GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

Technische Universität Dortmund

Fakultät Kulturwissenschaften
Institut für deutsche Sprache und Literatur
Lehrstuhl für Linguistik der deutschen
Sprache und Sprachdidaktik

Fakultät Informatik
Lehrstuhl für Künstliche Intelligenz

Technischer Bericht

Nr. 2013/2 (Meilenstein 1)

Disambiguierung in Suchtrefferlisten aus großen Textkorpora

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
Bearbeiter/innen: Thomas Bartz, Christian Pölitz

Dortmund, den 31.8.2013

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.

Disambiguierung in Suchtrefferlisten aus großen Textkorpora

1. Problemstellung und Projektkontext

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. Die in diesem Report dokumentierte Fallstudie bearbeitet einen lange bekannten, aber bislang nicht zufriedenstellend gelösten Problemtyp bei der Arbeit mit großen digitalen Textkorpora: Ein Wort, dessen Gebrauch empirisch-quantitativ 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 (z.B. weil nur bestimmte Bedeutungen relevant sind, oder die Vorkommen verschiedener Bedeutungen eines Wortes verglichen werden sollen), was aber mit der vorhandenen Korpustechnologie nicht automatisch möglich ist. Gesucht werden deshalb Data-Mining-Verfahren, die den Linguisten dabei unterstützen, Trefferlisten zu einem Wort nach verschiedenen Bedeutungen dieses Wortes zu partitionieren.

In einer ersten Fallstudie zeigen wir im Folgenden exemplarisch an zwei ausgewählten Wörtern Herausforderungen und Nutzen möglicher Data-Mining-Verfahren für diese Disambiguierungsaufgabe aus der Perspektive der korpusbasierten lexikographischen Sprachbeschreibung. Für die lexikographische Beschreibung von Stichwörtern in Wörterbüchern, anderen (digitalen) lexikalischen Ressourcen oder sprachwissenschaftlichen Studien zur Wortschatz- und Bedeutungsentwicklung werden schon seit langem Korpora genutzt (vgl. Engelberg & Lemnitzer 2009). In einem digitalen Referenzkorpus wie dem Kernkorpus des Projekts ‚Digitales Wörterbuch der deutschen Sprache (DWDS)‘ (vgl. Geyken 2007), das im Hinblick auf die Verteilung der enthaltenen Textbestände auf die Textsortenbereiche Belletristik, Gebrauchsliteratur, Wissenschaft und journalistische Prosa sowie auf die Dekaden des 20. Jahrhunderts ausgewogen ist, können Lexikographen zu einem Suchwort automatisch Daten zur Frequenzentwicklung über das 20. Jahrhundert hinweg gewinnen und die Gebräuchlichkeit des Wortes in verschiedenen Textsortenbereichen vergleichen. Wenn man allerdings Aussagen zur Textsortenspezifik und zur Bedeutungsentwicklung verschiedener oder einer speziellen Wortbedeutung treffen möchte, müssen die vom System ausgegebenen Trefferlisten bei Wörtern mit mehreren Bedeutungen (Polyseme oder Homonyme) bislang manuell disambiguiert werden.

Für diese Fallstudie haben wir Trefferlisten zu den Wörtern *Leiter* und *zeitnah* erhoben. Beide Wörter versprechen interessante Einblicke in Prozesse des Sprachwandels und der Bedeutungsentwicklung:

- *Der Leiter* und *die Leiter* sind Homonyme mit den möglichen weiteren Bedeutungen *Energieleiter* und *Tonleiter*, wobei *der Leiter* mit Lehnwörtern wie *Boss* oder *Chef* konkurriert. Aus linguistischer Sicht wäre beispielsweise eine Untersuchung zur Frage, ob *der Leiter* in der Bedeutung *Boss*/*Chef* im aktuellen Sprachgebrauch durch die genannten Lehnwörter verdrängt wird, sehr reizvoll.
- *Zeitnah*, ein Polysem, das bis ins 20. Jahrhundert hinein v.a. in der Bedeutung *zeitgenössisch*/*zeitkritisch* gebräuchlich war, scheint in der zweiten Hälfte des 20. Jahrhunderts eine bemerkenswerte Bedeutungsentwicklung durchlaufen zu haben und findet sich heute am häufigsten in der Bedeutung *unverzüglich*/*prompt*. Zu untersuchen, wann diese Entwicklung genau eingesetzt hat, welche Verwendungsdomänen sie zuerst bzw. überhaupt betrifft und inwiefern die erste Bedeutung heute noch gebräuchlich ist, stellt ebenfalls ein reizvolles Desiderat für die linguistische lexikographische Sprachbeschreibung dar.

