1 Das World Wide Web als computerlinguistische Ressource

Iryna Gurevych

1.1 Einleitung

Das *World Wide Web* (WWW) hat sich in den letzten Jahren einerseits zur wichtigsten Informations- und Kommunikationsstruktur und andererseits zur wichtigsten computerlinguistischen Ressource entwickelt. Durch Soziale Software ist für die Computerlinguistik ein Benutzer-definiertes semantisches Tagging-System von bisher nicht dagewesener Größe entstanden. Diese Entwicklung birgt das Potential, den Wissensakquisitionsproblemen in der Computerlinguistik Abhilfe zu schaffen. Zum einen handelt es sich beim Web um einen **enormen und reichen Datenbestand**. Zum anderen können aus diesem Datenbestand aufgabenspezifische Korpora für unterschiedliche Sprachen, Domänen, Textsorten, usw. gewonnen werden, um die Durchführung von computerlinguistischen Untersuchungen zu ermöglichen. Eine besondere Bedeutung spielen dabei die Benutzer-generierten Inhalte. Mit Benutzer-Tags ausgezeichnet, bilden sie die sogenannten **Folksonomien**. Sie beinhalten sehr wertvolle semantische Informationen, die mit computerlinguistischen Methoden weiter analysiert und erschlossen werden können. Das resultierende lexikalische und semantische Wissen sowie das Weltwissen kann in umgekehrter Richtung den computerlinguistischen Algorithmen zugeführt werden, um neue Anwendungen, z.B. im Bereich Textinformationsmanagement, zu ermöglichen. An dieser Stelle ist die *duale* Art des Verhältnisses zwischen dem WWW und der Computerlinguistik festzuhalten. Auf der einen Seite stellt das WWW **eine wichtige Ressource** für den Aufbau und die Verbesserungen von computerlinguistischen Systemen dar. Auf der anderen Seite bietet die Computerlinguistik dringend benötigte **Technologien**, um dem Benutzer den Umgang mit den Informationen im WWW zu ermöglichen, d.h. diese besser finden, filtern und auswerten zu können.

1.2 Web als Korpus und Webkorpora

Beim Web handelt es sich um ein **multilinguales Korpus**. Laut einer Untersuchung von Xu (2000) waren im Jahr 2000 71% aller Webseiten, die von der Suchmaschine Excite indiziert wurden, auf Englisch verfasst, gefolgt von Japanisch (6,8%), Deutsch (5,1%), Französisch (1,8%), Chinesisch (1,5%), Spanisch (1,1%), Italienisch (0,9%) und Schwedisch (0,7%). 2002 waren laut einer anderen Untersuchung von Ebbertz (2002) die Sprachen wie folgt verteilt: Englisch 56,4%, Deutsch 7,7%, Japanisch 4,9%, Spanisch 3,0%, Französisch 5,6%, Chinesisch 2,4% und Italienisch 2% .

Grundsätzlich können zwei *Vorgehensweisen* bei der Nutzung des WWW als computerlinguistisches Korpus unterschieden werden: (i) über Programmier-

schnittstellen gängiger Suchmaschinen kann auf statistische Informationen, beispielsweise die *Anzahl von Treffern*, zugegriffen werden, und (ii) die Programmierschnittstellen können eingesetzt werden, um *aufgabenspezifische Korpora* nach vorgegebenen Anforderungen zu erstellen. Neben vielen in der Einführung aufgezählten Vorteilen sind bei der erstgenannten Vorgehensweise eine Reihe von Herausforderungen zu berücksichtigen. Insbesondere sind experimentelle Ergebnisse auf der Grundlage des WWW **unzuverlässig** und oft **nicht reproduzierbar**. Sie können je nach Suchmaschine und je nach Ausführungszeit der Anfrage stark schwanken. Die genaue Zusammensetzung des Korpus, das von Suchmaschinen erfasst und indiziert wird, ist nicht bekannt. Das erschwert die Interpretation der Ergebnisse. Ebenso unbekannt ist, wie **vollständig** die Ergebnisse sind, da die Anzahl von korrekten Suchergebnissen unbekannt ist. Aus diesen Gründen kann es für bestimmte Einsatzszenarien sinnvoll sein, die zweite Vorgehensweise zu wählen. Diese hat zum einen den Vorteil, dass zuverlässige Statistiken auf konstanter Datengrundlage berechnet werden können, zum anderen hat der Korpusersteller zumindest eingeschränkt Kontrolle über die zugrundeliegenden Daten. Der Einsatz des Korpus kann dann je nach Domäne und Aufgabenstellung fokussierter erfolgen.

