Probabilistisches Clustering zur Identifikation von Verb-Nomen-Kollokationen

Detlef Prescher und Ulrich Heid
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart

11. Mai 2000

Zusammenfassung

Ausgehend von zwei der zentralen Annahmen des in der Lexikographie verbreiteten Kollokationsbegriffs, wie er etwa von Hausmann formuliert wurde, versuchen wir, Kollokationen aus Textcorpora zu extrahieren und den Anteil nicht-kollokatorischer Kombinationen in den Resultaten möglichst niedrig zu halten.

Die beiden definitionsrelevanten Aspekte sind folgende: (1) in der Kollokation ist die Wahl des Kollokators nicht frei; (2) die Bedeutung der Kollokatoren weicht von der 'regulären' Bedeutung derselben Lexeme in freien (Hausmann: 'trivialen') Kombinationen ab.

Unser Experiment ist wie folgt angelegt: Aus einem Zeitungscorpus (4.1 Millionen Wörter) werden Verb-Komplement- und Subjekt-Verb-Paare (Verb und Kopf der jeweiligen Nominalphrase) extrahiert, und zwar freie und potentiell kollokatorische Kombinationen ohne Unterschied. Mithilfe von probabilistischem Clustering werden statistisch motivierte Selektionsklassen ermittelt, von denen wir annehmen, daß sie die oben angesprochenen Subregularitäten und 'triviale' Kombinationen enthalten.

Nun werden alle im Corpus beobachteten Nomen-Verb-Paare daraufhin untersucht, ob sie in die Selektionsklassen fallen oder nicht. Diejenigen Kombinationen, die als individuelle Kombination besonders wahrscheinlich sind, aber außerhalb der Klassen liegen, werden speziell betrachtet: dort wo der Unterschied zu den vom Wahrscheinlichkeitsmodell beschriebenen Selektionsklassen besonders groß ist, vermuten wir Funktionsverbgefüge und Kombinationen mit synsemantisch gebrauchten Verben.

Die ersten Ergebnisse entsprechen relativ gut unserer Intuition. Zur Evaluierung vergleichen wir die mit unserem Verfahren gefundenen Kollokationskandidaten mit den Kollokationsangaben des deutschen

Teils von *Langenscheidts Handwörterbuch Deutsch-Englisch* und mit anderen manuell evaluierten Daten aus Corpora.

1 Einführung

Ziel der hier beschriebenen Arbeiten ist die Identifikation von Nomen-Verb-Kollokationen in Textcorpora, als Grundlage für den Wörterbuchaufbau.

Dabei gehen wir von einem Kollokationsbegriff aus, wie er z.B. von Hausmann ([Hau85], [Hau89], [Hau97]) formuliert und von Bergenholtz/Tarp (cf. [BT94]) ergänzt wurde: das Nomen wird als "semantisch autonome" Basis der Kollokation angesehen, das Verb als "synsemantischer" Kollokator.

Das als Kollokator benutzte Verb hat, so die Annahme, in der Kollokation eine spezifische, gegebenenfalls abgeschwächt-schematisierte Bedeutung (im Fall von Funktionsverbgefügen, cf. [Hel84]), jedenfalls eine, die nicht ganz der erwarteten kompositionell ableitbaren entspricht.

Kollokationen folgen weitgehend den syntaktischen Konstruktionsprinzipien für Verb-Komplement-Konstruktionen. Einer der spezifischen Aspekte von Kollokationen (und der Grund dafür, daß man Kollokationen eigens lernen muß) ist die Tatsache, daß die Kombination der Lexeme von Basis und Kollokator nicht vorhergesagt werden kann. Vorhersagbarkeit ist bei (im Rahmen der syntaktischen Constraints) freier Kombinierbarkeit gegeben, oder, im Sinne von Subregularitäten, bei Selektionsklassen von Basen (anhand Sorten beschreibbare Kookkurrenz, cf. [Hei94]).

