kamen 27 nicht in einem Bigramm mit *chips* vor. Die beste Alternative ist in beiden Verfahren *colored* (268/4.8); die restlichen möglichen Substitute hingegen müssen aufgrund ihrer hohen Werte verworfen werden (so z.B. *blue-blooded*, *dismal* oder *naughty*, die im log-likelihood allesamt Werte über 14 erreichten).

Fazit

Aus den bisher durchgeführten Stichproben ergibt sich, dass sowohl der differenzierende t-Test als auch das modifizierte log-likelihood Verfahren die signifikantesten Substitute bezüglich ihrer semantischen Nähe zum Ausgangswort richtig klassifiziert. Allein in Fällen, in denen das Ausgangswort Teil eines Idioms ist, werden aufgrund der Nicht-Substituierbarkeit von Elementen einer festen Wendung zwar statistisch einschlägige Voraussagen getroffen, die aber im Kontext nicht die intendierte Bedeutung des Ausdrucks widergeben.

Als weitere Aufgaben, die sich aus dieser Arbeit ergeben, steht die Überprüfung unserer Thesen anhand einer signifikanten Menge von Bigrammen an erster Stelle. Dazu müssen manuell die erzeugten Verwandtschaftsverhältnisse kontrolliert werden, sodass man anschließend vermittels der Signifikanzmaße *Precision* und *Recall* entscheiden kann, ob die vorgestellten Verfahren praktikabel sind.

Darüber hinaus müsste geklärt werden, ob sich das hier skizzierte Verfahren einerseits auf andere Wortarten (und somit auch auf andere syntaktische Verhältnisse) übertragen lässt, und inwieweit sich die Erweiterung des betrachteten Kontextes auf Trigramme oder weitere Wortkombinationen auswirkt. Sollten sich daraus keine signifikanten Veränderungen der Voraussagemächtigkeit ergeben, ließe sich - als mögliche Anwendung der vorgestellten Verfahren - z.B. ein elektronischer Thesaurus projektieren, der auch Kontextinformationen berücksichtigt.

Literatur

Evert, S. und Krenn, B. (2001): Methods for the Qualitative Evaluation of Lexical Association Measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*. Toulouse, Frankreich

Evert, S., Heid, U. und Lezius, W. (2000): Methoden zum Vergleich von Signifikanzmaßen zur Kollokationsidentifikation. In Zühlke, W. (Hrsg.): *Sprachkommunikation: KONVENS 2000*. Berlin: VDE-Verlag

Lehr, Andrea (1996): *Kollokationen und maschinenlesbare Korpora: Ein operationales Analysemodell zum Aufbau lexikalischer Netze.* Tübingen: Niemeyer

Manning, C. und Schütze, H. (1999): *Foundations of Statistical Natural Language Processing*. Cambridge (MA), London: MIT Press

Miller, George A. (1990): WordNet: An On-line Lexical Database. *International Journal of Lexicography*

Pearce, Darren (im Erscheinen): A Comparative Evaluation of Collocation Extraction Techniques. In *Third International Conference on Language Resources and Evaluation*. Las Palmas, Spanien

Pearce, Darren (2001): Synonymy in Collocation Extraction. In *WordNet and Other Lexical Resources: Applications, Extensions and Customizations (NAACL 2001 Workshop)*. Pittsburgh, USA

Rennie, Jason (2000): *WordNet::QueryData: A Perl Module for Accessing the WordNet Database.* `http://ai.mit.edu/people/jrennie/WordNet`