Bei der Nutzung des WWW als Korpus müssen einige Problemfelder addressiert werden: Zum einen werden erhebliche **Speicherplatz und Rechenkapazitäten** benötigt. Texte sind oft mit html-Code und anderen irrelevanten Inhalten wie sogenannten *boilerplates*, d.h. Navigationsmenüs, Werbung, usw. **vermischt**. Es ist wichtig, **Duplikate** in den Daten zu identifizieren und zu entfernen, um Bias zu vermeiden. Um ein monolinguales Korpus zusammenzustellen, müssen Webseiten in der **Zielsprache** zunächst automatisch identifiziert werden. Webseiten sind in der Regel nur **teilweise** mit Metainformationen versehen, und diese sind oft **uneinheitlich**. Auch **Autorenschaft** von Texten kann oft nicht hergestellt werden. Die Inhalte sind stark **diversifiziert** und von unterschiedlicher sprachlicher **Qualität**, die Daten müssen teilweise von **Spam** bereinigt werden. Je nach Art der Entstehung, z.B. Emails, Blog-Einträge oder Wikis, liegen verschiedene **Sprachregister** mit speziellen Eigenschaften vor. Es ist wünschenswert, die Webseiten mit Metadaten, beispielsweise in Bezug auf die **Genre** oder die **Themen**, automatisch zu annotieren. Die Nutzung von den gesammelten Daten sowie von den damit verknüpften Metainformationen und Medien wie Bildern für die computerlinguistische Forschung bedarf oft einer **rechtlichen Absicherung**, die in der Praxis problematisch ist. Nichtsdestoweniger sind die Webdaten im Hinblick auf ihre **Aktualität** sehr wertvoll und ermöglichen zum Beispiel die Erforschung von neuen Wörtern und sprachlichen Phänomenen. Die Weiterentwicklung von Methoden und Standards für die Erstellung Web-basierter Korpora ist insofern sehr berechtigt.

Ansätze für die Nutzung von WWW als Korpus Keller and Lapata (2003) beschreiben einen Ansatz, bei dem die Frequenzen für Bigramme auf Grundlage des WWW approximiert werden, die in einem herkömmlichen Korpus nicht vorkommen. Sie erhalten die Frequenzen für 'Adjektiv-Nomen', 'Nomen-

Nomen' und 'Verb-Objekt'-Bigramme aus dem Web via Anfragen an eine Suchmaschine. Die Evaluierung dieser Methode zeigt u. a., dass (i) die Web-basierten Frequenzen eine hohe Korrelation mit den Korpus-basierten Frequenzen aufweisen, (ii) die Web-basierten Frequenzen zuverlässig mit den menschlichen Bewertungen korrelieren, und (iii) die Web-basierten Frequenzen als gute Indikatoren für Disambiguierungsaufgaben dienen können.

Lapata and Keller (2005) beschreiben eine systematische Untersuchung der Nützlichkeit von Web-basierten Modellen für eine Reihe von computerlinguistischen Aufgaben, indem (i) Syntax und Semantik, (ii) Generierung und Analyse, und (iii) ein breites Spektrum an N-Grammen und Wortarten einbezogen sind. Für eine Mehrzahl der Aufgaben weisen einfache, unüberwachte Modelle der N-Gramme eine bessere Performanz auf, wenn sie auf den WWW-Daten und nicht auf einem Standard-Korpus berechnet werden. Eine weitere Verbesserung kann in einigen Fällen durch die Kombination von Web-basierten und Korpus-basierten Frequenzen mittels Back-off und Interpolierungstechniken erzielt werden.

Ein weiterer Ansatz für die Nutzung des WWW als Korpus ist die Sammlung von sogenannten *Text-Snippets*, also Textfragmenten. Die verfügbaren Programmierschnittstellen wie die Google API erlauben, den Kontext von Suchwörtern zu erhalten. So können Suchbegriff-zentrierte, aufgabenspezifische Korpora aufgebaut werden. Des Weiteren können die Webseiten komplett heruntergeladen werden. Baroni and Bernardini (2004) stellen ein System namens **BootCaT** vor, welches dazu dient, themenspezifische Web-Korpora zu erstellen. Der Benutzer legt im ersten Schritt die Suchbegriffe fest. Dann werden die Webseiten gesammelt, die die Kombination von diesen Suchbegriffen für eine gegebene Domäne enthalten. Anschließend werden Kollokationsstatistiken erstellt, um z.B. domänenspezifische Begriffe zu finden (dabei werden die gesammelten Webseiten mit einem allgemeinen Korpus verglichen).

