

Joachim Griesbaum, Thomas Mandl,  
Christa Womser-Hacker (Hrsg.)

# Information und Wissen: global, sozial und frei?

Proceedings des 12. Internationalen Symposiums  
für Informationswissenschaft (ISI 2011)

Hildesheim, 9.–11. März 2011

**vwh**

Verlag Werner Hülsbusch  
Fachverlag für Medientechnik und -wirtschaft

# Domänenübergreifende Phrasenextraktion mithilfe einer lexikonunabhängigen Analysekomponente

*Daniela Becks, Julia Maria Schulz*

Universität Hildesheim – Institut für Informations- und Sprachtechnologie  
Marienburger Platz 22, 31141 Hildesheim  
daniela.becks@uni-hildesheim.de, julia-maria.schulz@uni-hildesheim.de

## **Zusammenfassung**

Der vorliegende Artikel beschreibt einen neuartigen domänenübergreifenden Ansatz zur Extraktion von Phrasen, der sich mit geringem Aufwand und ohne komplexe Lexika umsetzen und auf andere Domänen übertragen lässt. Dies wird anhand von Kundenrezensionen und Patentschriften getestet.

## **Abstract**

This paper presents a new approach, which can be easily adapted to different domains without the existence of comprehensive lexica. As test documents customer reviews and patent documents are used.

## **Einleitung**

Die Extraktion sinntragender Phrasen aus Korpora setzt i.d.R. ein intensives Verständnis der Texte und der betrachteten Domäne voraus. Auch bedarf es in vielen Fällen der Adaption verwendeter Wissensbasen und zugrunde liegender Modelle. Dieser Prozess ist nicht selten zeit- und arbeitsintensiv. Der vorliegende Artikel beschreibt einen neuartigen domänenübergreifenden Ansatz, der Shallow und Deep Parsing kombiniert und sich mit wenig Aufwand und ohne komplexe Lexika realisieren und auf andere Domänen über-

tragen lässt. Als Beispiel werden zwei sehr unterschiedliche Textdomänen herangezogen: Kundenrezensionen und Patentschriften.

Im nächsten Abschnitt wird zunächst auf existierende Ansätze eingegangen, bevor in Kapitel 3 der domänenübergreifende Ansatz beschrieben wird. Es schließt sich eine Beschreibung der Evaluierungsansätze an, bevor das Paper mit einem Ausblick schließt.

## Verwandte Arbeiten

Im Information Retrieval zeigt sich seit einiger Zeit, dass der klassische Bag-of-Words-Ansatz sowohl innerhalb der Indexierung als auch im Anfrageprozess zunehmend abgelöst wird. Viele Wissenschaftler vertreten die Meinung, Phrasen seien häufig präziser als einfache Terme (vgl. z.B. Tseng et al. 2007: 1222). So kann bspw. die Bedeutung der beiden Terme „schwarzes Schaf“ nur bei gemeinsamer Betrachtung (als Phrase) erfasst werden.

Zu den gängigen Verfahren der Phrasenextraktion zählen regelbasierte Ansätze wie das wörterbuchunabhängige *Begrenzerverfahren* (vgl. Jaene/Seelbach 1975). Für die Inhaltserschließung werden hier Phrasen in Form von Mehrwortgruppen, die als mehrere eine syntaktisch-semantische Einheit bildende Wörter definiert werden (vgl. ebd.: 9), aus englischen Fachtexten extrahiert. Dafür werden sogenannte Begrenzerpaare definiert, die als Grenzen für die zu extrahierenden Nominalphrasen fungieren (vgl. ebd.: 7). Diese bestehen aus Kombinationen von Stoppwörtern oder Satzzeichen, die in Listen erfasst sind (vgl. ebd.: 51 ff.). Ein ähnliches Verfahren, das innerhalb der Patentdomäne Anwendung findet, beschreiben Tseng et al. 2007. Sie ermitteln Phrasen bzw. Schlüsselwörter mithilfe einer Stoppwortliste. Die Autoren stellen fest, dass die längsten sich wiederholenden Phrasen häufig besonders gute Kandidaten darstellen (vgl. Tseng et al. 2007: 1223).