Zudem ist insbesondere *Leiter* stark frequent. Im DWDS-Kernkorpus des 20. Jahrhunderts finden sich beispielsweise insgesamt 7.032 Treffer (Stand: 6.6.2013). Für diese Trefferzahl ist eine manuelle Disambiguierung kaum mit vertretbarem zeitlichem Aufwand möglich.

Der Report ist folgendermaßen aufgebaut: Im folgenden Abschnitt 2 beschreiben wir zunächst die verwendete Datengrundlage und die linguistischen Vorarbeiten, die in die Data-Mining-Experimente eingeflossen sind. Abschnitt 3 erläutert das Vorgehen bei den Experimenten und die eingesetzten Data-Mining-Methoden; in Abschnitt 4 werden die Ergebnisse der Evaluation dargestellt. Erste Verfahren wurden bereits in einem Masterarbeitsprojekt für den Vergleich der Verwendung von Anglizismen und möglichen indigenen Entsprechungen genutzt (Maria Ermakova, Berlin).

2. Datengrundlage und linguistische Vorarbeiten

2.1 Datenerhebung

Die in den Experimenten genutzten Daten stammen aus dem Kernkorpus des 20. Jahrhunderts des Projekts ‚Digitales Wörterbuch der deutschen Sprache (DWDS)' (s. 1). Für das in diesem Korpus weniger frequente *zeitnah* wurden zusätzlich die Vorkommen im ZEIT-Korpus des DWDS (Details s. Klein/Geyken 2010) erhoben. Das ZEIT-Korpus besteht aus den Ausgaben der Wochenzeitung *die Zeit* von 1946 bis 2009. Es handelt sich dabei also um ein reines Zeitungskorpus.

Die Datenerhebung fand am 6.6.2013 statt. Die Daten sind teilweise urheberrechtlich geschützt, standen aber für die Auswertungen im Projekt vollständig zur Verfügung. Tabelle 1 vermittelt einen Überblick über die für die Experimente verwendeten Datenbestände.

Wort	DWDS		
	Kernkorpus des 20. Jh.	ZEIT-Korpus	Gesamt
Leiter	7.032	0	7.032
zeitnah	37	251	288

Tabelle 1: Berücksichtigte Treffer der untersuchten Wörter *Leiter* und *zeitnah* im DWDS-Kernkorpus des 20. Jh. und im ZEIT-Korpus des DWDS

2.2 Datenaufbereitung

Die vom Korpusrecherchesystem ausgegebenen Textsegmente, die wir im Folgenden als ‚Treffer-Snippets' bezeichnen, wurden in Excel-Dateien bearbeitet. Wie der Ausschnitt in Abbildung 1 zeigt, belegt jedes Treffer-Snippet eine Tabellenzeile. Im Snippet ist das gesuch-

te Wort durch festgelegte Sonderzeichen hervorgehoben (z.B.: „eine zerbrochene &&*Leiter*&&“). Die Snippets umfassen jeweils drei Sätze. Die Metadaten zu den Snippets (Erscheinungsdatum, Textsorte etc.) sind in jeweils separaten Spalten vermerkt.