Das **WaCky-System**[1] (Web as Corpus kool ynitiative) bietet verschiedene Werkzeuge und Programmierschnittstellen, die einem Nutzer ermöglichen, einen Teil des Webs zu crawlen, zu verarbeiten, zu indexieren und darauf zu suchen. Das mit Hilfe dieses Systems erstellte Korpus für Deutsch (deWaC) mit 1,5 Milliarden Token und Italienisch (itWaC) mit 2 Milliarden Token stehen mit annotierten Wortarten und Lemmata zur Verfügung. Das englische Korpus besteht aus über mehr als 2 Milliarden Token und gehört derzeit zu den größten frei verfügbaren linguistischen Ressourcen im Web (Ferraresi *et al.* (2008)).

1.3 Sozio-Semantisches Web

Die Entwicklungen im WWW-Bereich waren in den letzten Jahren durch die sogenannten Sozio-Semantischen Technologien gekennzeichnet (Gruber (2008)). **Sozio-Semantisches Web** bezeichnet demnach die Vereinigung von umfangreichen Wissensdatenbanken, die von der Internet-Gemeinschaft kollaborativ erstellt werden, mit der Ausdrucksmächtigkeit und den Inferenzmechanismen

[1] http://wacky.sslmit.unibo.it/

des Semantic Web. Diese vereinigte Vision soll zu neuartigen Webanwendungen führen, die die in den Webdaten implizit repräsentierten semantischen Relationen automatisch identifizieren und daraus ein Netzwerk mit strukturiertem Wissen erstellen.

In der Computerlinguistik wurden im Bereich der lexikalischen Semantik und der semantischen Erschließung von Inhalten wichtige Schritte in Richtung des Sozio-Semantischen Webs gemacht. Insbesondere verschiebte sich der Fokus von herkömmlichen manuell erstellten Ressourcen, z.B. Wortnetzen, zur automatischen Erschließung und Nutzung des Wissens in den sogenannten **kollaborativen Wissensdatenbanken**. Letztere entstehen als Folge freiwilliger Benutzerbeiträge im Sozialen Web, also *bottom-up*. Für den Einsatz als computerlinguistische Ressource müssen solche Wissensquellen speziell aufbereitet werden, da sie nicht zu diesem Zweck geschaffen wurden und die Informationen dort meistens nicht geeignet strukturiert sind.

Im Folgenden werden wir uns mit zwei spezifischen Instanzen von kollaborativen Wissensdatenbanken beschäftigen: der multilingualen freien Internet-Enzyklopädie **Wikipedia** und dem freien Internet-Wörterbuch **Wiktionary**. Wikipedia und Wiktionary wurden in jüngster Zeit als besonders vielversprechende Ressourcen identifiziert. Analog zum *Web-Mining* (Chakrabarti (2002)) bezeichnen wir die Analyse von Wiki-basierten Wissensdatenbanken **WikiMining** und unterteilen sie in die folgenden drei Bereiche, die in absteigender Relevanz für die Computerlinguistik aufgeführt werden: (i) **Mining von Inhalten**, (ii) **Mining von Struktur**, und (iii) **Mining von Nutzungsdaten**.

Wikipedia ist eine durch Benutzer erstellte elektronische Enzyklopädie, die eine intensive Verlinkung der Inhalte aufweist. Zesch *et al.* (2007) analysieren die Inhalte und die Struktur von Wikipedia und identifizieren dort verschiedene Quellen lexikalisch-semantischer Informationen, wie in Tabelle 1 dargestellt. Infolge existierender Gestaltungsrichtlinien für Autoren beinhaltet Wikipedia überwiegend Begriffe von enzyklopädischem Interesse. Größtenteils handelt es sich hierbei um Nomen sowie relativ wenige Adjektive und Verben, von denen in den meisten Fällen auf die Nomen mittels der sogenannten Weiterleitungen (Engl. **redirects**) verwiesen wird, z.B. vom Verb *"sehen"* auf den Mehrwortbegriff *"visuelle Wahrnehmung"*.