Ein klassisches linguistisches Verfahren bildet das Dependenzparsing, das die Abhängigkeiten der Satzglieder ermittelt. Im Information Retrieval finden sich Dependenzrelationen häufig als sogenannte Head/Modifier-Relationen wieder. Diese Head/Modifier-Paare setzen sich aus einem Head, welcher den Kern der Phrase darstellt, und einem Modifier zusammen, der der Präzisierung des Heads dient (vgl. Koster 2004: 423), wie das nachfolgende Beispiel zeigt: linguistic (= **modifier**) approach (= **head**).

Der Vorteil von Head/Modifier-Relationen liegt insbesondere darin, dass diese neben syntaktischen auch semantische Information enthalten (vgl. u.a. Ruge 1989: 9). Daher erfreuen sie sich vor allem im Rahmen des Indexierungsprozesses großer Beliebtheit (vgl. u.a. Koster 2004). In Form von Head/Modifier-Tripeln (Term-Relation-Term) erweisen sich Dependenzrelationen u.a. für Klassifikationsaufgaben als hilfreich (vgl. Koster/Beney 2009).

## Domänenübergreifende Phrasenextraktion

Die im Folgenden vorgestellte Methode für die Phrasenextraktion vereinigt nun die beiden zuvor genannten Verfahrensansätze. Als Anwendungsbereiche werden Patentschriften und Kundenrezensionen gewählt, die in zwei Projekten mit unterschiedlichen Zielsetzungen verwendet werden (vgl. Kapitel 4). Das Ziel des neuen Extraktionsverfahrens besteht darin, für beide Projekte ein Werkzeug zur Identifikation linguistischer Phrasen bereitzustellen, das sich mit geringem Aufwand für unterschiedliche Domänen adaptieren lässt und auch auf umfangreichen Korpora performant arbeitet. Dabei ist die Semantik der extrahierten Phrasen nicht zu vernachlässigen. Demgemäß wird ein Mischverfahren entwickelt, das auf linguistische Regeln zurückgreift, aber eher die Funktionalität eines Shallow Parsers aufweist.

Es wird ein regelbasiertes Verfahren eingesetzt, das sich z. T. auf das Begrenzerverfahren (vgl. Jaene/Seelbach 1975) zurückführen lässt, jedoch mit Ansätzen des Dependenzparsings (vgl. z.B. Ruge 1989) kombiniert wurde. Um ressourcenintensives syntaktisches Parsen zu vermeiden, erfolgt die Phrasenextraktion mithilfe verschiedener Regeln, in denen jeweils Paare von Begrenzern definiert sind. Im Unterschied zu dem oben beschriebenen Begrenzerverfahren werden hier Wortklassen (POS-Tags) statt Stoppwörtern verwendet. Durch deren Einsatz werden bereits bestimmte Phrasentypen vorgegeben. Das POS-Tag *DT* (Artikel) leitet bspw. ausschließlich Nominalphrasen ein. Die so definierten Phrasentypen sind abstrahiert und können leichter auf andere Sprachen und Domänen übertragen werden, da die Kategorie *DT* sowohl die deutschen Artikel *der*, *die*, *das* als auch das englische Pendant *the* umfasst. Diese abstrahierte Version des Begrenzerverfahrens ist daher deutlich flexibler und benötigt keine komplexen Wortlisten. Außerdem wird auf Grundzüge des Dependenzparsings zurückgegriffen. Jede der extra-

hierten Phrasen verfügt daher über einen Head und einen Modifier (vgl. Koster 2004). Die Beispiele in Abb. 1 verdeutlichen, dass es sich bei den extrahierten Phrasen nicht nur um Head/Modifier-Paare im engeren Sinne handeln muss, sondern auch längere Phrasen abgebildet werden.