Mit Blick auf die geplante Evaluation der für die Disambiguierung entwickelten bzw. angepassten Data-Mining-Verfahren (s. 4) wurde für *Leiter* eine Zufallsstichprobe von 2.000 Treffer-Snippets, für *zeitnah* der gesamte Datenbestand (288 Treffer) manuell disambiguiert. Dazu erhielten zwei Hilfskräfte unabhängig von einander jeweils die Aufgabe, für die vorgelegten Vorkommen von *Leiter* und *zeitnah* jeweils die Bedeutung zu bestimmen, in der diese Wörter verwendet werden. Auf Basis der Bedeutungsdifferenzierung im Duden-Universalwörterbuch (Duden 2011) wurden die zu bestimmenden Bedeutungen beschrieben, durch jeweils einen passenden Korpusbeleg illustriert und den Annotatoren schließlich an die Hand gegeben. Durch dieses Vorgehen wurde ein sehr hohes Inter-Annotator-Agreement zwischen den beteiligten Hilfskräften erreicht (s. Tabellen 2 und 3).

	A	B	C	D	E	F	G
1	ID	Random	Treffer	Datum	Lesart	Bemerkungen	Texttyp
2	1	0.712261401	Die St. von Blasinstrumenten ist durch ihre Konstruktion (Grifflöcher, Ventile 2) relativ festgelegt. Von der leichten Umstimmbarkeit bei Saiteninstrumenten wird in der Scordatura Gebrauch gemacht; die normalen St. sind die in Quinten (Violinfamilie), Quarten mit Terz (Gambenfamilie, Lauten, Gitarren) sowie die seltenen in diatonischen und chromatischen &&Leitern&& (Erzlauten, Angelica, chromatische Harfe). Die in neuerer Zeit geforderte absolute Tonhöhe wird mit Stimmgabel oder elektrischem Generator nach dem geltenden Stimmton (Kammerton) eingestimmt. Im Orchester wird nach der Oboe oder Klarinette gestimmt, wenn nicht ein Tasteninstrument in temperierter St. mitwirkt.	1989-12-31	Tonleiter		Wissenschaft:Musikwissenschaft
3	2	0.929629808	Die Hauptsache ist, daß es sich nicht um einen zeitweiligen, sondern um einen dauernden Anstaltsaufenthalt handelt. Nach Schmittmann, dem &&Leiter&& der rheinischen Versicherungsanstalt, der warm für die Asylisierung der infektiösen Invaliden eintritt, wurden von seiner Verwaltung aus im Jahre 1910 bereits 400 tuberkulöse Invalide gegen Abtretung ihrer Rente in dauernde Anstaltspflege übernommen. Auch von privaten und unter geistlicher Leitung stehenden Wohlfahrtsvereinen und neuerdings auch von Kommunalverwaltungen sind recht befriedigende Erfahrungen mit Pflegeheimen für Tuberkulöse gemacht worden.	1912-12-31	Führungsperson		Gebrauchsliteratur:Wissenschaft
4	3	0.19666777	Albrecht Schaeffer hat eine durch die Jahre seines Schaffens hindurch sich immer gleichbleibend glückliche Feder, wenn er zwei Menschen eins werden läßt. Als ich zum ersten Male den> Helianth <las, blieb mir der Atem aus im Kapitel » Rausch «: Georg steigt auf die &&Leiter&& und klettert zur Anna ins Zimmer, und dann verlieren weder der Dichter noch die Liebenden viel Zeit. Es geht aufs Ganze - » er hörte sie leise aufstöhnen, fühlte selber Schmerz, war ratlos, aber da kam die anschluchzende Sekunde, und jählings fühlte er sich von der unsichtbaren Riesenfaust zu rasenden Zuckungen der Lust… « und so weiter.	1953-12-31	Stiege		Belletristik

Abbildung 1: Excel-Tabelle mit importierten Treffer-Snippets aus dem DWDS-Kernkorpus des 20. Jh. für *Leiter*, Hervorhebung durch festgelegte Sonderzeichen („&&“); Metadaten und manuelle Disambiguierung in separaten Spalten.

