GEFÖRDERT VOM

Technische Universität Dortmund

Fakultät Informatik
Lehrstuhl für Künstliche Intelligenz

Universität Mannheim

Seminar für deutsche Philologie
Lehrstuhl Germanistische Linguistik

Technischer Bericht

Nr. 2014/1 (Meilenstein 2)

Überwachte und unüberwachte Disambiguierung von KwiC-Snippets bei der Suche in großen Textkorpora

Data-Mining-Verfahren des KobRA-Projekts, Stand 08/2014

BMBF-Verbundprojekt:

Korpus-basierte linguistische Recherche und Analyse mithilfe von Data-Mining (KobRA)

Förderkennzeichen: 01UG1245A
Projektlaufzeit: 01.09.2012 bis 31.08.2015
Projektkoordination: Prof. Dr. Angelika Storrer
Koordination des Teilprojekts: Prof. Dr. Katharina Morik
Bearbeiter: Christian Pölitz, Thomas Bartz, Michael Beißwenger

Dortmund, den 31.8.2014

Das diesem Bericht zugrunde liegende Vorhaben wird mit Mitteln des Bundesministeriums für Bildung und Forschung (BMBF) unter den Förderkennzeichen 01UG1245A-D gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autoren.

Überwachte und unüberwachte Disambiguierung von KWIC-Snippets bei der Suche in großen Textkorpora

Data-Mining-Verfahren des KobRA-Projekts, Stand 08/2014

1. Projektkontext und Aufgabenstellungen

Das übergreifende Ziel des KobRA-Projekts besteht darin, durch den Einsatz innovativer Data-Mining-Verfahren (insbesondere Verfahren des maschinellen Lernens) die Möglichkeiten der empirischen linguistischen Arbeit mit strukturierten Sprachressourcen zu verbessern. Die Verfahren werden an linguistischen Fallstudien erprobt, die an konkrete Forschungsprojekte angebunden sind.

Dieser Report dokumentiert den Kernbestand der im KobRA-Projekt bis Projektmonat 24 (08/2014) entwickelten, angepassten und evaluierten Data-Mining-Verfahren. Die Verfahren lassen sich zur teilweisen bzw. vollständigen Automatisierung von routinisierten Aufgabenstellungen einsetzen, die in vielen korpusbasierten linguistischen Untersuchungen auftreten. Bei diesen Aufgabenstellungen, die in der ersten Projektphase abgeleitet aus konkreten Forschungsvorhaben im Verbund spezifiziert worden sind, handelt es sich um folgende:

Filtern

Zu einem sprachlichen Phänomen lassen sich zwar umfangreiche Trefferlisten aus Korpora gewinnen. Diese Trefferlisten sind aber nicht unmittelbar nutzbar, weil sie viele falsch positive Treffer enthalten, die mit der vorhandenen Korpustechnologie auch nicht weiter ausgefiltert werden können. Zur Verfügung gestellt werden deshalb Data-Mining-Verfahren, die den Korpus-Nutzer dabei unterstützen, falsch positive Treffer aus großen Suchtrefferlisten auszusondern.

Klassifizieren/Annotieren

Ein sprachliches Phänomen lässt sich mithilfe einer vorhandenen Korpusabfragesprache bzw. auf Basis der vorhandenen linguistischen Aufbereitung eines Korpus noch nicht präzise beschreiben und zielgenau erheben oder ein Analyseschritt erfordert eine feinere (Nach-)Klassifizierung der erhobenen Daten. Zur Verfügung gestellt werden deshalb Verfahren, die eine effiziente aufgabenbezogene Annotation ermöglichen.

Disambiguieren/Visualisieren

Ein Wort, dessen Gebrauch untersucht werden soll, tritt im Korpus mit hoher Frequenz auf. Die bei der Korpussuche generierten Trefferlisten sind aber nicht unmittelbar nutzbar, weil das gesuchte Wort in verschiedenen Bedeutungen vorkommt, die im Rahmen der Untersuchung zu unterscheiden und ggf. einzeln zu zählen sind, was aber mit der vorhandenen Korpustechnologie nicht automatisch möglich ist. Zur Verfügung gestellt werden deshalb Data-

Mining-Verfahren, die den Korpus-Nutzer dabei unterstützen, Trefferlisten zu einem Wort nach verschiedenen Bedeutungen dieses Wortes zu partitionieren.

Die Partitionierung ermöglicht zugleich anspruchsvolle Visualisierungen, die den Gebrauch von Wörtern über Zeitspannen und Textsortenbereiche hinweg in seiner Entwicklung auch grafisch sichtbar machen sowie neue Fragestellungen und Hypothesen induzieren können. Ein Werkzeug zur Visualisierung disambiguierter lexikalischer Einheiten wird daher ergänzend bereitgestellt.

Die für die Aufgabenstellungen eingesetzten Data-Mining-Verfahren lassen sich zwei übergeordneten Typen zuordnen:

1) Für das **Filtern** oder **Klassifizieren**, bei dem Korpus-Nutzer auf der Grundlage ihrer Expertise Merkmale oder Hypothesen in Bezug auf die Beschaffenheit der Ergebnismenge vorgeben, werden **überwachte Verfahren** eingesetzt.

2) Beim **Disambiguieren** nach Bedeutungen wird die Beschaffenheit der verschiedenen Ergebnismengen automatisch analysiert. Korpus-Nutzer machen hierzu keine Vorgaben, um auch bislang nicht bekannte bzw. dokumentierte Bedeutungen eines gesuchten Wortes entdecken zu können. Für diese Aufgabe sind **unüberwachte Verfahren** am besten geeignet.

Die Verfahren wurden als Plug-in für die Data-Mining-Umgebung „RapidMiner“ (früher „YALE“, Mierswa et al. 2006) implementiert und evaluiert. Mit der Implementierung ist auch die Voraussetzung für die Integration der Verfahren in die Infrastrukturen der KobRA-Projektpartner gegeben.

Der Report ist folgendermaßen aufgebaut: Im folgenden Abschnitt 2 beschreiben wir zunächst die zur Verfügung gestellten überwachten und unüberwachten Data-Mining-Verfahren sowie ihre Integration in die Data-Mining-Umgebung RapidMiner. Abschnitt 3 illustriert die Nutzung der Verfahren in linguistischen Fallstudien des Projekts und dokumentiert ihren Nutzwert anhand von Evaluationsergebnissen. Abschnitt 4 gibt eine Zusammenfassung und einen Ausblick auf geplante Anschlussarbeiten.

2. Data-Mining-Verfahren und -Umgebung, Werkzeuge zur Annotation und Visualisierung

2.1 Einlesen, Repräsentation und Nutzung der Korpusdaten

Die Data-Mining-Verfahren des KobRA-Projekts setzen direkt an der von einem Korpusabfragesystem zu einem gesuchten Ausdruck ausgegebenen Keyword-in-Context-Ergebnisliste (KwiC-Liste) an (s. Abbildung 1). Diese besteht aus kurzen Text-Snippets für jeden Treffer der Abfrage, die das Suchwort in einem Kontext von einigen Sätzen erhalten (je nach Fragestellung und genutztem Korpus variabel, meist 1-3 Sätze). Grundlage für das maschinelle Lernen sind also nicht die vollständigen Korpora, sondern eine vom Korpus-Nutzer auf Grundlage seiner Expertise schon auf die hochrelevanten Daten konzentrierte Auswahl. Die gängigen Abfragesysteme bieten dazu heute über ausgefeilte Abfragesprachen bereits umfangreiche Möglichkeiten an, das Suchergebnis abhängig von bestimmten Merkmalen möglichst präzise einzuschränken. Zu diesen Merkmalen zählen Wortformen und Phrasen, Wortabstände und -fenster bis hin zu regulären Ausdrücken für die Mustersuche, Lemmata (Rückführung der flektierten Wortformen auf die Grundform), morphosyntaktischen (Wortarten) und syntaktischen Informationen.

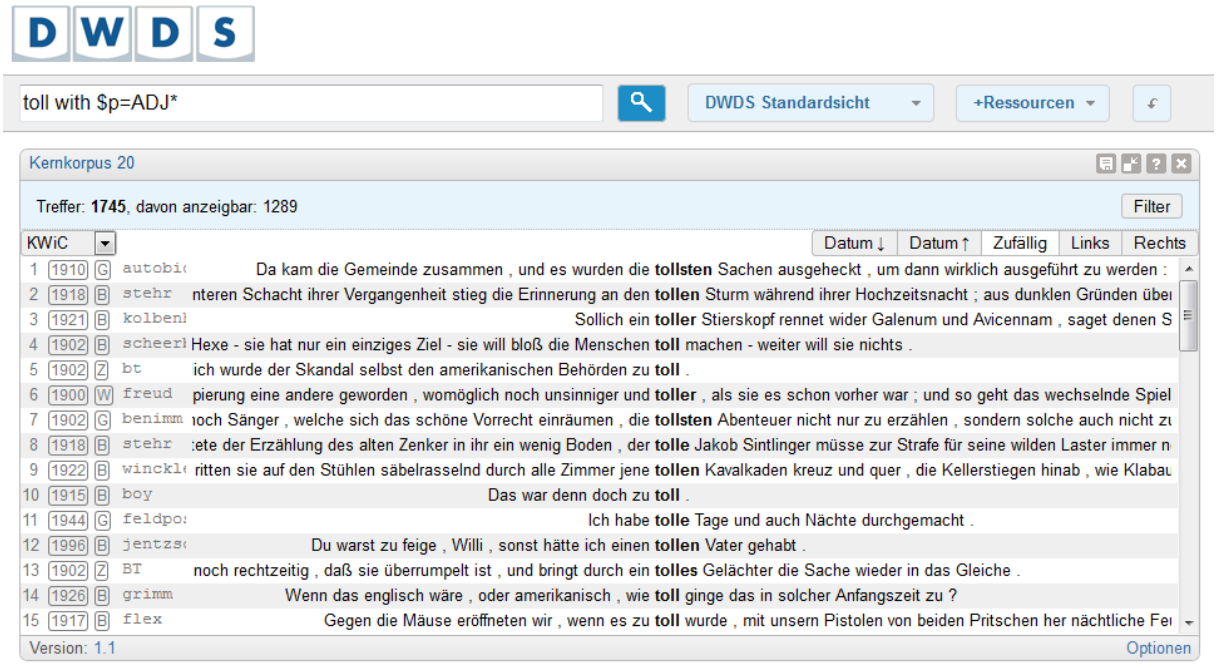

Abbildung 1: Abfrage zum Suchwort „toll“ im DWDS-Kernkorpus des 20. Jh. über das Abfragesystem des DWDS mit Nutzung des Wortarten-Filters (vgl. Geyken 2007, Klein & Geyken 2010).

Die durch Korpusabfrage gewonnenen Daten können unabhängig vom verwendeten Korpus in die im Projekt genutzte und angepasste Data-Mining-Umgebung (s. 2.4) eingelesen werden. Einzige Voraussetzung ist die Formatierung in einem Tabellenformat (z.B. als Komma-separated-Values/CSV oder XLS). Alternativ steht mit dem im Projekt entwickelten Plug-in ein Werkzeug („LinguisticQuery-Operator“) zur Verfügung, das die Abfrage der durch die KobRA-Projektpartner bereitgestellten Ressourcen direkt aus der Data-Mining-Umgebung heraus ermöglicht. Das Werkzeug unterstützt auch die oben genannten erweiterten Abfragemöglichkeiten gängiger Korpusabfragesysteme. Für das Auslesen der in vielen Korpora genutzten XML-Datenbasis (nach TEI-Standard, z.B.: Beißwenger et al. 2012) wurde als Bestandteil des RapidMiner-Plug-ins außerdem ein TEI-Reader bereitgestellt.