Wir versuchen, diese beiden Annahmen (in der Kollokation ist die Wahl des Kollokators nicht frei; die Bedeutung der Kollokatoren weicht von der 'regulären' Bedeutung derselben Lexeme in freien (Hausmann: 'trivialen') Kombinationen ab) für ein corpus-basiertes Experiment zur Kollokationsextraktion nutzbar zu machen.

2 Probabilistisches Clustering zur Identifikation von Verb-Nomen-Kollokationen

2.1 Probabilistisches Clustering

In der maschinellen Sprachverabeitung wird probabilistisches Clustering hauptsächlich für die Disambiguierung und Klassifizierung dyadischer linguistischer Daten, z.B. von Paaren aus grammatischen Prädikat-Argument-Relationen (Verb-Nomen, Adjektiv-Nomen, etc.) eingesetzt. Während für die

erste Aufgabe die Daten-Cluster nur als generelle Hintergrundquelle für Informationen über ungesehene oder seltene Datenpaare dienen, sind sie für die Klassifizierung als den Daten unterliegende statistische "Semantik" relevant ([RRP+98]).

Im Ansatz von [Roo98] und [RRP+99] wird Clustering als Klassen-basiertes Wahrscheinlichkeitsmodell über Wortpaare konzipiert, in unserem Fall Nomen-Verb-Paare aus den wahrscheinlichsten Parse-Bäumen ([BCP+98], [BCP+99]) des 4.1-Millionen-Wort-Corpus. Klassen werden hierbei als versteckte Variablen gehandhabt, die die beobachteten Verb-Nomen-Paare komplettieren.

Die Einschätzung der Klassen-Parameter geschieht in einem mathematisch wohldefiniertem statistischen Inferenzverfahren, dessen wesentliche Grundannahme ist, dass die Verben einer jeden Klasse frei mit den Nomina dieser Klasse kombinieren können. Gerade diese Annahme trifft aber auf Kollokationen nicht zu: die jeweilige Kombination ist (lexikalisch) präferiert. Somit liegt die Idee nahe, dass das Klassen-basierte Wahrscheinlichkeitsmodell den Kollokationen Wahrscheinlichkeitsmasse "stiehlt", um sie den Paaren zuweisen zu können, die sich Modell-konform verhalten, d.h. die in ihrer jeweiligen Klasse frei mit anderen Partnern kookkurrieren ('Selektion').

Im folgenden werden wir diese Idee formalisieren, um ein Bewertungs-Kriterium dafür zu erhalten, ob ein gegebenes Nomen-Verb-Paar als Kollokation anzusehen ist.

2.2 Kollokationen: Empirische und Modell-Wahrscheinlichkeit

Für eine gegebene Grundmenge (ein 'Corpus') von Nomen-Verb-Paaren sei V bzw. N die Menge der in dieser Grundmenge vorkommenden Verben bzw. Nomina. Für ein beliebiges Verb-Nomen-Paar $(v, n) \in V \times N$ sei $f(v, n)$ die beobachtete Anzahl Vorkommen des Paares in der Grundmenge (entsprechend den syntaktischen Extraktionsresultaten).

Dann bezeichnen wir mit:

$$\tilde{P}(v, n) = \frac{f(v, n)}{\sum_{(v,n) \in V \times N} f(v, n)}$$

die *empirische Wahrscheinlichkeit* des Verb-Nomen-Paares $(v, n) \in V \times N$.

Als *Klassen-Modell* wollen wir jedes Modell:

$$\begin{aligned} P_{lc}(v,n) &= \sum_{c \in C} P_{lc}(c) P_{lc}(v,n|c) \\ &= \sum_{c \in C} P_{lc}(c) P_{lc}(v|c) P_{lc}(n|c) \end{aligned}$$

bezeichnen, welches die Corpus-Wahrscheinlichkeit $\prod_{v \in V, n \in N} P_{lc}(v,n)^{f(v,n)}$ maximiert.