Leiter: **Belege (übereinstimmend)**	
Führungsperson	1.653
Sprossenstiege	294
Energieleiter	29
Tonleiter	10
Gesamt	1.986

zeitnah: **Belege (übereinstimmend)**	
gegenwartsnah	145
unverzüglich	130
Gesamt	275

Inter-Annotator-Agreement	
prozentual	0.993
Kappa (Cohen 1960)	0.972

Inter-Annotator-Agreement	
prozentual	0.955
Kappa (Cohen 1960)	0.910

Tabellen 2 und 3: Anzahl der übereinstimmend bestimmten Bedeutungen zu Vorkommen von *Leiter* und *zeitnah* und Inter-Annotator-Agreement (prozentualer und konservativer Kappa-Wert, vgl. Cohen 1960)

3. Beschreibung der Data-Mining-Experimente

3.1 Vorüberlegungen und Aufbau der Experimente

Wie unter 1. bereits erläutert, ist es bislang mithilfe der einschlägigen großen Korpora und ihrer Abfragesysteme nicht möglich, separate Trefferlisten für verschiedene Bedeutungen eines gesuchten Wortes zu erzeugen bzw. gezielt nach bestimmten Bedeutungen eines Wortes zu suchen. Beim manuellen Sichten der Suchtreffer lassen sich verschiedene Bedeutungen

eines gesuchten Wortes aber meist leicht an den Kontexten erkennen, in denen sie verwendet werden. Verwendungen eines Wortes in einer bestimmten Bedeutung korrespondieren offenbar mit überzufällig häufigen Vorkommen bestimmter anderer Wörter bzw. sprachlicher Strukturen im Umfeld dieser Wörter. Data-Mining-Verfahren können diese im sprachlichen Kontext eines Suchtreffers gegebenen latenten Informationen für die automatische Disambiguierung nutzbar machen. Dazu werden um alle Vorkommen eines betreffenden Wortes Kontextfenster in einer bestimmten Größe gelegt und mithilfe von Wort- und Kookkurrenzstatistiken Verteilungen von Kontextwörtern ermittelt, die als Repräsentationen von Bedeutungen aufgefasst werden können. Für jedes einzelne Kontextfenster lässt sich daraufhin die Wahrscheinlichkeit berechnen, mit der ein Vorkommen des betreffenden Wortes einer bestimmten Bedeutung zugeordnet werden kann. Ein großer Vorteil solcher induktiv von den Kontexten betreffender Wörter ausgehender Verfahren ist die Tatsache, dass sich auf diese Weise auch unerwartete oder bislang lexikographisch nicht erfasste Bedeutungen identifizieren lassen.

Die Induktion von Wortbedeutungen ist in der Forschung zu Data-Mining-Verfahren bereits gut erforscht. 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 Latent-Dirichlet-Allocation (LDA, vgl. Blei et al. 2003) tendenziell die besten Ergebnisse erzielen lassen. Sie erweiterten zudem das Verfahren um die Möglichkeit, neben den reinen Wortvorkommen verschiedene weitere Kontextmerkmale zu berücksichtigen (z.B. Part-of-Speech-Tags, Syntax, etc.). LDA wurde ursprünglich zum thematischen Clustern 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 Clustering der Trefferlisten von Web-Suchmaschinen.

Der in diesem Report vorgestellte Ansatz unterscheidet sich von diesen Vorarbeiten v.a. dadurch, dass LDA auf Trefferlisten aus Korpussuchen angewendet wird. Während sich die Ergebnisse einer Abfrage in einer Web-Suchmaschine meist auf (Web-)Texte beziehen, die mit dem Suchwort in einem engen thematischen Zusammenhang stehen, ermitteln Korpusabfragesysteme Vorkommen des gesuchten Wortes im ganzen Korpus, unabhängig von der thematischen Relevanz der Fundstellen. Dadurch erscheinen die gesuchten Wörter öfter in weniger typischen, semantisch tendenziell weniger eindeutigen Kontexten. Im Textsortenbereich Belletristik und in Zeitungstexten finden sich nicht selten metaphorische Verwendungen. Möglichkeiten und Grenzen der Anwendung von Clusteringverfahren wie LDA zur automatischen Disambiguierung von Suchtreffern aus Korpora sind noch kaum erforscht. Die Computerlinguistik stellt Expertise in Bezug auf die linguistische Aufbereitung der Korpusdaten durch Wortarten- und Syntaxannotationen bereit. Metadaten ermöglichen zudem die Zuordnung von Belegen zu Textsorten und Zeiträumen (z.B. im DWDS-Kernkorpus). Welche dieser Merkmale als sogenannte ‚Features' die Ergebnisse von Clusteringverfahren verbessern und wie Treffer-Snippets und Merkmale idealerweise für die Verfahren zu repräsentieren sind, sind interessante und größtenteils noch offene Fragen.