Die KwiC-Snippets werden für das maschinelle Lernen als Sequenzen von Wörtern repräsentiert (‚Bags-of-Words‘; grundlegende Repräsentation). Jedes Snippet wird als großer Vektor mit Einträgen für jedes Wort der Gesamtmenge aller Wörter in einer KwiC-Liste dargestellt (ein sogenannter ‚Wortvektor‘). In einer KwiC-Liste mit *N* Wörtern ist der Vektor *N*-dimensional. Die Elemente der Wortvektoren können binär sein und das bloße Vorkommen eines Wortes in einem Snippet oder Häufigkeiten des Wortes in einem Snippet und in allen Snippets der KwiC-Liste darstellen. Formal ist ein Wortvektor *v* für einen endlichen Text definiert als ein *N*-dimensionaler Vektor, d.h. alle möglichen Texte enthalten *N* unterschiedliche Wörter. Für *v* gilt, dass die *i*-te Komponente die Anzahl der Vorkommen oder (normalisierte) Frequenz von Wort *i* im Text ist. Ordnet man diese Wörter, so kann man jedes Wort über einen Index *i* identifizieren. Damit definieren wir eine Abbildung *Phi*, die die Snippets (hier wie ‚Texte‘ behandelt) als Wortvektoren abbildet. Dies geschieht formal so:

$\varphi(d) = (f(w1,d), f(w2,d), \ldots, f(wN,d))$, wobei $f(wi,d)$ die Anzahl oder (normalisierte) Frequenz von Wort *i* in Text *d* (für ‚document‘) angibt.

Für eine erweiterte Repräsentation, die die Berücksichtigung weiterer Merkmale (z.B. N-Gramme, Phrasen, morphosyntaktische Informationen, Dependenzen, Syntaxbäume) über die reinen Wortvorkommen hinaus beim maschinellen Lernen erlaubt, nutzen wir Kernmethoden

(Shawe-Taylor & Cristianini 2004), die die Ähnlichkeit für jedes mögliche Paar von Snippets angeben, indem sie die Snippets in einem Hilbertraum abbilden. Mithilfe der Stützvektormethode (auch ‚Support-Vector-Machine', kurz: SVM, Joachims 1998; s. 2.2) lässt sich daraufhin eine klassifizierende Hyperebene lernen (s. 2.2). Beispielsweise werden Parse-Bäume über sogenannte ‚Treekernels' in einen Hilbertraum gemappt, der von allen möglichen Teilbäumen aufgespannt wird. Mittels des sogenannten ‚Kerneltricks' kann dann eine Support-Vector-Maschine gelernt werden, ohne explizit alle möglichen Teilbäume aufzählen zu müssen (Collins & Duffy 2001).

2.2 Überwachte Verfahren

Überwachte Verfahren können zur Klassifikation von KwiC-Snippets auf das Vorkommen bestimmter sprachlicher Phänomene (für die Aufgabenstellung ‚Klassifizieren/Annotieren', s. 1.; z.B.: Verben in Verwendung als Stützverben, s. Bartz et al. 2013a; Aktionswörter oder nicht-kanonische Verwendungen von „weil", s. 3.1) bzw. bestimmter falsch positiver Treffer (für die Aufgabenstellung ‚Filtern', s. 1.) eingesetzt werden. Die Verfahren müssen dazu auf einer zuvor manuell klassifizierten Beispielmenge an Snippets trainiert werden. Formal soll ein Classifier $c(d)$ gelernt werden, der auf Basis der klassifizierten Beispielmenge für ein gegebenes KwiC-Snippet voraussagt, ob dieses einen sprachlichen Ausdruck in der gesuchten Verwendung enthält oder nicht.

Stützvektormethode

Ein für diese Aufgabe geeignetes Verfahren ist die ‚Stützvektormethode' (kurz: ‚SVM'), deren Überlegenheit auch für Aufgaben der Dokumentklassifikation in der Dortmunder Informatik bereits Joachims (1998) gezeigt hat. Die Anwendbarkeit dieses Verfahrens auch auf KwiC-Snippets aus Korpora haben wir bereits in Bartz et al. (2013a, Technischer Bericht 2013/1) gezeigt. Formal wird dabei eine lineare Hyperebene für den Raum gesucht, der durch die bei der Vorverarbeitung (s. 2.1) erzeugte Repräsentation der Snippets aufgespannt ist. Die manuell klassifizierten Trainingsdaten bestimmen die Lage dieser Hyperebene, die so definiert ist, dass sie den Raum der KwiC-Snippets mit dem gesuchten Phänomen vom Raum der KwiC-Snippets ohne das gesuchte Phänomen trennt und möglichst weit von den jeweils am nächsten liegenden Snippets entfernt ist. Dies hat verschiedene Vorteile: Für die exakte Lagebestimmung der Hyperebene werden nicht alle Snippets benötigt, sondern nur die am nächsten liegenden sogenannten ‚Stützvektoren'. Außerdem garantiert der möglichst breite Rand um die Hyperebene, dass auch solche Snippets noch zutreffend klassifiziert werden können, die von den Trainingsdaten geringfügig abweichen.

Verwendet und evaluiert wurde beispielsweise ein binärer Classifier, der definiert ist auf Basis einer linearen Funktion $g(d) = \langle w,\varphi(d)(d)\rangle + b$, wobei w ein Vektor in Raum RN ist, b ein Bias-Term und $\langle .,. \rangle$ das Skalarprodukt in R. Der Classifier ist weiterhin definiert durch $c(d) = 1$, falls $g(d) >= 0$ und $c(d) = -1$, falls $g(d) < 0$. Dabei steht 1 für das Vorhandensein der gesuchten Phänomens und -1 für dessen Nicht-Vorhandensein. Die Aufgabe ist nun, den optimalen Vektor w zu bestimmen. Dieser soll so gewählt werden, dass $g(d) >= 0$ ist für alle Sätze d, die ein Stützverb enthalten, und $g(d)<0$ ist für alle Sätze, die kein Stützverb enthalten. Dazu werden die manuell klassifizierten Trainingsdaten benötigt. Der Vektor w wird so gewählt, dass die Hyperebene $g(d)$ die Menge der Trainingsdaten wie oben verlangt trennt. Weiterhin muss w so gewählt werden, dass die Klassifikation neuer, ungesichteter Snippets mit hoher Wahrscheinlichkeit richtig vorhergesagt wird. Dies kann man bei einer Bags-of-Words-Repräsentation (s.2.1) beispielsweise gewährleisten, wenn die Trainingsdaten im Raum der Wortvektoren, also $\{\varphi(d)\}$, einen maximalen Abstand zu $g(d)$ haben. Details zum Verfahren siehe Cristianini & Shawe-Tayler (2004) und Bartz et al. (2013a, Technischer Report 2013/1).

Aktives Lernen

Die Idee des Aktiven Lernens ist es, das Lernverfahren bei denjenigen Klassifikationsentscheidungen durch manuelles Nachannotieren aktiv zu unterstützen, bei denen auf Basis einer vorhandenen Trainingsdatenmenge noch keine hinreichend sichere Klassifizierung möglich ist. Dabei kommt es darauf an, diejenigen Snippets auszuwählen, von denen die größtmögliche Verbesserung des Lernverfahrens zu erwarten ist.

Im Rahmen des KobRA-Projekts wurde dazu ein sogenanntes konfidenzbasiertes Aktives Lernverfahren implementiert, welches sukzessive Snippets aussucht, die von einer gelernten Support-Vector-Maschine nur mit geringer Verlässlichkeit klassifiziert wurden (vgl. Rüping 2004). Abbildung 2 zeigt einen Screenshot der Aktiven Lernumgebung in RapidMiner (s. 2.4) zur anschaulichen Erläuterung. Links werden die Snippets mithilfe des implementierten „ActiveLearning-Operators“ ausgewählt und mit den bisher schon annotierten Snippets für das Training der Support-Vector-Maschine genutzt. In der Mitte wird diese auf die noch nicht annotierten Snippets angewendet, wobei erneut die Konfidenzwerte bestimmt werden, die den Grad der Verlässlichkeit der Klassifikation angeben. Diese werden dann wieder an den ActiveLearning-Operator übergeben und so weiter. Am Ende wird mittels des „Performance-Operators“ ganz rechts die Güte des Verfahrens bestimmt.

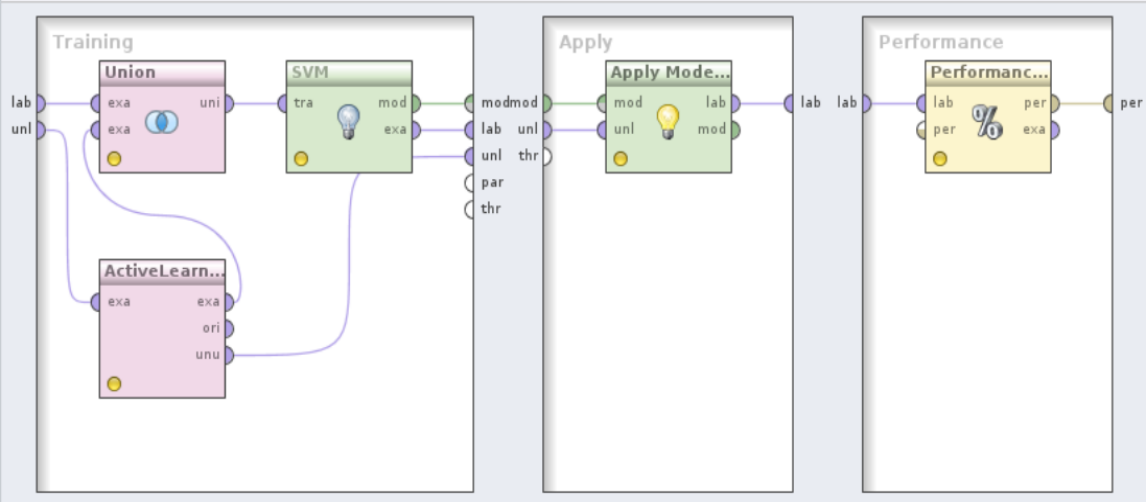

Abbildung 2: Aktives Lernen

2.3 Unüberwachte Verfahren

Unüberwachte Verfahren lassen sich einsetzen, um die im Kontext eines gesuchten sprachlichen Ausdrucks vorhandenen latenten Informationen (z.B. Verteilungen von Kontextwörtern) für eine vollautomatische Disambiguierung seiner Vorkommen in KwiC-Listen zu nutzen (z.B. für die Disambiguierung nach Lesarten bei lexikographisch interessanten Wörtern, s. 3.2 und Bartz et al. 2013b, Technischer Bericht 2013/2). Dabei ist keine manuelle Annotation von Snippets notwendig; diese kann allerdings vorgenommen werden, um das Verfahren zu evaluieren. Formal sollen auf Basis der sprachlichen Merkmale der Snippets einer KwiC-Liste ein Modell für unterschiedliche Verwendungsweisen eines in den Snippets gegebenen sprachlichen Ausrucks gelernt werden, mit dem für ein gegebenes KwiC-Snippet vorausgesagt werden kann, mit welcher Wahrscheinlichkeit dieses einer bestimmten Verwendungsweise zuzuordnen ist.