In [Roo98] und [RRP+99] wird ein einfacher Inferenzalgorithmus angegeben, um für vorgegebene Klassenanzahl $|C|$ ein *Klassen-Modell* zu induzieren. Dort wird auch gezeigt, dass solche Modelle ausgezeichnete Disambiguierungs- und Klassifizierungs-Eigenschaften haben.

Ein Klassen-Modell macht für jede seiner Klassen $c \in C$ die folgende probabilistische Unabhängigkeitsannahme:

$$P_{lc}(v,n|c) = P_{lc}(v|c) P_{lc}(n|c)$$

Für Kollokationen gilt diese Unabhängigkeitsannahme nun aber genau nicht, sodass wir annehmen können, dass eine Kollokation (v,n) von ihrer ursprünglichen Wahrscheinlichkeit – der empirischen Wahrscheinlichkeit $\tilde{P}(v,n)$ – besonders viel "Masse" an andere Verb-Nomen-Paare abgeben muss. In vereinfachter Überlegung erwarten wir deshalb, dass die abgegebene Wahrscheinlichkeitsmasse:

$$\tilde{P}(v,n) - P_{lc}(v,n)$$

für Kollokationen vergleichsweise grössere Werte als für modell-konforme (in Hausmann's Sinn 'triviale' oder den Selektions-Subregularitäten entsprechende)Verb-Nomen-Paare annimmt.

3 Experimente

In einem Experiment extrahierten wir 318086 Typen von Verb-Nomen- und Adjektiv-Nomen-Paare aus den wahrscheinlichsten Parse-Bäumen eines 4,1 Mio. Wort-Korpus deutscher Verb-Letzt-Sätze ([BCP+98], [BCP+99]) und trainierten damit ein 35-Klassen-Modell ([RRP+98], [RRP+99]).

Für die Verb-Nomen-Paare aus dem Trainingskorpus wurde die Bewertung $\tilde{P}(v,n) - P_{lc}(v,n)$ berechnet und absteigend sortiert.

Ein erster Blick auf die Ergebnisse legt nahe, daß in der Tat Kombinationen, die besonders viel Wahrscheinlichkeitsmasse abgegeben haben, intuitiv als signifikate Kollokationen angesehen werden: Beispiele sind:

- *Rolle + spielen, Problem + lösen, Erfolg + haben, Gebrauch + machen, Ziel + erreichen, Anspruch + haben, Spaß + machen* etc.

Am Ende der sortierten Liste treten primär kompositionell analysierbare, der jeweiligen Selektionsklasse entsprechende Wortgruppen auf:

- *der Polizei vorliegen, Garten + betreten, der Stadt guttun, Arbeitsplatz + aufgeben, Gegner + sehen.*

4 Evaluierung

Ziel der Untersuchungen ist es, Anhaltspunkte für die Grenzen zwischen Kollokation und Selektion zu bekommen; daß im allgemeinen Fall keine ganz klare Grenze gezogen werden kann, ist weithin akzeptiert. Trotzdem können Lexikographen relativ gut zwischen kollokatorischen und nicht-kollokatorischen Kombinationen unterscheiden. Benutzen sie ein Werkzeug, das ihnen aus Corpora extrahierte Kollokationskandidaten vorschlagen soll, so ist natürlich ihr Interesse, möglichst viel kollokatorische Kombinationen und wenig Triviales zu erhalten.

Unsere Ergebnisse werden durch Abgleich gegen verschiedene andere Wissensquellen mindestens qualitativ evaluiert; aufgrund der Annahme, daß Wörterbücher die deskriptive Intuition der Redakteure widerspiegeln, wurden unsere Kandidaten-Kombinationen daraufhin überprüft, ob sie im deutschen Teil eines zweisprachigen Wörterbuchs Deutsch-Englisch vorkommen [1]

Daneben wurden unsere Ergebnisse mit Daten verglichen, die mit partiellem Parsing aus den Verb-Letzt-Sätzen eines Corpus von 300 Millionen Wortformen extrahiert und nachfolgend mithilfe des Kookkurenzmaßes Log-likelihood nach Signifikanz sortiert worden waren (cf. [Hei98]).