Die folgenden Abschnitte erläutern die Experimente, die zur Lösung der in Abschnitt 1 dargestellten Problemstellung durchgeführt wurden. Ein LDA-Clusteringverfahren wurde in fünf unterschiedlichen Treatments auf die in Abschnitt 2 dargestellten ungesichteten Daten angewendet und anschließend anhand der manuell disambiguierten Daten evaluiert. Die Treatments unterscheiden sich hinsichtlich der Größe der berücksichtigten Kontextfenster und der Features, die für das Clustering genutzt wurden:

1. Bags-of-Words-Ansatz mit einem Kontext von jeweils 10 Wörtern vor und nach dem betreffenden Wort: **w10**,

2. Bags-of-Words-Ansatz mit einem Kontext von jeweils 40 Wörtern vor und nach dem betreffenden Wort: **w40**,
3. Bags-of-Words-Ansatz mit einem Kontext von insgesamt 80 Wörtern vor und nach dem betreffenden Wort: **w80**,
4. Bags-of-Words-Ansatz unter Berücksichtigung der kompletten Treffer-Snippets (drei Sätze, das betreffende Wort im zweiten Satz): **all**,
5. Bags-of-Words-Ansatz, bei dem nur diejenigen Kontextwörter berücksichtigt werden, die syntaktisch unmittelbar vom betreffenden Wort abhängig sind: **syntax**.

Als Maß für die Zuverlässigkeit der Verfahren dient das gewichtete harmonische Mittel aus Präzision (Precision) und Ausbeute (Recall), wobei Genauigkeit und Ausbeute gleich gewichtet werden. Der auf diese Weise ermittelte F_1-Wert stellt ein Standardmaß für die Beurteilung automatischer Disambiguierungsverfahren dar (vgl. Navigli & Vannella, 2013).

3.2 Technische Beschreibung der Experimente

3.2.1 Vorverarbeitung

Die Treffer-Snippets liegen als Sequenzen von Wörtern vor, die zunächst vorverarbeitet werden müssen, um als Eingabe für das Clusteringverfahren dienen zu können. Wir repräsentieren die Snippets als Bags-of-Words, wobei jedes Treffer-Snippet als großer Vektor mit Einträgen für jedes Wort der Gesamtmenge aller Wörter in einer Suchergebnisliste dargestellt wird (ein sogenannter ‚Wortvektor'). In einer Trefferliste mit *N* Wörtern ist der Vektor *N*-dimensional. Die Elemente der Wortvektoren können binär sein und das bloße Vorkommen eines Wortes in einem Treffer-Snippet oder Häufigkeiten des Wortes in einem Snippet und in allen Snippets der Trefferliste 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 Treffer-Snippets (hier wie ‚Texte' behandelt) als Wortvektoren abbildet. Dies geschieht formal so:

$\varphi(d) = (f(\mathrm{w}1,d), f(\mathrm{w}2,d), \ldots, f(\mathrm{w}N,d))$, wobei $f(\mathrm{w}i,d)$ die Anzahl oder (normalisierte) Frequenz von Wort *i* in Text *d* (für ‚document') angibt.

Weil untersucht werden soll, inwiefern Kontextinformationen von unterschiedlicher Größe und Zusammenstellung das Ergebnis des automatischen Clustering beeinflussen, werden verschiedene Bags-of-Words-Repräsentationen erprobt (s. 3.1), wobei einmal die Menge der berücksichtigten Wortvorkommen im Kontext und einmal ihre syntaktische Abhängigkeit vom zu disambiguierenden Wort entscheidend ist. Für die syntaktische Annotation der Treffer-Snippets wurde der Stanford-Konstituentenparser genutzt (Klein & Manning 2003).