Der **erste Absatz** eines Wikipedia-Artikels beinhaltet typischerweise eine kurze Definition des im Artikel beschriebenen Begriffs. Im Volltext eines Artikels sind zahlreiche **verwandte Begriffe** enthalten, die die Bedeutung des Begriffs weiter präzisieren. Zum Teil sind auch **Übersetzungen** des Begriffs mit Links zu den entsprechenden Wikipedias in anderen Sprachen enthalten. Somit stellt Wikipedia eine vielversprechende Ressource für **multilinguale** computerlinguistische Anwendungen dar.

Eine weitere Quelle der lexikalisch-semantischen Relationen in Wikipedia sind die **Links**, die verschiedene Artikel in Wikipedia untereinander verbinden. Ein Link deutet typischerweise auf eine semantische Relation zwischen den beiden verlinkten Begriffen hin. Der Typ dieser Relation sowie ihre Stärke sind

Quelle	*Art von lexikalisch-semantischen Informationen*
Artikel	
- Erster Absatz	Definition
- Volltext	Beschreibung der Bedeutung; verwandte Begriffe; Übersetzungen
- Weiterleitungen	Synonyme; (teilweise inkorrekte) Schreibvarianten; Abkürzungen
- Titel	Eigennamen; domänenspezifische Begriffe und ihre Bedeutungen
Artikel-Links	
- Kontextfenster	verwandte oder zusammen vorkommende Begriffe;
- Label	Synonyme; Schreibvarianten; verwandte Begriffe
- Ziel-Artikel	verwandte Begriffe
Kategorien	
- dort beinhaltete Artikel	semantisch verwandte Begriffe (meistens Hyponyme)
- Hierarchie	semantische Relationen, wie Hyponyme und Meronyme
Disambiguierungsseiten	
- Artikel-Links	häufigste Bedeutung, Bedeutungsvokabular

Tabelle 1: Beispiele der lexikalisch-semantischen Informationen in Wikipedia.

jedoch nicht explizit kodiert und müssen ggf. mit computerlinguistischen Methoden automatisch erschlossen werden (Krötzsch *et al.* (2005)). Zusammen bilden alle verlinkten Begriffe und die Links einen **Artikel-Graphen**. Jeder Link hat zusätzlich ein **Label**, dessen Wortlaut sich vom verlinkten Begriff durchaus unterscheiden kann. Beispielsweise haben viele Begriffe, die auf den Artikel *"Deutschland"* verweisen, das Label *"Bundesrepublik Deutschland"*. Infolgedessen können die Labels als Quelle für **Synonyme**, **Schreibvarianten** oder andere **semantisch verwandte Begriffe** genutzt werden. Aus dem **Kontextfenster** um das Label herum können mittels computerlinguistischer Techniken weitere verwandte Begriffe gewonnen werden.

Das Kategoriensystem in Wikipedia resultiert daraus, dass jeder Artikel eine beliebige Anzahl an **semantischen Tags**, also **Kategorien** von Benutzern bekommen kann. Insofern ist das Kategoriensystem eine **Folksonomie**. Jede Kategorie kann eine beliebige Anzahl an Artikeln zugewiesen bekommen. Sie kann auch **Unterkategorien** haben, die typischerweise über die Hyponymie oder Meronymie mit der **Oberkategorie** verknüpft sind. Die Kategorie *"Fahrzeug"* hat beispielsweise Unterkategorien wie *"Luftfahrzeug"* oder *"Wasserfahrzeug"*. Insofern bildet das Kategoriensystem von Wikipedia eine Art **Thesaurus**.

Polyseme, also mehrdeutige Wörter sind in Wikipedia mittels der Disambiguierungsseiten repräsentiert. Eine **Disambiguierungsseite** listet alle Artikel auf, die für einen mehrdeutigen Begriff vorhanden sind. Da die Bezeichnung jedes Artikels eindeutig sein muss, werden die Artikel für polyseme Begriffe meistens unterschieden, indem jeder Artikel mit dem disambiguierenden Begriff in Klammern versehen wird, z.B. *"Wald"* und *"Wald (Graphentheorie)"*. Der Artikel

ohne Disambiguierungstag beschreibt zumeist die **häufigste Bedeutung** eines Begriffs. Alle aufgelisteten Bedeutungen bilden ein **Bedeutungsvokabular** für den gegebenen Begriff.