Topic-Modelle/Latente Dirichlet-Analyse

Die oben formulierte Aufgabe wurde in der Forschung zu Data-Mining-Verfahren vor allem im Bereich der Induktion von Wortbedeutungen schon in zahlreichen Ansätzen bearbeitet. Ein früher statistischer Ansatz wurde bereits 1991 von Brown et al. vorgelegt, einen umfassenden Überblick über den gegenwärtigen Forschungsstand gibt Navigli (2009). Brody und Lapata (2009) konnten zeigen, dass sich mithilfe der Latenten Dirichlet-Analyse (auch ‚Latent-Dirichlet-Allocation', kurz: ‚LDA', vgl. Blei et al. 2003) tendenziell die besten Ergebnisse erzielen lassen. LDA wurde ursprünglich zum thematischen Partitionieren von Dokumentsammlungen genutzt. Navigli und Crisafulli (2010) konnten aber bereits zeigen, dass sich das Verfahren auch für die Disambiguierung kleiner Text-Snippets erfolgreich nutzen lässt, z.B. für das Partitionieren der Trefferlisten von Web-Suchmaschinen. Besonderheiten der Anwendung von LDA auf KwiC-Listen aus Korpora und weitere Details zum Verfahren sind in Batz et al. (2013, Technischer Bericht 2013/2) beschrieben.

Im Rahmen des KobRA-Projekts wurde LDA für die Nutzung in RapidMiner (s. 2.4) implementiert, wie es von Blei et al. (2003) vorgestellt wurde. LDA schätzt die Wahrscheinlichkeitsverteilungen von Wörtern und Dokumenten (hier: KwiC-Snippets) über eine bestimmte Anzahl überzufällig häufig auftretender Kontextwörter, sogenannter ‚Topics', die als Repräsentationen für verschiedene Verwendungsweisen (z.B. Bedeutungen) eines gegebenen sprachlichen Ausdrucks aufgefasst werden. Dabei wird angenommen, dass die Wahrscheinlichkeit für die Zuordnung zu den Topics einer Dirichletverteilung folgt, die von den gegebenen Metaparametern α und β abhängt. Die Wahrscheinlichkeit eines bestimmten Topics für ein gegebenes Snippet ist modelliert als multinomiale Verteilung, die von der Dirichletverteilung der Snippets über die Topics abhängt. Formal sei $\phi \sim Dirichlet(\beta)$ die Wahrscheinlichkeitsverteilung eines Snippets und $p(z_1|\ \phi(j)) \sim Multi(\phi(j))$ die Wahrscheinlich-keit des Topics z_1 für ein gegebenes Snippet j.

Wir verwenden einen Gibbs-Sampler (Griffiths & Steyvers 2004), um die Verteilungen zu schätzen. Der Gibbs-Sampler modelliert die Wahrscheinlichkeitsverteilungen für ein gegebenes Topic z_1 in Abhängigkeit zu allen anderen Topics und den Wörtern eines Snippets als Markov-Reihe. Diese nähert sich der A-posteriori-Verteilung der Topics für die in einem Snippet gegebenen Wörter an. Die A-posteriori-Verteilung kann schließlich genutzt werden, um das wahrscheinlichste Topic für ein gegebenes Snippet zu ermitteln. Auf dieser Basis wird im Rahmen des stochastischen Prozesses die Generierung von Topics simuliert. Abhängig davon, wie häufig ein bestimmtes Topic für ein gegebenes Snippet gezogen wird, ermitteln wir die Wörter, die das Topic am wahrscheinlichsten indizieren. Diese repräsentieren das Topic und damit die Verwendungsweise/Bedeutung des gesuchten Ausdrucks.

Berücksichtigung zeitlicher Entwicklungen

Die Analyse von Aspekten des Sprachwandels über die Zeit ist aus der linguistischen Anwenderperspektive ein besonderer Fokus des KobRA-Projekts. Deshalb wurde das oben beschriebene Verfahren für die Analyse zeitlicher Entwicklungen erweitert. Dafür haben wir zum einen eine Möglichkeit geschaffen, die in den verwendeten Korpora als Metadaten vorhandenen zeitlichen Informationen zu den Snippets (z.B. Veröffentlichungsdatum) unabhängig von den Wahrscheinlichkeitsverteilungen der Topic-Wörter und der Topics über die Snippets auszuwerten. Dies ermöglicht uns, zu erfassen, wie häufig ein bestimmtes Topic des Topic-Modells in einem bestimmten Zeitabschnitt vorkommt. Abbildung 3 zeigt eine solche Verteilung der Topics für das Wort „Platte" über die Zeit (Korpusbasis: DWDS-Kernkorpus des 20. Jh., s. 3.2):

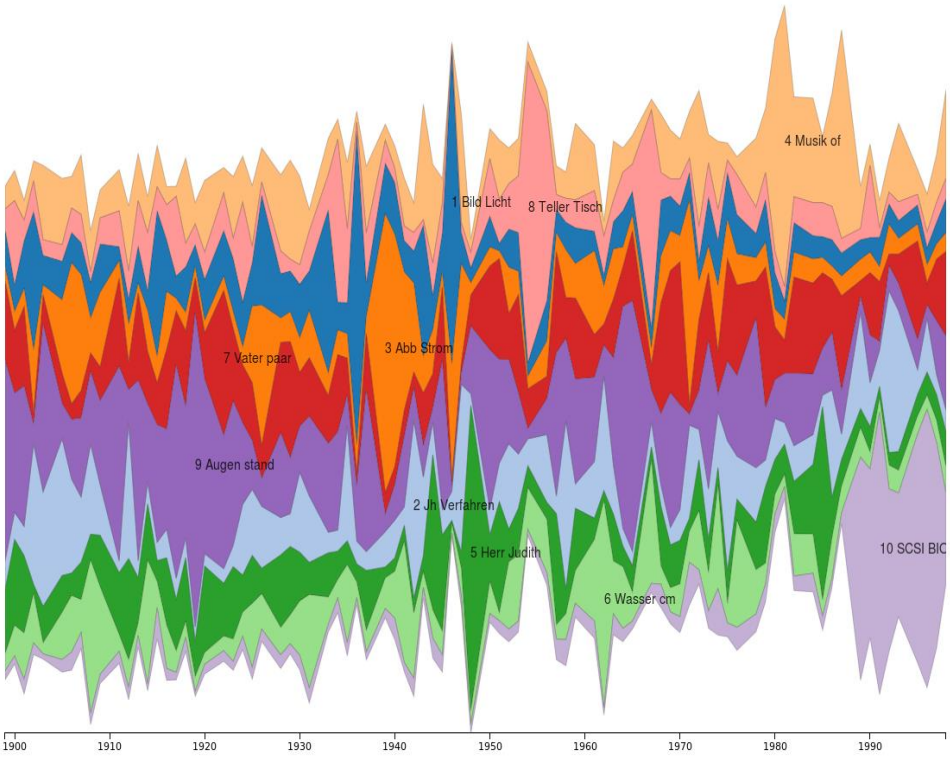

Abbildung 3: Verteilung der LDA-Topics für Snippets mit dem Wort „Platte“ über die Zeit unter der Unabhängigkeitsannahme

Eine weitere Möglichkeit zur Integration von zeitlichen Informationen in Topic-Modelle ist die Modellierung von Zeit explizit als Zufallsvariable (Wang & McCallum 2006). Dafür nehmen wir an, dass die Zeit eine Beta-verteilte Zufallsvariable ist und die Wahrscheinlichkeit, dass ein bestimmtes Wort in einem Snippet zu einem Topic gehört, auch von dieser Variable abhängig ist. Dies hat vor allem den Vorteil, dass wir die Zeit nicht in Intervalle einteilen müssen, sondern dynamische Perioden der Topics modellieren können. Abbildung 4 zeigt die Verteilung der Topics für das Wort „Platte“ über die Zeit, wenn Zeit als abhängige Beta-verteilte Zufallsvariable modelliert wird (gleiche Korpusbasis: DWDS-Kernkorpus des 20. Jh., s. 3.2). Im Vergleich zu Abbildung 3 sieht man sehr schön, dass wir nun die Topics über die Zeit viel eindeutiger trennen können.

Visualisierungen wie die Abbildungen 3 und 4 wurden mithilfe des Werkzeugs „dfr-browser“ (Goldstone o.J.) generiert, das die Entwicklung von Topics über die Zeit und auch die Verteilung von Kontextwörtern und Snippets über die Topics veranschaulichen kann. Eine Schnittstelle zum Visualisierungswerkzeug wurde für die Nutzung in RapidMininer implementiert (s. 2.4).

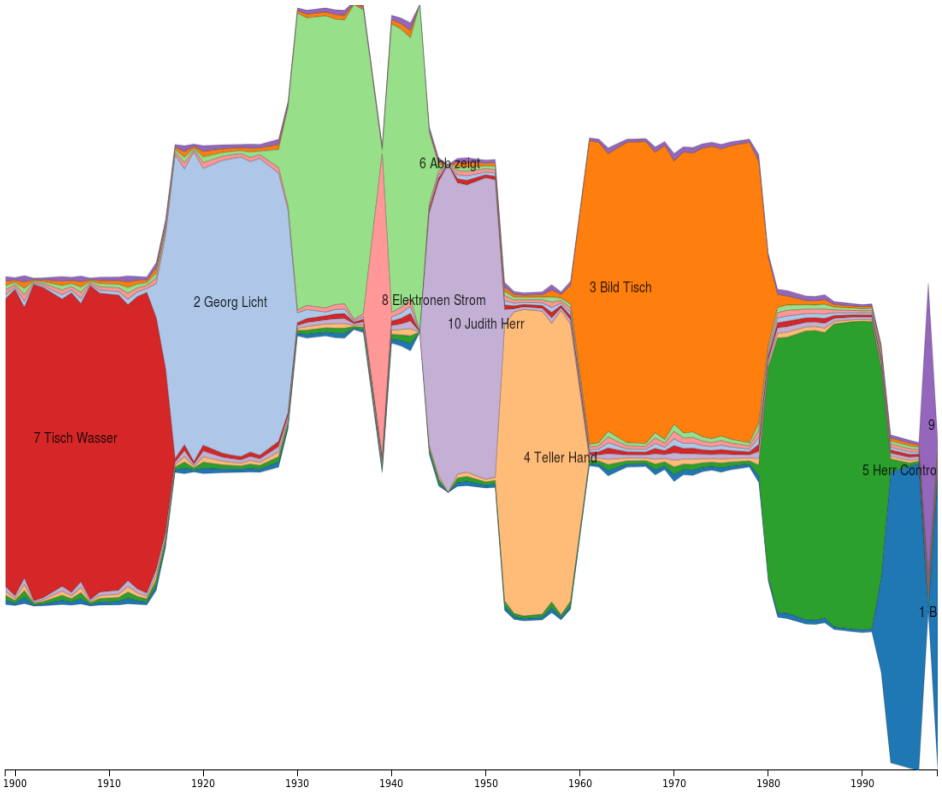

Abbildung 4: Verteilung der LDA-Topics für Snippets mit dem Wort „Platte“ über die Zeit unter der Abhängigkeitsanahme

2.4 Erweiterung der Data-Mining-Umgebung „RapidMiner“

Alle im KobRA-Projekt implementierten und evaluierten Verfahren und Werkzeuge sind als Plug-in für die Data-Mining-Umgebung „RapidMiner“ (früher „YALE“, Mierswa et al. 2006) verfügbar. RapidMiner ermöglicht auf einfache Weise die Ausführung vielfältiger, leistungsfähiger Methoden zur Analyse großer Datenmengen und enthält standardmäßig bereits eine Vielzahl von Werkzeugen für Datenimport, -transformation, -analyse und -visualisierung.