3.2.2 Disambiguierung

Für die automatische Disambiguierung nutzen wir das Verfahren der Latent-Dirichlet-Allocation (LDA, s. 3.1), wie es von Blei et al. (2003) vorgestellt wurde. LDA schätzt die Wahrscheinlichkeitsverteilungen von Wörtern und Dokumenten (hier: Treffer-Snippets) über eine bestimmte Anzahl überzufällig häufig auftretender Kontextwörter, sogenannter ‚Topics', die als Repräsentationen für Bedeutungen 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 Wahrscheinlichkeit 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 Bedeutung des gesuchten Wortes.

4. Evaluation

4.1 Quantitative Evaluation

Zur Evaluation des in Abschnitt 3 beschriebenen Verfahrens werden die durch das automatische Verfahren gebildeten Cluster mit den von zwei Hilfskräften manuell übereinstimmend disambiguierten Datensätzen abgeglichen (s. 2). Überprüft wird jeweils Präzision und Ausbeute der automatischen Disambiguierung im Vergleich zu den manuell disambiguierten Daten. Als Gütekriterium für das Clusteringverfahren dient der F_1-Score, das gewichtete harmonische Mittel aus Präzision (Precision) und Ausbeute (Recall), wobei Genauigkeit und Ausbeute gleich gewichtet werden; formal: $F_1 = 2*(\text{Präzision}*\text{Ausbeute})/(\text{Präzision}+\text{Ausbeute})$. Die Tabellen 4 und 5 zeigen die für *Leiter* und *zeitnah* in den einzelnen Treatments (s. 3.1) ermittelten Werte:

Leiter

Kontext (Wörter)	w10	w40	w80	all	syntax
F_1	0.727	0.749	0.741	0.742	0.690

Tabelle 4: F_1-Scores für die Güte der automatischen Disambiguierung der Treffer mit *Leiter*

zeitnah

Kontext (Wörter)	w10	w40	w80	all	syntax
F_1	0.777	0.692	0.763	0.749	0.458

Tabelle 5: F_1-Scores für die Güte der automatischen Disambiguierung der Treffer mit *zeitnah*

Die Ergebnisse zeigen, dass die automatische Disambiguierung von Treffer-Snippets aus Korpusrecherchen bereits mit einfachen Bags-of-Words-Repräsentationen der Snippets mit einer Güte (F_1) von zwischen 70% und 78% möglich ist. Generell scheint die Berücksichtigung eines möglichst großen Kontextfensters (‚all') robust eine mittlere Güte zu erzielen, während optimale Kontextfenster wortspezifisch variieren. Überraschend ist die schlechte Güte des Verfahrens, bei dem die Auswahl des zu berücksichtigenden Wortkontexts auf unmittelbaren syntaktischen Abhängigkeiten beruht (‚syntax'). Offensichtlich bergen auch solche Wörter für die Disambiguierung essenzielle latente Informationen, die syntaktisch von dem zu disambiguierenden Wort nicht abhängen bzw. obligatorisch sind. Inwiefern eine parallele Repräsentation von Wortarten- oder syntaktischen Merkmalen eine Verbesserung der Güte bewirken kann, ist in weiteren Experimenten zu untersuchen.

4.2 Qualitative Evaluation

Die durch das automatische Verfahren ermittelten, ein Topic am wahrscheinlichsten induzierenden Wörter (s. 3.2) sind aufschlussreich für den Nutzwert der erzeugten Cluster für anknüpfende lexikographische Untersuchungen. Tabelle 6 zeigt die für *Leiter* ermittelten Topics und diese repräsentierende Kontextwörter:

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
DDR	politisch	Berlin	Musik	hinauf
Abteilung	Partei	Prof.	München	Mann
Regierung	Korps	Dr.	New_York	oben
Minister	Führer	Hochschule	Dirigent	gehen
ZK	Arbeit	Institut	Oper	Sprosse
SED	NSDAP	Lehrer	Komponist	Wand

Tabelle 6: Automatisch induzierte Topics und wahrscheinlichste Kontextwörter (Auszug aus Top 50)