Wiktionary wird von Nutzern als multilinguales web-basiertes Wörterbuch und Thesaurus im Web kollaborativ erstellt und ist komplementär zur Online-Enzyklopädie Wikipedia. Zesch *et al.* (2008a) stellen erstmalig eine systematische Analyse von Wiktionary als computerlinguistische Ressource vor. Im Unterschied zur Wikipedia zielt Wiktionary demnach eher auf allgemeines Vokabular ab. Es deckt mehrere Wortarten ab und verzichtet auf detaillierte faktische Informationen enzyklopädischen Charakters, die in Wikipedia zu finden sind.

Im Oktober 2008 beinhaltete Wiktionary etwa 3,5 Mil. Einträge in 272 sprachspezifischen Editionen. Jede solche sprachspezifische Wiktionary-Edition beinhaltet auch Einträge für fremdsprachliche Begriffe. Folglich stellt sie ein **multilinguales Wörterbuch** mit einem substanziellen Anteil an Einträgen in Fremdsprachen dar. Das englische Wiktionary beinhaltet beispielsweise den deutschen Eintrag *"Haus"*, der mit dem englischen Eintrag *"house"* verknüpft ist. Die Größe von kollaborativ erstellten Ressourcen hängt von der Größe und dem Engagement der Internet-Gemeinde ab, die zum Projekt beiträgt. Die englische Wiktionary-Edition, die am 12. Dezember 2002 ins Leben gerufen wurde, ist die älteste, aber nicht die größte (über 900.000 Einträge im Februar 2008). Die größte Wiktionary-Edition ist die Französische, die ein Jahr später gestartet wurde (über 923.000 Einträge im Februar 2008).

Einträge in Wiktionary beinhalten ein breites Spektrum an lexikalischen und semantischen Informationen wie **Wortart**, **Wortbedeutung**, **Gloss**, **Etymologie**, **Aussprache**, **Deklination**, **Beispiele**, **Zitate**, **Übersetzungen**, **Kollokationen**, **abgeleitete Begriffe** und **Hinweise zum Sprachgebrauch**. Ebenso enthalten sind lexikalisch oder semantisch verwandte Begriffe verschiedener Art, wie **Synonyme**, **Antonyme**, **Hyperonyme** und **Hyponyme**. Darüber hinaus beinhaltet Wiktionary eine beeindruckende Menge an Informationen, die in klassischen Wissensdatenbanken nicht immer vorhanden sind. Dazu zählen **Komposita**, **Abkürzungen**, **Akronyme** und **Namensabkürzungen**, verbreitete falsche **Schreibvarianten** (z.B. Engl. *basicly - basically*), **vereinfachte Schreibvarianten** (z.B. Engl. *thru - through*), **Kontraktionen** (z.B. Engl. *o - of*), **Sprichwörter** (z.B. Engl. *no pain, no gain*), **umstrittene Wortverwendungen** (z.B. Engl. *irregardless - irrespective or regardless*), **Protologismen** (z.B. Engl. *iPodian*), **Onomatopoeia** (z.B. Engl. *grr*), und sogar **umgangssprachliche Formen** oder **Slang**. Die meisten solchen Relationen sind in Wiktionary explizit kodiert. Dies' ist ein prinzipieller Unterschied zu Wikipedia, wo die Art der semantischen Relationen zwischen Begriffen meistens mittels spezieller Verfahren inferiert werden muss. Des Weiteren muss berücksichtigt werden, dass sprachspezifische Wiktionary-Editionen uneinheitlich strukturiert sind. Z.B. enthält das deutsche Wiktionary im Unterschied zum englischen Wiktionary charakteristische Wortkombinationen, jedoch keine Zitate, die in der englischen Version sehr wohl vorhanden sind.

Ähnlich wie in Wikipedia, sind Wiktionary-Einträge zusätzlich mit **Kategorien** versehen. Schließlich sind die Einträge massiv mit anderen Einträgen verlinkt, sowohl innerhalb einer sprachspezifischen Wiktionary-Edition als auch sprachenübergreifend. Die Links verweisen zusätzlich auf weitere externe Wissensdatenbanken oder Web-basierte Wörterbücher.