Im KobRA-Projekt wurden neben den oben bereits beschriebenen Data-Mining-Verfahren für die Klassifikation (s. 2.2) und das Partitionieren (s. 2.3) von Daten zusätzlich Methoden implementiert, die einen effizienten Zugriff auf die im Projekt verfügbaren Sprachressourcen (s. 1) und die Extraktion sowie Analyse von Dokument- und sprachlichen Merkmalen ermöglichen. Eine integrierte Annotationsumgebung erlaubt Korpus-Nutzern, ihre Expertise durch Annotation von Daten direkt aus der Data-Mining-Umgebung heraus in maschinelle Lernprozesse einzubringen, z.B. in Szenarien des Aktiven Lernens (s. 2.2). Eine Schnittstelle zur CLARIN-Annotationsumgebung „WebLicht“ (Hinrichs et al. 2010) eröffnet Nutzern die Möglichkeit, alle automatischen Sprachverarbeitungswerkzeuge zur Anreicherung der Daten zu verwenden, die über die CLARIN-Infrastruktur verfügbar sind. Eine weitere Schnittstelle zu einem leistungsfähigen Visualisierungswerkzeug (Goldstone o.J.) erschließt aktuelle Verfahren zur visuellen Aufbereitung der Analyseergebnisse. Abbildung 5 zeigt eine Auswahl der zur Verfügung gestellten Werkzeuge in der Anwendung in einem Prozess zur automatischen Disambiguierung von Korpusbelegen zum Adjektiv „toll“, wobei das Ergebnis der Disambiguierung an einer manuell annotierten Stichprobe direkt evaluiert wird.

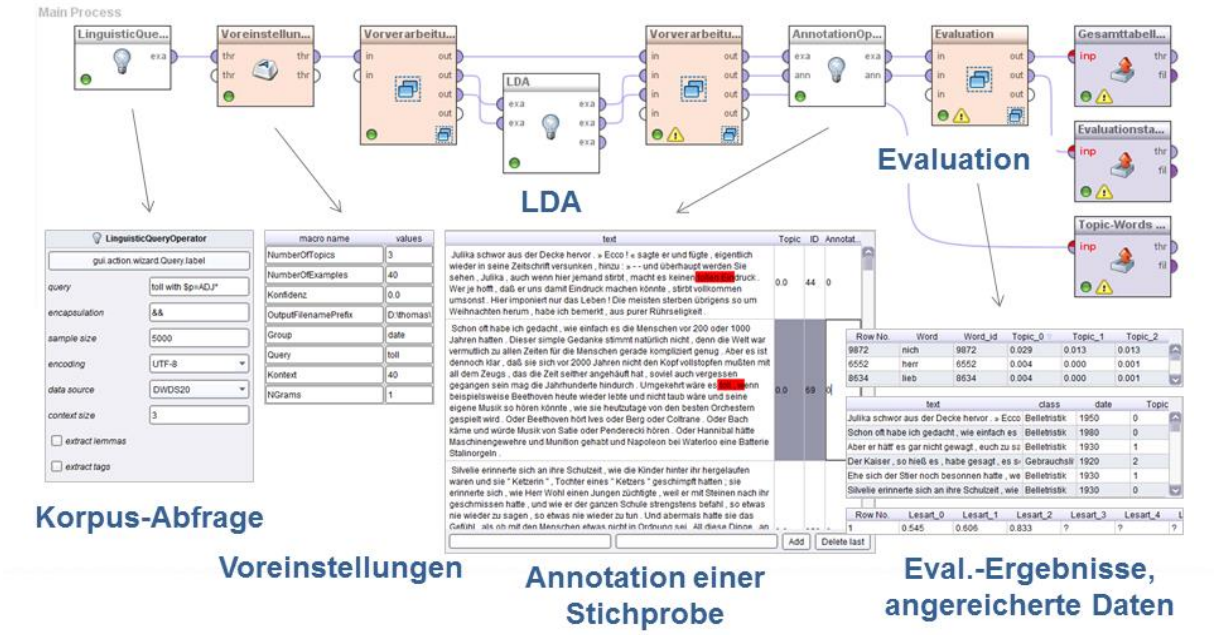

Abbildung 5: Operatoren des KobRA-Plug-ins im Einsatz: „LinguisticQuery-Operator“, „LDA-Operator“, „Annotation-Operator“

3. Evaluation der Verfahren in linguistischen Fallstudien

Die im Projekt bereitgestellten Verfahren wurden allesamt an linguistischen Fallstudien evaluiert, die aus konkreten aktuellen Forschungsvorhaben der Projektbeteiligten stammen. Es ist zu erwarten, dass der Nutzwert und noch offene Anforderungen an die Verfahren, die sich in den Fallstudien gezeigt haben, für viele Korpus-basierte Studien ebenfalls zu erwarten sind.

Im Folgenden werden die überwachten und unüberwachten Verfahren (s. 2.2 und 2.3) an je einer Fallstudie exemplarisch evaluiert. Für weitere Fallstudien verweisen wir auf die zwei vorangehenden Technischen Berichte Bartz et al. (2013a, Technischer Bericht 2013/1; 2013b, Technischer Bericht 2013/2) und unsere zahlreichen projektbegleitenden Veröffentlichungen.

3.1 Sprachliche Besonderheiten auf den Diskussionsseiten der deutschsprachigen Wikipedia (KobRA-Anwendungsbereich Varietätenlinguistik)

Korpus-basierte Studien zu sprachlichen Besonderheiten internetbasierter Kommunikation (kurz: ‚IBK‘) sind bislang noch sehr aufwändig oder nur an sehr kleinen Stichproben möglich, weil aktuell noch kaum strukturierte Korpora mit Sprachdaten aus IBK-Genres vorhanden sind. Forscher müssen in den meisten Fällen selbst Korpora zusammenstellen und zudem die Daten meist manuell analysieren, da die gängigen automatischen Sprachverarbeitungswerkzeuge auf den IBK-Daten nicht zufriedenstellend funktionieren. Ziel der im Folgenden beschriebenen Fallstudie ist es, Data-Mining-Verfahren zu erproben, die den Korpus-Nutzer dabei unterstützen, aus einer Menge von Snippets genau diejenigen auszuwählen, die Belege für ein relevantes sprachliches Phänomen enthalten; in diesem Fall: 1) Aktionswörter und 2) nicht-kanonische Verwendungen von „weil“ (Details s. Beißwenger et al. 2014).

Aktionswörter

Aktionswörter sind einzelsprachlich gebundene symbolische Einheiten, die strukturell auf einem Wort basieren – zumeist einem unflektierten Verbstamm (‚Inflektiv‘) –, das entweder alleine steht oder um weitere Einheiten erweitert sein kann – im Falle von Inflektiven um vom

Verb geforderte Ergänzungen sowie Angaben (z.B. „grins“, „lach“, „freu“, „lautlach“, „diabolischgrins“, „kopf schüttel“, „schulterzuck“, „nachlinksrutsch“). Im Falle solcher Konstruktionen werden die einzelnen Wortformen sehr häufig zusammengeschrieben, so dass sie formal als ein Token erscheinen; sehr verbreitet ist zudem die Markierung mit ein- und ausleitenden Asterisken (z.B. *grins*, *freu*). Aktionswörter werden zur Beschreibung von Gesten, mentalen Zuständen oder Handlungen verwendet. Sie dienen als Emotions- oder Illokutionsmarker, als Ironiemarker, zur spielerischen Nachbildung fiktiver Handlungen oder dazu, sich selbst (oder dem eigenen virtuellen Charakter) Charaktermerkmale oder innere Zustände zuzuschreiben. Einige sehr gebräuchliche Aktionswörter haben die Form von Akronymen („lol“, „rofl“, „g“).

Nicht-kanonische Verwendungen von „weil“

Im geschriebenen Deutsch in redigierten, am schriftsprachlichen grammatischen Standard orientierten Texten wird die Konjunktion „weil“ als Einleitung für untergeordnete Nebensätze verwendet mit Endstellung des finiten Verbs verwendet. Unter Bedingungen der konzeptionellen Mündlichkeit erscheint „weil“ jedoch ebenso im Vor-Vorfeld von Sätzen, in denen das finite Verb nicht am Satzende steht, z.B.: „ja toll aber so richtig steht es nicht drin weil damals sollten wir nämlich eine arbeit in informatik machen über das dualsystem”, „Ja ich bin auch 96 Fan aber trotzdem, er hätte auch im Spiel sein fehler noch ändern können. Weil ich bin selber Schiedsrichter, und hatte auch schon so eine Situation”. Die Forschung im Feld der gesprochenen Sprache konnte zeigen, dass solche nicht-kanonischen Verwendungen Funktionen erfüllen, die sich von denen der kanonischen Verwendungen unterscheiden (vgl. Gohl & Günthner 1999; Günthner & Auer 2005; Imo 2012). Es ist eine offene Frage, inwiefern nicht-kanonische Verwendungen von „weil“ in der internetbasierten Kommunikation ähnliche Funktionen erfüllen wie entsprechende Verwendungen in der gesprochenen Sprache.

Korpusbasis

Als Datenbasis für die Erprobung der Data-Mining-Verfahren werden KwiC-Ergebnislisten 1) zu den Aktionswörtern „lol“, „lach“, „freu“, „grins“, „wink“ und „seufz“ und 2) zu allen Vorkommen von „weil“ aus dem deutschen Wikipedia-Korpus (Datenstand: 2013) erhoben, das als Bestandteil von DeReKo (Deutsches Referenzkorpus, Kupietz & Lüngen 2014) vom Projektpartner am Institut für Deutsche Sprache in Mannheim bereitgestellt wird. Die KwiC-Snippets haben jeweils eine Größe von maximal 999 Zeichen. Als Trainingsdatenmenge wurden zu 1) 600 Snippets von zwei unabhängigen Annotatoren manuell klassifiziert (Klasse 1: Snippet enthält ein Aktionswort, Klasse 0: Snippet enthält kein Aktionswort; Inter-Annotator-Agreement: 0,900 Kappa, vgl. Cohen 1960), zu 2) 1200 Snippets (Klasse 1: nicht-kanonische Verwendung, Klasse 0: kanonische Verwendung; Inter-Annotator-Agreement: 0,971 Kappa).

Experimente und Evaluation

Für die Klassifikation von 1) Formen wie „lach“, „freu“, „grins“ etc. als Aktionswörter bzw. Nicht-Aktionswörter und 2) Formen von „weil“ als nicht-kanonische bzw. kanonische Verwendungen wird jeweils ein überwachtes Verfahren eingesetzt, wie es unter 2.2. beschrieben ist. Zur Anwendung kommt jeweils eine SVM, die unter Verwendung verschiedener Kernels trainiert wird (Tree-Kernel für syntaktische Informationen, Moschitti 2006; Substring-Kernel für Zeichenketten, Lodhi et al., 2002; linearer Kernel für Bags-of-Words, s. 2.2). Für Training und Evaluation des Verfahrens wurden die manuell klassifizierten Daten herangezogen, wobei das Training auf einem anderen Teil des Datensatzes ausgeführt wurde als die Evaluation. Mittels einer 10-fachen Kreuzvalidierung wurde der F_1-Wert als Standardmaß für die Leistungsfähigkeit der Verfahren ermittelt. Der F_1-Wert gibt das gewichtete harmonische Mittel

aus Präzision (Precision) und Ausbeute (Recall) an, wobei Präzision und Ausbeute gleich gewichtet werden (Navigli & Vanella 2013).

Ergebnisse

Die Evaluation der oben genannten Verfahren ergab bereits zufriedenstellende Ergebnisse für die Klassifikation von 1) Aktionswörtern vs. Nicht-Aktionswörtern; noch nicht zufriedenstellend ist hingegen das Ergebnis für die Klassifikation von 2) nicht-kanonischen vs. kanonischen Verwendungen von „weil“. In allen Treatments erwies sich das Bags-of-Words-Verfahren bislang als das erfolgreichste (Ergebnisse siehe Tabellen 1 und 2). Dies mag damit zusammenhängen, dass morphosyntaktische/syntaktische Merkmale in Sprachdaten aus Genres internetbasierter Kommunikation mit den bislang verfügbaren automatischen Sprachverarbeitungswerkzeugen noch nicht zufriedenstellend automatisch analysierbar sind (vgl. Bartz et al. 2014).