Es zeigt sich, dass die ermittelten Bedeutungen nicht den Bedeutungsbestimmungen entsprechen, die sich z.B. in gängigen Wörterbüchern oder anderen lexikalischen Ressourcen finden. Insbesondere für die frequenteste Bedeutung *Führungsperson*/*Boss* (vgl. Tabelle 2) wurde eine sehr feinkörnige Disambiguierung vorgenommen; die Kontextwörter weisen auf Belege für *Leiter* i.S.v. *politischer Leiter, DDR/Drittes Reich* (Topic 1/2), *Leiter einer Bildungsinstitution* (Topic 3) und *musikalischer Leiter* (Topic 4) hin. Dies ist als Vorteil zu werten: Korpusnutzer können bei Bedarf eine sehr feinkörnige Bedeutungsunterscheidung erhalten. Eine grobkörnigere Disambiguierung lässt sich je nach Fragestellung leicht durch Zusammenführen einzelner Cluster erreichen.

5. Fazit und Anschlussarbeiten

Bereits die bislang getesteten Verfahren ermöglichen eine Disambiguierung der Gesamttrefferlisten beliebiger Wörter mit akzeptabler Genauigkeit bzw. zumindest mit bekannter ‚Grauzone'. Damit wird es künftig möglich sein, auch solche quantitative empirische Untersuchungen zu lexikographisch interessanten hochfrequenten Wörtern durchzuführen, die eine Disambiguierung homonymer oder polysemer Wortformen erfordern. Von den Verfahren können auch statistische Visualisierungs- und Analysewerkzeuge profitieren (z.B. ‚Wortverlauf' und ‚Wortprofil' des DWDS), die bislang noch überwiegend formbasiert arbeiten.

In Bezug auf die Frage der Anwendbarkeit von Data-Mining-Verfahren – genauer: Clusteringverfahren auf Basis der Latent-Dirichlet-Allocation – auf Treffer-Snippets aus Korpussuchen können die Experimente ersten Aufschluss darüber geben, mit welcher Repräsentation der Korpusdaten die beste Güte des evaluierten Verfahrens zu erreichen ist.

Aufbauend auf den in diesem Report dargestellten Erkenntnissen wird diese Fragestellung in weiteren Experimenten vertieft. In den weiterführenden Arbeiten soll insbesondere erprobt werden, durch welche weiteren Merkmale und ggf. Merkmalskombinationen (z.B. N-Gramme, vollständige oder teilweise syntaktische Annotation der Treffer, Berücksichtigung weiterer manuell annotierter Merkmale wie typische prädikative Nomina/Suffixe etc., Textsorten-Metadaten) die Verfahren in ihrer Güte noch verbessert werden können. Darüber hinaus sollen auch innovative Möglichkeiten der Visualisierung von Bedeutungsentwicklungen erprobt werden.

6. Zitierte Literatur

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

Samuel Brody & Mirella Lapata (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. Association for Computational Linguistics.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra & Robert L. Mercer (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. Association for Computational Linguistics.

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

Duden (2011): Deutsches Universalwörterbuch. 7, überarb. u. erw. Aufl., Dudenverlag, Berlin.

Stefan Engelberg & Lothar Lemnitzer (2009): Lexikographie und Wörterbuchbenutzung. Stauffenburg, Tuebingen.

Alexander Geyken (2007): The DWDS corpus. A reference corpus for the German language of the twentieth century. In Christiane Fellbaum, editor, Idioms and collocations. Corpus-based linguistic and lexicographic studies, pages 23–40. Continuum, London.

T. L. Griffiths & M. Steyvers (2004): Finding scientific topics. Proceedings of the National Academy of Sciences, 101(Suppl. 1):5228–5235, April.

Dan Klein & Christopher D. Manning (2003): Accurate unlexicalized parsing. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics – Volume 1, ACL '03, pages 423–430, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Christopher D. Manning, Prabhakar Raghavan & Hinrich Schütze (2008): Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA.

Roberto Navigli and Giuseppe Crisafulli (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, pages 116–126, Stroudsburg, PA, USA. Association for Computational Linguistics.

Roberto Navigli & Daniele Vannella (2013): Semeval-2013 task 11: Word sense induction and disambiguation within an end-user application. In Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), pages 193–201, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

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

Mark Steyvers, Padhraic Smyth, Michal Rosen-Zvi & Thomas Griffiths (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, pages 306–315, New York, NY, USA. ACM.