Programmatischer Zugriff auf Wikipedia und Wiktionary Die Nutzung von Wikipedia und Wiktionary in computerlinguistischen Anwendungen bedarf effizienter Methoden für den strukturierten Zugriff auf die dort enthaltenen Informationen. Zesch *et al.* (2008a) beschreiben eine Reihe von spezialisierten Werkzeugen für den Zugriff auf Wikipedia und stellen einen optimierten Ansatz vor, bei dem die Inhalte von Wikipedia und Wiktionary zunächst in eine Datenbank importiert werden. So können spezielle Funktionalitäten von Datenbanken, z.B. eine effiziente Indexierung von Inhalten, voll ausgenutzt werden. Die für die computerlinguistischen Anwendungen relevanten Informationen, wie Links oder Kategorien, werden explizit auf ein Datenbankschema abgebildet. Das ermöglicht einen verbesserten Zugriff auf diese Informationen in den darauf aufbauenden computerlinguistischen Anwendungen. Die Java-basierten Programmierschnittstellen **JWPL** (Zesch (2008)) und **JWKTL** (Müller (2008)) sind für die nicht-kommerzielle Forschung frei verfügbar.

1.4 Sprachverarbeitungsanwendungen mit Nutzung von WWW als Ressource

Lapata and Keller (2005) geben einen Überblick darüber, in welchen computerlinguistischen Anwendungen das Web als Ressource eingesetzt wurde. Dazu zählen beispielsweise maschinelle Übersetzung, Entdeckung von semantischen Relationen, Disambiguierung von Wortlesarten und die Beantwortung natürlichsprachlicher Fragen. In **maschineller Übersetzung** diente das Web als Quelle bilingualer Korpora sowie zur Nachbearbeitung von Übersetzungskandidaten. Andere Arbeiten **entdecken semantische Relationen** wie Hyponymy, Ähnlichkeit, Antonymy oder logische Folgerung mittels lexikalisch-semantischer Muster als Suchanfragen.

Bisherige Anwendungen von Wikipedia und Wiktionary in der computerlinguistischer Forschung sind exemplarisch von Zesch *et al.* (2008a) beschrieben. So setzen Gabrilovich and Markovitch (2006) Wikipedia für die Aufgabe der **automatischen Textklassifikation** ein. Ruiz-Casado *et al.* (2005) befassen sich mit **automatischer Informationsextraktion** und beschreiben einen Ansatz, um Wikipedia-Einträge automatisch mit Konzepten in einer Ontologie oder einem lexikalisch-semantischen Wortnetz zu verknüpfen. Ahn *et al.* (2004) verwenden das Wissen aus Wikipedia im Rahmen des *TREC 2004* ***Question Answering*** Wettbewerbs sowohl als eine Quelle für Antworten auf faktische Fragen. Es existieren dagegen nur noch wenige Arbeiten, in denen **Wiktionary** als eine lexikalisch-semantische Wissensdatenbank verwendet wird. Chesley *et al.* (2006) verwenden Adjektive aus Wiktionary für eine Analyse der **Orientierung von**

Meinungen in Blogs. Eine weitere Arbeit, die Wiktionary verwendet, ist im Bereich der **diachronischen Phonologie** (Bouchard *et al.* (2007)). Zesch *et al.* (2008b) verwenden sowohl Wikipedia als auch Wiktionary für die Berechnung der **semantischen Verwandtschaft** zwischen zwei Wörtern. Schließlich wurde das kombinierte Wissen aus Wikipedia und Wiktionary in jüngster Zeit für die Verbesserung **natürlichsprachlicher Informationsrecherche** eingesetzt (Müller and Gurevych (2008)).