Klasse	F_1
1: Snippet enthält ein Aktionswort	89,493
0: Snippet enthält kein Aktionswort	76,247

Tabelle 1: Evaluationsergebnis für die Klassifikation Aktionswort vs. Nicht-Aktionswort, Treatment: Bags-of-Words

Klasse	F_1
1: Snippet enthält nicht-kanonische Verwendung von „weil“	21,679
0: Snippet enthält kanonische Verwendung von „weil“	76,505

Tabelle 2: Evaluationsergebnis für die Klassifikation nicht-kanonisches „weil“ vs. kanonisches „weil“, Treatment: Bags-of-Words

3.2 Entwicklung und Ausdifferenzierung von Bedeutungen (KobRA-Anwendungsbereich Lexikographie)

Die Entwicklung und Ausdifferenzierung von Bedeutungen ist für Linguisten in zweierlei Hinsicht interessant: Lexikographen verfolgen Wortentwicklungen, um adäquate lexikographische Beschreibungen erstellen bzw. vorhandene Wörterbucheinträge aktualisieren zu können (Storrer, 2011). Forscher im Bereich der Historischen Semantik fragen nach den Möglichkeiten, Bedingungen und Folgen semantischer Innovationen (Fritz, 2012; Fritz 2005; Keller & Kirschbaum 2003). Für den Erkenntnisgewinn entscheidend ist in beiden Fällen die Verfügbarkeit strukturierter Textkorpora, die es erlauben, die Verwendung eines Wortes über größere Zeiträume hinweg nachzuvollziehen. Während insbesondere im Rahmen von CLARIN umfangreiche synchrone und diachrone Textkorpora mit Metadaten zu Erscheinungsdatum und Textsorte sowie komfortable Abfrage- und Analysewerkzeuge zur Verfügung stehen, ist die großflächige automatische semantische Annotation der Korpora nach gegenwärtigem Stand der Technik noch nicht zufriedenstellend möglich (Rayson & Stevenson, 2008). Bei der Korpus-basierten Untersuchung von Bedeutungswandel müssen deshalb bislang die zu einem Wort gefundenen Einzelbelege manuell disambiguiert werden. Verbreitung und Prozesse des Bedeutungswandels können daher aktuell lediglich anhand weniger Beispiele und auf einer vergleichsweise geringen Datenbasis beschrieben werden (Fritz 2005; Keller & Kirschbaum 2003). Ziel der im Folgenden beschriebenen Fallstudie ist es, Data-Mining-Verfahren zu erproben, die den Korpus-Nutzer dadurch unterstützen, dass sie eine Menge von Snippets zu einem lexikographisch interessanten Wort nach den Bedeutungen dieses Wortes partitionie-

ren, sodass die Snippets für einzelne Bedeutungen auch einzeln zählbar sind (Details s. Bartz et al. im Erscheinen).

Auswahl der Wörter und Korpora

Wir haben Korpusabfragen zu einer Auswahl an Wörtern gestellt, die aus linguistischer Sicht interessant sind, weil sie in letzter Zeit oder über eine längere Zeitspanne hinweg neue Bedeutungen entwickelt oder ihre prototypische Bedeutung gewechselt haben. Je nach angenommenem Zeitraum der Bedeutungsveränderungen wurden unterschiedliche Korpora abgefragt. Bei der Auswahl der Beispielwörter haben wir zudem unterschiedliche Wortarten berücksichtigt, um auch Einsichten in mögliche wortartenspezifische Unterschiede in der Leistungsfähigkeit der evaluierten Data-Mining-Verfahren zu erhalten. Folgende Beispielwörter bilden die Basis für die unten dargestellten Experimente. Details zu den verwendeten Korpora finden sich direkt im Anschluss.

Das Substantiv „Platte“ hat im Zuge technischer Innovationen im Laufe des 20. Jahrhunderts sein Bedeutungsspektrum stark ausdifferenziert. Neben den Bedeutungen *flaches Werkstück* oder *Teller* finden sich nach und nach zunehmend auch Verwendungen in den Bedeutungen *fotografische Platte*, *Schallplatte/CD* oder Festplatte. Eine Suche nach dem Lemma „Platte“ im DWDS-Kernkorpus des 20. Jh. ergibt 2886 KwiC-Snippets.

Das Verb „anrufen“ hat mit Beginn der kommerziellen Verbreitung des Telefons in den 20er/30er Jahren des 20. Jahrhunderts neben seiner ursprünglichen Bedeutung *rufen*/*bitten* auch die Bedeutung *telefonieren* erhalten. Eine Suche nach dem Verb „anrufen“ im DWDS-Kernkorpus des 20. Jh. ergibt 2085 KwiC-Snippets.

Das Substantiv „Heuschrecke“ scheint spätestens seit der Finanz- und Bankenkrise (ab 2007) neben seiner prototypischen Bedeutung *Grashüpfer* auch als Bezeichnung für eine am sogenannten “Heuschreckenkapitalismus” beteiligte *Person* verwendet zu werden. Eine Suche nach „Heuschrecke“ im DWDS-Zeitungskorpus ‚Die ZEIT‘ ergibt 715 KwiC-Snippets.

Das Adjektiv „zeitnah“ scheint in Laufe der letzten 20 bis 30 Jahre neben seiner ursprünglichen Bedeutung *zeitgenössisch*/*zeitkritisch* eine neue prototypische Bedeutung erhalten zu haben: *unverzüglich*. Eine Suche nach „zeitnah“ im DWDS-Zeitungskorpus ‚Die ZEIT‘ ergibt 597 KwiC-Snippets.

Das Adjektiv „toll“ hat im Laufe der letzten Jahrhunderte einen bemerkenswerten Bedeutungswandel durchlaufen, wobei sich die ursprüngliche Bedeutung *irre* über *ausgelassen*/*wild* bis hin zum positiv attribuierenden *sehr gut* wandelte. Eine Suche nach dem Adjektiv „toll“ in der Tübingen Baumbank des Deutschen Diachron (TüBa-D/DC) ergibt 5793 KwiC-Snippets, eine entsprechende Suche im DWDS-Kernkorpus des 20. Jh. 1745 KwiC-Snippets.

Die Konjunktion „da“ wurde nach frühen Belegen zunächst ausschließlich in temporaler Bedeutung genutzt, heute finden sich häufiger Belege in kausaler Verwendung. Eine Suche nach der Konjunktion „da“ in der Tübingen Baumbank des Deutschen Diachron (TüBa-D/DC) ergibt 123496 KwiC-Snippets.

Mit der Auswahl des englischen Substantivs „cloud“ soll schließlich ein erster Eindruck zur Anwendbarkeit des Verfahrens auch auf nicht deutsche Sprachdaten gewonnen werden. Das Wort scheint mit der Entstehung großer Computernetzwerke in den letzten Jahrzehnten neben seiner ursprünglichen Bedeutung *Wolke* eine neue Bedeutung entwickelt zu haben. Eine Suche nach „cloud“ in den Korpora der Leipzig Corpora Collection ergibt 1486 KwiC-Snippets.

Das DWDS-Kernkorpus des 20. Jh. (DWDS-KK), das an der Berlin-Brandenburgischen Akademie der Wissenschaften gepflegt wird, enthält ca. 100 Millionen laufende Wörter, die ausgewogen über die Dekaden des 20. Jh. und die Textsortenbereiche Belletristik, Zeitung, Wissenschaft und Sachtexte verteilt sind. Das Zeitungskorpus ‚Die ZEIT' (ZEIT) umfasst alle Ausgaben der gleichnamigen Wochenzeitung von 1946 bis 2009, ca. 460 Millionen laufende Wörter (Klein & Geyken, 2010; Geyken, 2007).

Die Tübingen Baumbank des Deutschen Diachron (TüBa-D/DC) ist ein syntaktisch annotiertes Korpus (Konstituentenbäume) mit ausgewählten diachronen Sprachdaten aus dem deutschen Gutenberg-Projekt (http://gutenberg.spiegel.de/); dabei handelt es sich um eine Initiative einer Gemeinschaft von Interessierten, die Copyright-freie Literatur von 1210 bis 1930 über eine Web-Schnittstelle öffentlich zugänglich macht. Die TüBa-D/DC wird vom CLARIN-D-Center an der Universität Tübingen gepflegt und enthält etwa 250 Millionen laufende Wörter (Hinrichs and Zastrow, 2012).

Die Leipzig-Corpora-Collection (LCC) besteht aus Korpora für verschiedene Sprachen, die zufällig ausgewählte Sätze aus Zeitungstexten und einer Web-Stichprobe enthalten (Quasthoff, Richter & Biemann, 2006). Für diese Fallstudie haben wir das englischsprachige Korpus mit Sprachdaten aus Zeitungstexten und der englischen Wikipedia verwendet, das eine Zeitspanne von 2005 bis 2010 abdeckt.

Die Korpusabfragen ergeben KwiC-Snippets mit Vorkommen der untersuchten Wörter (einschließlich ihrer flektierten Formen) in einem Kontext von bis zu drei Sätzen (von bis zu einem Satz bei den Daten aus der LCC). Zusätzlich werden für jedes Snippet das Veröffentlichungsdatum sowie weitere Metadaten (bei der TüBa-D/DC: Publikationstitel und Autorname; beim DWDS-KK: Textsortenbereiche) ausgegeben.

Experimente und Evaluation

Für die automatische Disambiguierung der KwiC-Snippets zu den untersuchten Beispielwörtern wird jeweils ein unüberwachtes Verfahren eingesetzt, wie es unter 2.3. beschrieben ist. Zur Anwendung kommt jeweils das LDA-Verfahren, das in acht verschiedenen Treatments evaluiert wird, die sich durch die Auswahl der Beispielwörter und Korpora (s.o.) sowie unser Erkenntnisinteresse in Bezug auf die optimale Repräsentation der KwiC-Snippets ergeben. Die Treatments unterscheiden sich hinsichtlich folgender Aspekte:

1) **Abgefragtes Wort und Wortart:** Substantiv, Verb, Adjektiv oder Konjunktion?
2) **Menge der Bedeutungen:** Zwei oder mehr Bedeutungen?
3) **Abgefragtes Korpus:** Gegenwartssprachlich (DWDS-KK, ZEIT) oder diachron (TüBa-D/DC)?
4) **Sprache des Korpus:** Deutsch oder Englisch?
5) **Menge der KwiC-Snippets:** Weniger oder mehr als 1000 Snippets?