Begrenzer: a(DT) & with(IN)	Begrenzer: a(DT) & ,(,)
a shank-like stud with (EP-1120530-B1)	very good front panel button layout (Hiu&Liu 2004)

Abb. 1: Visualisierte Beispielfrasen beider Domänen

## Evaluierungsansätze

Das Ziel im Opinion Mining Projekt ist das Extrahieren von Phrasen, die aus Meinungen bezüglich der rezensierten Produkte und deren Eigenschaften bestehen. Für Evaluierungszwecke liegt im Projekt ein Korpus vor, das auf Satzebene annotierte Produkteigenschaften und die diesbezüglich ausgedrückte Meinung enthält (vgl. Hu/Liu 2004; Ding et al. 2008).

Für explizit genannte Produkteigenschaften, wie „picture quality“ in folgendem Satz: „The picture quality is great.“ soll im Rahmen der Evaluierung überprüft werden, ob die jeweilige Phrase die annotierte Produkteigenschaft enthält. Ist dies der Fall, wird die Phrase als Treffer gewertet. Da implizit genannte Produkteigenschaften, wie „size“ im Satz „It fits in every pocket.“, so nicht evaluiert werden können, wird das Korpus um Markierungen der entsprechenden Textstellen, die die Produkteigenschaft aufweisen, erweitert. Für die Evaluierung werden jeweils *Recall* und *Precision* ermittelt.

Im Patent Retrieval-Projekt liegt der Fokus auf der Evaluierung der Genauigkeit der extrahierten Phrasen. Zu diesem Zweck wird auf einen Ansatz von Verbene et al. 2010 zurückgegriffen. Als Evaluierungsbasis verwenden die Autoren eine manuell annotierte Stichprobe von 100 Sätzen, die als Goldstandard betrachtet werden kann. Ein Abgleich der extrahierten Phrasen mit den intellektuellen Annotationen ermöglicht die Berechnung der *Accuracy*. In diesem Projekt bietet sich ein solcher Evaluierungsansatz ebenfalls an, da eine Stichprobe manuell annotierter Patentschriften bereits existiert.

## Ausblick

Erste Experimente und manuelle Auswertungen weisen auf eine viel versprechende und Ressourcen sparende Methode zum Extrahieren von Phrasen aus verschiedenen Korpora hin. Im Rahmen der beiden zuvor angesprochenen Projekte sollen im nächsten Schritt die in Kapitel 4 erläuterten Evaluierungsverfahren umgesetzt werden, um auf diese Weise den Eindruck der ersten manuellen Auswertungen empirisch zu stützen.

## References/Literaturverzeichnis

- Ding, X.; Liu, B.; Yu, Ph. S. (2008): A holistic lexicon-based approach to opinion mining. In: Proceedings of the international conference on Web search and web data mining. Palo Alto, California, USA: ACM, S. 231–240.
- Hu, M.; Liu, B. (2004): Mining Opinion Features in Customer Reviews. In: Proceedings of the 19th National Conference on Artificial intelligence. San Jose, California, USA: AAAI Press/The MIT Press, S. 755–760.
- Jaene, H.; Seelbach, D. (1975): Maschinelle Extraktion von zusammengesetzten Ausdrücken aus englischen Fachtexten. Berlin u.a.: Beuth.
- Koster, C. H. A. (2004): Head/Modifier Frames for Information Retrieval. In: Proceedings of the 5th Conference on Intelligent Text Processing and Computational Linguistics. Seoul, Korea: Springer (LNCS 2945), S. 420–432.
- Koster, C. H. A.; Beney, G. Jean (2009): Phrase-Based Document Categorization Revisited. In: Proceedings of the 18<sup>th</sup> Conference on Information and Knowledge Management. Hong Kong, China: ACM, S. 49–55.
- Ruge, G. (1989): Generierung semantischer Felder auf der Basis von Frei-Texten. In: LDV Forum 6, H. 2, S. 3–17.
- Tseng, Y.-H.; Lin, Ch.-J.; Lin, Y.-I. (2007): Text Mining Techniques for Patent Analysis. In: Information Processing and Management 43, H. 5, S. 1216–1247.
- Verbene, S.; D’hondt, E.; Oostdijk, N. (2010): Quantifying the Challenges in Parsing Patent Claims. In: Proceedings of the 1st International Workshop on Advances in Patent Information Retrieval (AsPIRe’10). Milton Keynes, S. 14–21.