Literatur

[BCP+98] Franz Beil, Glenn Carroll, Detlef Prescher, Stefan Riezler, and Mats Rooth. Inside-outside estimation of a lexicalized PCFG for German. –Gold–. In *Inducing Lexicons with the EM Algorithm*, AIMS Report 4(3). IMS, Universität Stuttgart, 1998.

[1] Es handelt sich um *Langenscheidts Handwörterbuch Deutsch-Englisch.* Vincent Docherty und der Redaktion Wörterbücher der Langenscheidt KG möchten wir besonders Dank für die Überlassung der Daten für diesen Zweck danken.

[BCP+99] Franz Beil, Glenn Carroll, Detlef Prescher, Stefan Riezler, and Mats Rooth. Inside-outside estimation of a lexicalized PCFG for German. In *Proceedings of the 37th Annual Meeting of the ACL*, Maryland, 1999.

[Ben89] Morton Benson: "The Structure of the Collocational Dictionary", in: *International Journal of Lexicography*, Oxford, Oxford University Press, 1989.

[BT94] Henning Bergenholtz, Sven Tarp: *Mehrworttermini und Kollokationen in Fachwörterbüchern*, in [SB94]: 385 – 419

[Hau85] Franz Josef Hausmann: "Kollokationen im deutschen Wörterbuch. Ein Beitrag zur Theorie des lexikographischen Beispiels", in: [Bergenholtz/Mugdan 1985], pp. 118-129, 1985.

[Hau89] Hausmann F.J.: "Le dictionnaire de collocations", in: [Hausmann e.a. 1989], pp. 1010-1019, 1989.

[Hau+89] Hausmann F.J., Reichmann O., Wiegand H.E., Zgusta L. (eds.): *Wörterbücher: ein internationales Handbuch zur Lexikographie. Dictionaries. Dictionnaires. Wörterbücher*, Berlin, New York, de Gruyter, vol. 1, 1989.

[Hau97] Franz Josef Hausmann: "Semiotaxis und Wörterbuch"; in: [KL97], pp. 171-179.

[Hei98] Ulrich Heid:"Towards a corpus-based dictionary of German noun-verb collocations"; in: *Proceedings of the* EURALEX *International Congress 1998*, (Liège), 1998

[Hei94] Ulrich Heid: "On Ways Words Work Together – Topics in Lexical Combinatorics", in: W. Martin, W. Meijs, M. Moerland, E. ten Pas, P. van Sterkenburg, P. Vossen (Eds.), *Euralex 1994, Proceedings*, Amsterdam, 1994, SS. 226-257

[Hel84] Gerhard Helbig: "Probleme der Beschreibung von Funktionsverbgefügen im Deutschen", in: Gerhard Helbig: *Studien zur deutschen Syntax*, Bd.2, (Leipzig) 1984

[KL97] Klaus-Peter Konerding, Andrea Lehr: *Linguistische Theorie und lexikographische Praxis*; Symposiumsvorträge, Heidelberg 1996; Lexicographica Series Maior 82; Niemeyer 1997.

[Roo98] Mats Rooth. Two-dimensional clusters in grammatical relations. In *Inducing Lexicons with the EM Algorithm*, AIMS Report 4(3). Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 1998.

[RRP+98] Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. EM-based clustering for NLP applications. In *Inducing Lexicons with the EM Algorithm*, AIMS Report 4(3). Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 1998.

[RRP+99] Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. Inducing a semantically annotated lexicon via EM-based clustering. In *Proceedings of the 37th Annual Meeting of the ACL*, Maryland, 1999.

[SB94] Burkhard Schaeder/Henning Bergenholtz (Eds.) *Fachlexikographie. Fachwi ssen und seine Repräsentation in Wörterbüchern*, Tübingen: Narr, 1994.