Für jedes Treatment wurde zudem überprüft, ob ein Kontext von 20, 30 oder 40 Wörtern um das zu disambiguierende Wort zu den besten Ergebnissen führt. Die folgende Tabelle 3 zeigt eine Übersicht über die Evaluations-Treatments:

Treatment	Wort	Wortart	Bedeutungen	Korpus	Sprache	Snippets	Kontext		
							20	30	40
1	Platte	Substantiv	5	gegenwartssprachlich	deutsch	> 1000	X	X	X
2	toll	Adjektiv	3	gegenwartssprachlich	deutsch	> 1000	X	X	X
3	anrufen	Verb	2	gegenwartssprachlich	deutsch	> 1000	X	X	X
4	Heuschrecke	Substantiv	2	gegenwartssprachlich	deutsch	< 1000	X	X	X
5	zeitnah	Adjektiv	2	gegenwartssprachlich	deutsch	< 1000	X	X	X
6	toll	Adjektiv	2	diachron	deutsch	> 1000	X	X	X
7	da	Konjunktion	2	diachron	deutsch	> 1000	X	X	X
8	cloud	Substantiv	3	gegenwartssprachlich	englisch	> 1000	X	X	X

Tabelle 3: Treatments für die Evaluation der unüberwachten Verfahren zur Disambiguierung

Für die Evaluation wurden jeweils 30% der für die untersuchten Wörter erhobenen KwiC-Snippets von zwei unabhängigen Annotatoren manuell disambiguiert. Tabelle 4 zeigt das erreichte Inter-Annotator-Agreement (kappa: Cohen, 1960):

Treatment	Wort	IAA
1	Platte	0,82
2	toll	0,76
3	anrufen	0,97
4	Heuschrecke	0,98
5	zeitnah	0,91
6	toll	0,71
7	da	0,75
8	cloud	0,92

Tabelle 4: Inter-Annotator-Agreement für die manuelle Disambiguierung durch zwei unabhängige Annotatoren

Das Disambiguierungsverfahren wurde auf Basis der manuell annotierten Datensätze evaluiert. Dazu wurden Topic-Modelle (s. 2.3) generiert, um die verschiedenen Bedeutungen der Vorkommen der untersuchten Wörter automatisch zu bestimmen. Diese wurden mit den Bedeutungszuweiseungen verglichen, die die Annotatoren manuell vorgenommen haben. Als Maß für die Zuverlässigkeit der automatischen Disambiguierung haben wir jeweils den F_1-Wert bestimmt. Der F_1-Wert gibt das gewichtete harmonische Mittel aus Präzision (Precision)

und Ausbeute (Recall) an, wobei Präzision und Ausbeute gleich gewichtet werden (Navigli & Vanella 2013; s. auch 3.1).

Ergebnisse

Die folgenden Tabellen 5-12 zeigen die mit dem oben beschriebenen Verfahren erzielten Ergebnisse:

„Platte“		flaches Werkstück	Teller	fotografische Platte	Schallplatte/CD	Festplatte
F_1 für Kontext (Wörter)	20	0,800	0,800	0,667	0,287	0,857
	30	0,998	0,875	0,500	0,381	0,988
	40	0,733	0,600	0,750	0,353	0,800

Tabelle 5: Ergebnisse für Treatment 1

„toll“		irre	ausgelassen/wild	sehr gut
F_1 für Kontext (Wörter)	20	0,519	0,571	0,167
	30	0,714	0,615	0,632
	40	0,625	0,667	0,500

Tabelle 6: Ergebnisse für Treatment 2

„anrufen“		rufen/bitten	telefonieren
F_1 für Kontext (Wörter)	20	0,727	0,667
	30	0,800	0,800
	40	0,909	0,889

Tabelle 7: Ergebnisse für Treatment 3

„Heuschrecke“		Grashüpfer	Person
F_1 für Kontext (Wörter)	20	0,857	0,842
	30	0,800	0,933
	40	0,667	0,727

Tabelle 8: Ergebnisse für Treatment 4

„zeitnah“		unverzüglich	zeitgenössisch/zeitkritisch
F_1 für Kontext (Wörter)	20	0,727	0,667
	30	0,888	0,800
	40	0,895	0,818

Tabelle 9: Ergebnisse für Treatment 5

„toll“		irre	ausgelassen/wild
F_1 für Kontext (Wörter)	20	0,526	0,571
	30	0,625	0,750
	40	0,556	0,636

Tabelle 10: Ergebnisse für Treatment 6

„da“		temporal	kausal
F_1 für Kontext (Wörter)	**20**	0,471	0,556
	30	0,353	0,529
	40	0,400	0,611

Tabelle 11: Ergebnisse für Treatment 7

„cloud“		Wolke	Netzwerk	Name
F_1 für Kontext (Wörter)	**20**	0,526	0,500	0,471
	30	0,783	0,631	0,615
	40	0,467	0,545	0,684

Tabelle 12: Ergebnisse für Treatment 8

Die Evaluation zeigt, dass die avisierte Aufgabenstellung der automatischen Disambiguierung von KwiC-Snippets aus Korpusabfragen mit dem oben beschriebenen Ansatz (s. 2.3) zu überwiegend zufriedenstellenden Ergebnissen führt. In den günstigsten Treatments liegen die F_1-Werte für die Zuverlässigkeit des Verfahrens im Durchschnitt bei 0,732. Je nach untersuchtem Wort und gewünschter Bedeutung variieren die Werte allerdings zum Teil relativ stark in einem Bereich zwischen 0,381 und 0,998 (wiederum im günstigsten Treatment). Generelle Aussagen über die Leistungsfähigkeit des Verfahrens sind also nur schwer möglich. Abhängig von den oben formulierten systematischen Unterschieden der Treatments lassen sich aber folgende Trends feststellen:

Wortart

Den untersuchten Beispielen zufolge scheint die automatische Disambiguierung bei Substantiven, Verben und Adjektiven grundsätzlich mit ähnlichem Erfolg möglich zu sein. Bei „Heuschrecke“ (Tabelle 8) erzielte das Verfahren ebenso gute Werte wie bei „zeitnah“ (Tabelle 9) oder „anrufen“ (Tabelle 7). Die Spitzenwerte wurden jedoch allesamt bei Substantiven (s. auch Tabelle 5) erreicht. Die feineren Bedeutungsunterschiede bei der Konjunktion „da“ ließen sich nicht zufriedenstellend erkennen (Tabelle 11). Erfolgversprechend ist das Verfahren also vor allem bei Inhaltswörtern. Dies ist aufgrund ihrer semantisch referenzierenden Funktion auch erwartbar. Die Eignung bei grammatischen Funktionswörtern muss in zusätzlichen Studien weiter untersucht werden.

Anzahl der Bedeutungen

Hingegen scheint die Anzahl der Bedeutungen bei den untersuchten Beispielen die Ergebnisse systematisch zu beeinflussen. Bei den Beispielen „toll“ (Tabelle 6) und „cloud“ (Tabelle 12) erzielte das Verfahren schlechtere Ergebnisse als bei den Beispielen mit nur zwei Bedeutungen. Dies trifft auch für einzelne Lesarten des Beispiels „Platte“ (see Table 5) zu, während für andere jedoch Spitzenwerte erreicht wurden. Grundsätzlich scheinen unterschiedliche Bedeutungen unterschiedlich gut erkennbar zu sein.

Korpus und Sprache

Die ausgewählten Korpora (gegenwärtiges Deutsch vs. diachron, Deutsch vs. Englisch) scheinen grundsätzlich für die Aufgabe der automatischen Disambiguierung ähnlich gut geeignet zu sein. Die Ergebnisse für die Snippets zu „toll“ aus dem DWDS-KK (Tabelle 6) sind mit denen aus der TüBa-D/DC (Tabelle 10) etwa vergleichbar; dies gilt auch für die Ergebnisse zum englischen Beispiel “cloud” (Tabelle 12). Dieses Evaluationsergebnis ist insofern er-

wartbar, als die Texte der diachronen TüBa-D/DC in orthographisch normalisierter Form vorliegen. Um die Leistungsfähigkeit des Verfahrens auch für diachrone Korpora mit orthographisch nicht normalisierten Sprachdaten überprüfen zu können, sind weitere Studien notwendig.

Anzahl an Snippets und Größe des Kontexts

Während die Anzahl der vom Verfahren genutzten KwiC-Snippets (500-1000 vs. 1000-5000) für die untersuchten Beispiele keine systematischen Auswirkungen auf das Ergebnis zu haben scheint – „zeitnah“ (Tabelle 9) und „Heuschrecke“ (Tabelle 8) werden ähnlich gut disambiguiert wie „Platte“ (Tabelle 5), „toll“ (Tabelle 10) oder „anrufen“ (Tabelle 7) – erweist sich für die Größe des Kontexts ein Umfang von 30 Wörtern vor und nach dem untersuchten Wort in den meisten Fällen als ideal. Beim Verb „anrufen“ (Tabelle 7) scheint jedoch der größte Kontext am erfolgversprechendsten zu sein. Dies könnte damit zusammenhängen, dass das Verb in seiner Funktion eher auf den Satz als größere Einheit bezogen ist, während Substantive und Adjektive bereits im näheren Kontext spezifiziert werden. Dafür sprechen auch die leicht besseren Ergebnisse beim hauptsächlich adverbiell gebrauchten „zeitnah“ (Tabelle 9) im Treatment mit einem Kontext von 40 Wörtern.

Anwendbarkeit im Rahmen der Forschung zum Bedeutungswandel

Nach der automatischen Disambiguierung lassen sich auf einfachem Wege die Häufigkeiten der einzelnen Bedeutungen der untersuchten Wörter ermitteln und visualisieren. Die Abbildungen 6-10 veranschaulichen den Nutzen der Integration zeitlicher Informationen beim Generieren der Topic-Modelle: Forscher können auf dieser Basis leicht die Entwicklung disambiguierter lexikalischer Einheiten über die Zeit verfolgen:

„Platte“

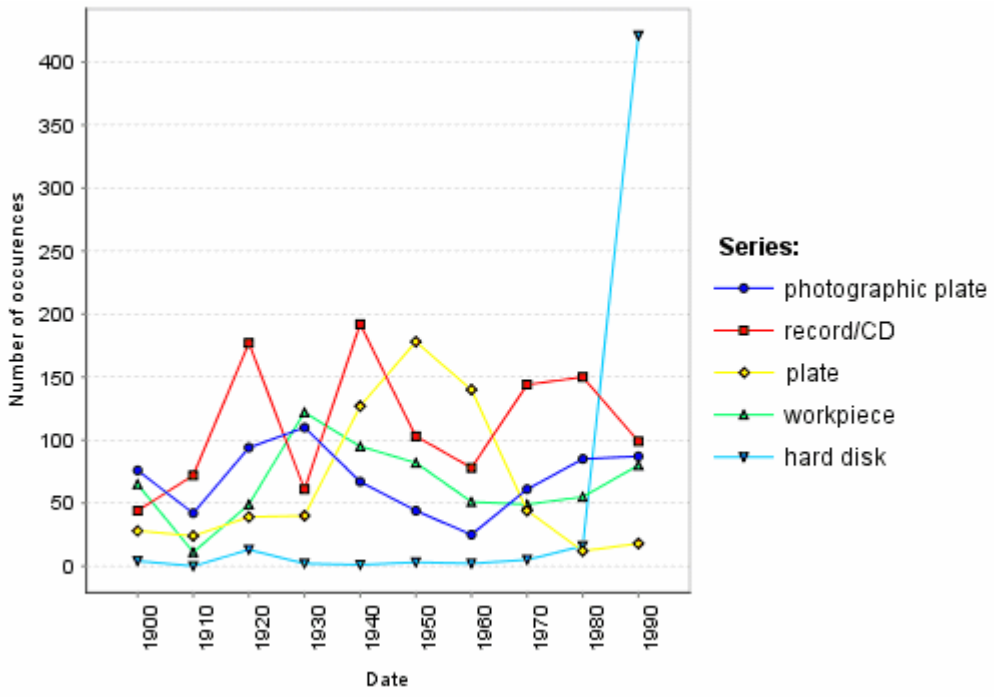

Abbildung 6: Vorkommen des Wortes „Platte“ mit seinen unterschiedlichen Bedeutungen in den Dekaden des 20. Jh.

Abbildung 6 veranschaulicht nachvollziehbar die Bedeutungsentwicklung von „Platte“. Die Bedeutung *Festplatte* wird in den 90er Jahren sprunghaft frequent, während sich die anderen Bedeutungen bei einzelnen Phasen häufigerer Verwendung auf einem einigermaßen gleichbleibenden Niveau bewegen. Die Phasen häufigerer Verwendung (z.B. in der Bedeutung *Teller* in den 40er bis 60er Jahren oder in der Bedeutung *fotografische Platte* in den 80er/90er Jahren) bieten Anlass für genauere Untersuchungen unter Berücksichtigung der zugrundeliegenden KwiC-Snippets.

„toll“

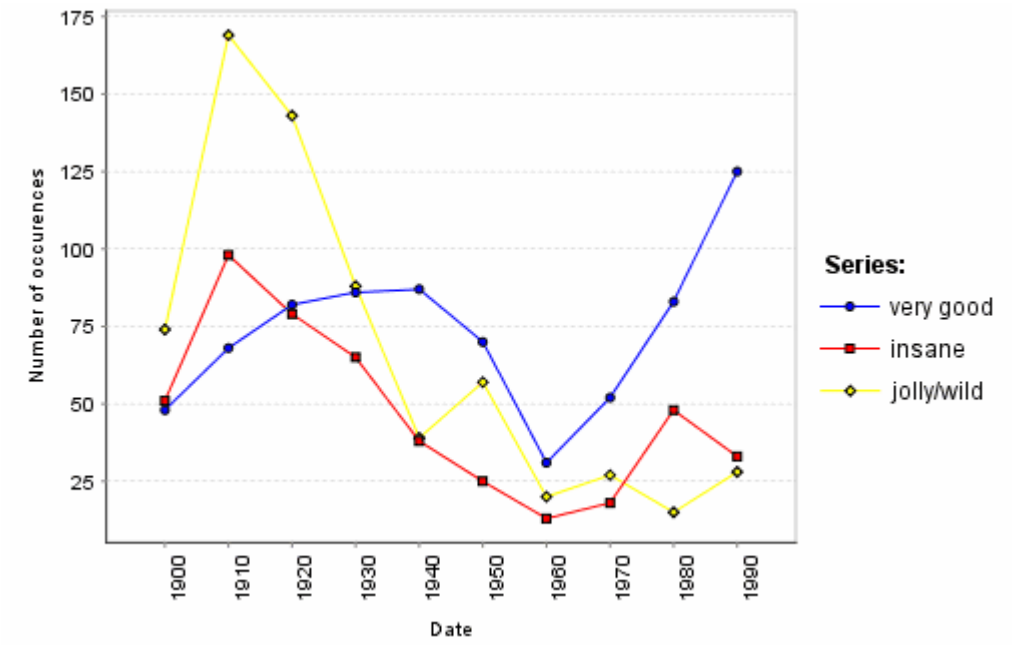

Abbildung 7: Vorkommen des Wortes „toll“ mit seinen unterschiedlichen Bedeutungen in den Dekaden des 20. Jh.

Abbildung 7 macht die Bedeutungsentwicklung des Wortes „toll“ im 20. Jahrhundert deutlich. In dem Maße, wie die älteren Bedeutungen *irre* und *ausgelassen/wild* in der Frequenz zurückgehen, wird die neuere Bedeutung *very good* mehr und mehr prominent.

„anrufen“

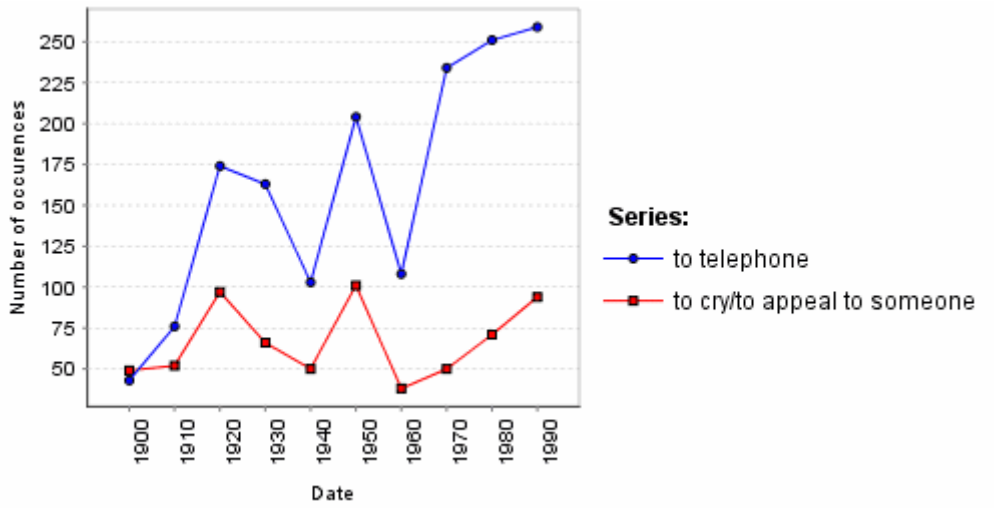

Abbildung 8: Vorkommen des Wortes „anrufen“ mit seinen unterschiedlichen Bedeutungen in den Dekaden des 20. Jh.

Abbildung 8 zeigt den starken Anstieg der Verwendung des Wortes „anrufen” in der Bedeutung *telefonieren* parallel zur kommerziellen Verbreitung des Telefons. Der in beiden Bedeutungen auftretende sägezahnartige Frequenzverlauf zwischen 1930 und 1970 könnte auf Unregelmäßigkeiten in der Ausgewogenheit der Korpusbasis hinweisen.

„Heuschrecke“

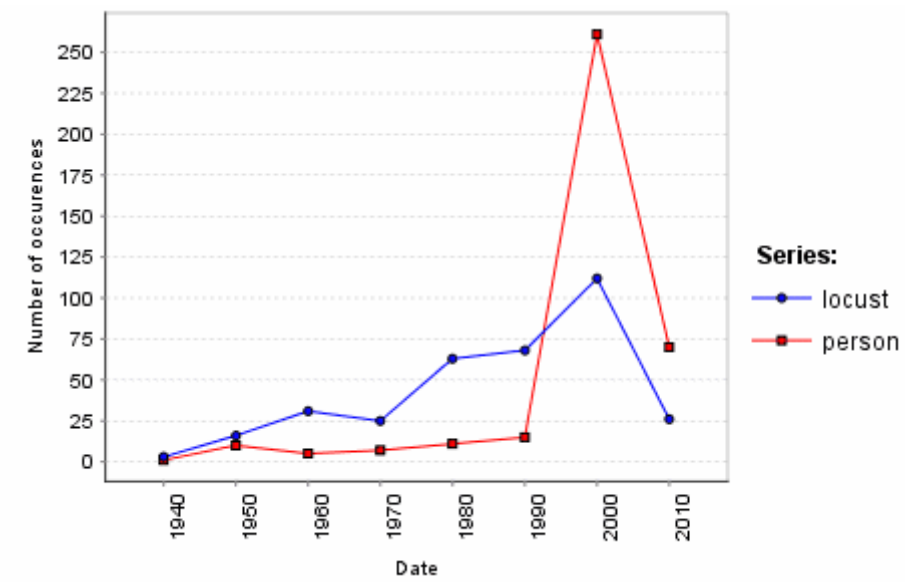

Abbildung 9: Vorkommen des Wortes „Heuschrecke“ mit seinen unterschiedlichen Bedeutungen im Zeitraum 1940-2010

Abbildung 9 verdeutlicht den sprunghaften Anstieg der Verwendung von „Heuschrecke“ in der Bedeutung *Person* in den 2000er Jahren, der Dekade, an deren Ende die internationale Finanz- und Bankenkriese steht. Auffällig ist auch der schnelle Rückgang der Frequenz zur 2010er-Dekade hin. Dabei ist jedoch zu berücksichtigen, dass zu dieser Dekade bislang noch deutlich weniger Dokumente vorliegen als zu den übrigen Dekaden.

„zeitnah“

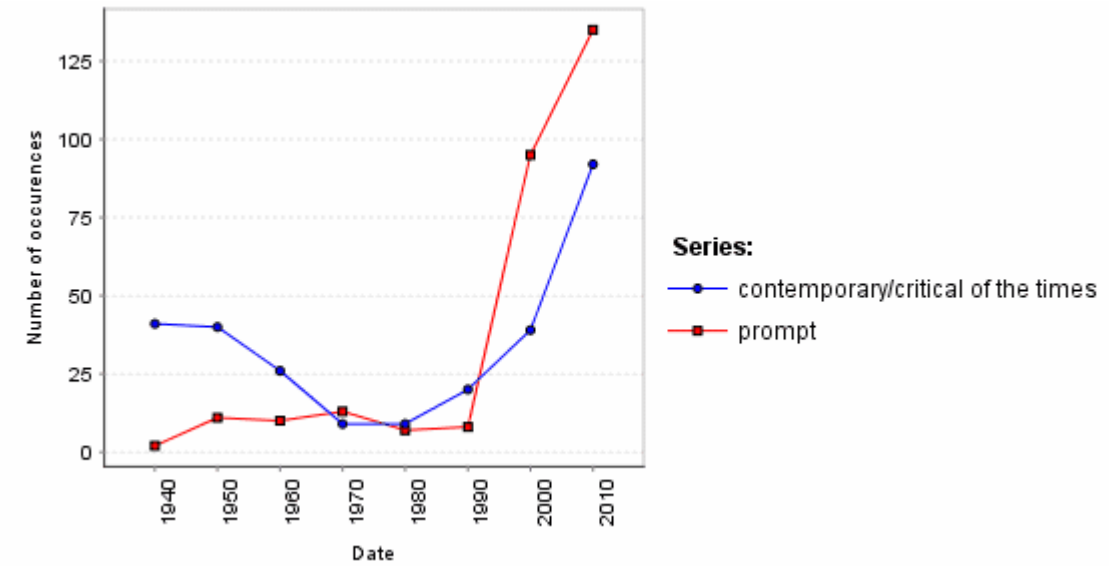

Abbildung 10: Vorkommen des Wortes „zeitnah“ mit seinen unterschiedlichen Bedeutungen im Zeitraum 1940-2010

Abbildung 10 zeigt schließlich die sprungartige Entwicklung der Bedeutung *unverzüglich* zu einer neuen, als prototypisch zu betrachtenden Bedeutung von „zeitnah“ ab den 2000er Jahren. Interessant ist allerdings auch der gleichzeitige Anstieg der Verwendung des Wortes in seiner älteren Bedeutung *zeitgenössisch/zeitkritisch*. Ob dieser tatsächlich existent ist, oder ob es sich dabei um eine Kumulierung von falsch positiven Bedeutungszuordnungen handelt, wäre auf Basis der KwiC-Snippets noch zu prüfen.

4. Fazit und Ausblick

In diesem Bericht wurde der Kernbestand der im KobRA-Projekt bis Projektmonat 24 (08/2014) entwickelten, angepassten und evaluierten Data-Mining-Verfahren dokumentiert. Dabei handelt es sich um überwachte und unüberwachte maschinelle Lernverfahren, die geeignet sind, typische, zeitaufwändige Routinearbeiten bei der Korpus-basierten Sprachanalyse teilweise oder vollständig zu automatisieren, sodass künftig größere Datenmengen in kürzerer Zeit mit präziserem Ergebnis analysiert werden können. Im Einzelnen werden Verfahren angeboten zum Filtern falsch positiver Suchtreffer, zum Klassifizieren und Annotieren von Suchtreffern sowie zur (semantischen) Disambiguierung der Suchtreffermenge, sodass eine Visualisierung von Häufigkeitsverteilungen disambiguierter lexikalischer Einheiten möglich wird. Die Interoperabilität mit Datenin- und -exportwerkzeugen sowie weiteren Lernverfahren ist durch die Integration der evaluierten Verfahren in die Data-Mining-Umgebung RapidMiner sichergestellt. Nutzer profitieren zudem von einem integrierten Annotationseditor für das Aktive Lernen und Schnittstellen zur CLARIN-D-Annotationsumgebung WebLicht sowie zu dem innovativen Visualisierungswerkzeug dfr-browser.

Die dokumentierten Verfahren wurden an konkreten linguistischen Fallstudien der Projektbeteiligten mit überwiegend zufriedenstellenden Ergebnissen evaluiert. Mithilfe eines überwachten Verfahrens lassen sich beispielsweise Aktionswörter wie „lach“, „freu“, „grins“ etc. auf Diskussionsseiten der deutschsprachigen Wikipedia mit einer Zuverlässigkeit von F_1=89,493 erkennen. Die automatische Disambiguierung von Wörtern mit mehreren Bedeutungen lässt sich bei Inhaltswörtern wie Substantiven, Verben oder Adjektiven mit ähnlicher Aussicht auf

Erfolg durchführen. Bei der Evaluation des Disambiguierungsverfahrens hat sich gezeigt, dass die Güte der Ergebnisse vor allem von der Anzahl der Bedeutungen des zu untersuchenden Wortes (je weniger desto besser) abhängt. Außerdem scheint in den meisten Fällen ein mittelgroßer Wortkontext zu den besten Ergebnissen zu führen. Die Anzahl der berücksichtigten KwiC-Snippets hatte in einem Bereich zwischen 500-5000 keine erkennbare Auswirkung auf das Ergebnis der automatischen Disambiguierung, ebensowenig das verwendete (orthographisch normalisierte) Korpus. Aus linguistischer Sicht überraschend ist die Erkenntnis, dass bislang alle evaluierten Lernverfahren mit einer Bags-of-Words-Repräsentation der Daten bessere Ergebnisse erzielt haben als mit einer Repräsentation, die weitere sprachliche Merkmale wie Wortartenzuordnungen oder Syntax berücksichtigt.

Im Anschluss an diesen Report kann die Integration der Verfahren in die Infrastrukturen der Sprachressourcenpartner im Projekt beginnen. Die Verfahren werden begleitend noch an weiteren Fallstudien und in verschiedenen Anwendungsszenarien erprobt. Es ist zudem geplant, für wiederkehrende Aufgabenstellungen vorgefertigte Prozesskonfigurationen zur Verfügung zu stellen und zu dokumentieren, um die Anwendung der Verfahren für die Nutzer in Forschung und Lehre künftig noch einfacher zu machen und die notwendige Einarbeitungszeit zu reduzieren.

5. Zitierte Literatur

Bartz, Thomas/Pölitz, Christian/Radtke, Nadja (2013a): Automatische Klassifikation von Stützverbgefügen mithilfe von Data-Mining. Technischer Bericht 2013/1, Technische Universität Dortmund.

Bartz, Thomas/Pölitz, Christian (2013b): Disambiguierung in Suchtrefferlisten aus großen Textkorpora. Technischer Bericht 2013/2, Technische Universität Dortmund.

Bartz, Thomas/Beißwenger, Michael/Storrer, Angelika (2014): Optimierung des Stuttgart-Tübingen-Tagset für die linguistische Annotation von Korpora zur internetbasierten Kommunikation: Phänomene, Herausforderungen, Erweiterungsvorschläge. In: Journal for Language Technology and Computational Linguistics 28 (1), 157-198.

Bartz, Thomas/Pölitz, Christian/Morik, Katharina/Storrer, Angelika (im Erscheinen): Using Data Mining and the CLARIN Infrastructure to Extend Cor□pus-based Linguistic Research. Proceedings of the Clarin Annual Conference, Soesterberg, NL.

Beißwenger, Michael/Ermakova, Maria/Geyken, Alexander/Lemnitzer, Lothar/Storrer, Angelika (2012): A TEI Schema for the Representation of Computer-mediated Communication. In: Journal of the Text Encoding Initiative (jTEI), 3.

Beißwenger, Michael/Lüngen, Harald/Margaretha, Eliza/Pölitz, Christian (2014): Mining corpora of computer-mediated communication: Analysis of linguistic features in Wikipedia talk pages using machine learning methods. In: Faaß, Gertrud/Ruppenhofer, Josef (Hg): Workshop Proceedings of the 12th Edition of the Konvens Conference. Hildesheim, Germany, October 8-10, 2014. Hildesheim: Universitätsverlag, 42-47.

Blei, David M./ Ng, Andrew Y./Jordan, Michael I. (2003): Latent dirichlet allocation. J. Mach. Learn. Res., 3:993–1022.

Brody, Samuel/Lapata, Mirella (2009): Bayesian word sense induction. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, EACL '09, pages 103–111, Stroudsburg, PA, USA.

Brown, Peter F./Della Pietra, Stephen A./Della Pietra, Vincent J. /Mercer, Robert L. (1991): Word-sense disambiguation using statistical methods. In Proceedings of the 29th Annual Meeting on Association for Computational Linguistics, ACL '91, pages 264–270, Stroudsburg, PA, USA.

Cohen, Jacob (1960): A coefficient of agreement for nominal scales. In Educational and Psychological Measurement 20, 37–46.

Collins, Michael/Duffy, Nigel (2001): Convolution kernels for natural language. In: Advances in neural information processing systems.

Engelberg, Stefan/Lemnitzer, Lothar (2009): Lexikographie und Wörterbuchbenutzung. Tübingen: Stauffenburg.

Fritz, Gerd (2012): Theories of meaning change – an overview. In: Maienborn, Claudia et al. (Hg): Semantics. An International Handbook of Natural Language Meaning. Volume 3. Berlin: de Gruyter, 2625-2651.

Fritz, Gerd (2005): Einführung in die historische Semantik. Tübingen: Niemeyer.

Geyken, Alexander (2007): The DWDS corpus: A reference corpus for the German language of the 20th century. In: Fellbaum, Christiane (Hg.): Idioms and Collocations. Corpus-based Linguistic and Lexicographic Studies. London u.a.: Continuum, 23–41.

Goldstone, Andrew (o.J.): dfr-browser. Take a MALLET to disciplinary history. Online: http://agoldst.github.io/dfr-browser/, abgerufen am 29.4.2015.

Gohl, Christine/Günthner, Susanne (1999): Grammatikalisierung von weil als Diskursmarker in der gesprochenen Sprache. Zeitschrift für Sprachwissenschaft, 18(12):39–75.

Griffiths, Thomas L./Steyvers, Mark (2004): Finding scientific topics. In Proceedings of the National Academy of Sciences, 101 (Suppl. 1), 5228–5235.

Günthner, Susanne/Auer, Peter (2005): Die Entstehung von Diskursmarkern im Deutschen – ein Fall von Grammatikalisierung? In: Leuschner, Torsten et al. (Hg): Grammatikalisierung im Deutschen. Berlin: de Gruyter, 335–362.

Günthner, Susanne (2008): Geht die Nebensatzstellung im Deutschen verloren? In: Denkler, Markus et al. (Hg): Frischwärts und unkaputtbar. Sprachverfall oder Sprachwandel im Deutschen, Münster: Aschendorff, 103–128.

Hinrichs, Erhard/Hinrichs, Marie/Zastrow, Thomas (2010): WebLicht: Web-based LRT services for German. In Proceedings of the ACL 2010 System Demonstrations, 25–29.

Hinrichs, Erhard/Zastrow, Thomas. (2012): Automatic Annotation and Manual Evaluation of the Diachronic German Corpus TüBa-D/DC. In Proceedings of the 8th International Conference on Language Resources and Evaluation, 1622–1627.

Imo, Wolfgang (2012): Wortart Diskursmarker? In: Björn Rothstein (Hg): Nicht-flektierende Wortarten. Berlin: de Gruyter, 48–88.

Joachims, Thorsten (1998): Text Categorization with Support Vector Machines: Learning with Many Relevant Features. Proceedings of the European Conference on Machine Learning, Berlin, Heidelberg: Springer.

Keller, Rudi/Kirschbaum, Ilja (2003): Bedeutungswandel. Eine Einführung. Berlin: de Gruyter.

Klein, Wolfgang/Geyken, Alexander (2010): Das Digitale Wörterbuch der Deutschen Sprache (DWDS). In: Heid, Ulrich et al. (Hg.): Lexikographica. Berlin u.a.: de Gruyter, 79–93.

Kupietz, Marc/ Lüngen, Harald (2014): Recent developments in DEREKO. In: Calzolari, Nicoletta et al. (Hg): Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, may.

Lodhi, Huma/Saunders, Craig/ Shawe-Taylor, John/Cristianini, Nello/ Watkins, Chris (2002): Text classification using string kernels. Journal of Machine Learning Research, 2:419–444.

Manning, Christopher D./Raghavan, Prabhakar/Schütze, Hinrich (2008): Introduction to Information Retrieval. Cambridge: Cambridge University Press.

Mierswa, Ingo/Wurst, Michael/Klinkenberg, Ralf/Scholz, Martin/Euler, Timm (2006): YALE: Rapid Prototyping for Complex Data Mining Tasks. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006), Philadelphia, USA.

Moschitti, Alessandro (2006): Making tree kernels practical for natural language learning. In: McCarthy, Diana/Wintner, Shuly (Hg): Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006), April 3-7, 2006, Trento, Italy, 113–120.

Navigli, Roberto (2009): Word sense disambiguation: A survey. ACM Comput. Surv., 41(2):10:1–10:69.

Navigli, Roberto/Crisafulli, Giuseppe (2010): Inducing word senses to improve web search result clustering. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP '10, Stroudsburg, PA, USA, 116–126.

Navigli, Roberto/Vannella, Daniele (2013): Semeval-2013 task 11: Word sense induction and disambiguation within an end-user application. In: Second Joint Conference on Lexical and Computa-

tional Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), Atlanta, Georgia, USA, 193–201.

Nello Cristianini & John Shawe-Taylor (2004): Kernel Methods for Pattern Analysis. Cambridge: Cambridge University Press.

Platt, John (1999): Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Smola, Alexander et al. (Hg.): Advances in Large Margin Classifiers. Cambridge: MIT Press.

Quasthoff, Uwe/Richter, Matthias/Biemann, Chris (2006): Corpus Portal for Search in Monolingual Corpora. In Proceedings of the fifth international conference on Language Resources and Evaluation, 1799-1802.

Rayson, Paul/Stevenson, Mark (2008): Sense and semantic tagging. In: Lüdeling, Anke/ Kytö, Merja (Hg): Corpus Linguistics. Volume 1. Berlin: de Gruyter, 564–578.

Rüping, Stefan (2006): Robust Probabilistic Calibration. In: Proceedings of the European Conference on Machine Learning (ECML), Berlin, Heidelberg: Springer, 743–750.

Rüping, Stefan (2004): A Simple Method for Estimating Conditional Probabilities in SVMs. In: Abecker, Andreas et al. (Hg): Lernen - Wissensentdeckung - Adaptivität, Berlin: Humboldt-Universität.

Shawe-Taylor, John/Cristianini, Neollo (2004): Kernel Methods for Pattern Analysis. New York: Cambridge University Press.

Steyvers, Mark/Smyth, Padhraic/Rosen-Zvi, Michal/ Griffiths, Thomas (2004): Probabilistic author-topic models for information discovery. In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’04, New York, NY, USA. ACM, 306–315.

Storrer, Angelika (2011): Korpusgestützte Sprachanalyse in Lexikographie und Phraseologie. In: Knapp, Karlfried et al. (Hg): Angewandte Linguistik. Ein Lehrbuch. 3. vollst. überarb. und erw. Aufl. Tübingen: Francke, 216–239.

Wang, Xuerui/McCallum, Andrew (2006): Topics over time: a non-Markov continuous-time model of topical trends. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. 424–433.