1.5 Computerlinguistik und Sprachtechnologie für das Web

Das vorliegende Buchkapitel fokussierte sich auf das Potenzial des World Wide Web als computerlinguistische Ressource. Mit dem Wachstum des WWW werden jedoch computerlinguistische Methoden und Werkzeuge zu einer unabdingbaren Voraussetzung, um dem Benutzer einen **effizienten Umgang** mit explosionsartig anwachsenden Mengen an Informationen zu ermöglichen. Die Globalisierung der Informationsflut führte mittlerweile dazu, dass **Suchmaschinen** zu einer Grundsatztechnologie geworden sind. Die primäre Aufgabe von Suchmaschinen ist es, die im Web vorhandenen Dokumente nach ihrer Relevanz zur Benutzeranfrage zu ordnen. Für ein optimales Suchergebnis ist es jedoch notwendig, **Informationen über den Benutzer** beim Ranking zu berücksichtigen. Solche Informationen können wiederum mit computerlinguistischen Methoden aus Benutzer-generierten Inhalten gewonnen werden. Eine weitere wichtige Technik ist die Bereinigung der im Retrieval eingesetzten zugrundeliegenden Dokumentenbasis. Hier ist es wichtig, **Spam und Duplikate zu identifizieren** und zu entfernen. Mit der Verbreitung von Sozialer Software hat die Aufgabe der **Qualitätsbewertung der Inhalte** stark an Bedeutung gewonnen. Da es keine redaktionelle Kontrolle über die Inhalte im Web gibt, müssen automatische Verfahren entwickelt werden, um vertrauenswürdige Informationen hoher Qualität dem Benutzer vorrangig anzubieten. Eine weitere Herausforderung, bei deren Bewältigung computerlinguistische Verfahren eine Schlüsseltechnologie darstellen, ist die **kontextbezogene Aufbereitung** und die **Präsentation der Informationen** für den Benutzer. Insbesondere mit starker Zunahme von Technologien des *Ubiquitous Computing* (Mühlhäuser and Gurevych (2007)) beim Zugriff auf sprachliche Informationen müssen computerlinguistische Verfahren für die **automatische Zusammenfassung** im Hinblick auf verschiedene Geräte, Formate, Inhalte und weitere Arten von Präsentationskontexten weiterentwickelt und optimiert werden.

Zusammenfassend kann festgehalten werden, dass das Verhältnis zwischen dem WWW und der Computerlinguistik **dualer Natur** ist. Die Computerlinguistik profitierte enorm vom **WWW als Ressource**, indem (i) die dort abrufbaren Informationen als ein einzigartiges, multilinguales Korpus für computerlinguistische Verfahren eingesetzt werden, und (ii) die dort kollaborativ entstehenden Wissensdatenbanken, wie z.B. Wikipedia und Wiktionary, als semantisch strukturierte und teilweise ausgezeichnete Korpora anstelle von konventionellen lexikalisch-semantischen Ressourcen und Korpora eingesetzt werden. Andererseits stellt das WWW ein **äußerst attraktives Anwendungsge-**

biet für mehrere zentrale Verfahren der angewandten Computerlinguistik und der Sprachtechnologie mit einem enormen wissenschaftlichen und wirtschaftlichen Potenzial dar. Dies' sichert der Computerlinguistik als Forschungsgebiet eine Schlüsselrolle in Gesellschaft und Politik.

1.6 Literaturhinweise

Umfangreiche weiterführende Informationen, u.a. die Links zu den Online-Proceedings der einschlägigen Workshops sowie zur Software können auf den folgenden Webseiten abgerufen werden WAC (2008); SIGWAC (2008). Vor zwei Jahren stellte Google eine Kollektion mit den aus dem Web (ca. 1 Billion Token) gewonnenen N-Grammen über das *Linguistic Data Consortium* zur Verfügung. Das Material zur Nutzung von Wikipedia als computerlinguistische Ressource ist ebenso im Web zu finden[2] sowie in den Online-Proceedings (Bunescu *et al.* (2008)). Die Webseite vom *Ubiquitous Knowledge Processing Lab* enthält einige wichtige Publikationen zu kollaborativen Wissensdatenbanken in der Computerlinguistik (Zesch *et al.* (2007, 2008a,b)) sowie die dazugehörige Software (Zesch (2008); Müller (2008)).

Literatur

Ahn, D., Jijkoun, V., Mishne, G., Müller, K., de Rijke, M., and Schlobach, S. (2004). Using Wikipedia at the TREC QA Track. In *Proceedings of TREC 2004*.

Baroni, M. and Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. In *Proceedings of Fourth International Conference on Language Resources and Evaluation*, pages 1313–1316, Lisbon, Portugal.

Bouchard, A., Liang, P., Griffiths, T., and Klein, D. (2007). A probabilistic approach to diachronic phonology. In *Proceedings of Conference on Empirical Methods in Natural Language Processing and Conference on Computational Natural Language Learning*, pages 887–896, Prague, Czech Republic.

Bunescu, R., Gabrilovich, E., and Mihalcea, R. (2008). AAAI 2008 Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy . URL `http://lit.csci.unt.edu/~wikiai08/`.

Chakrabarti, S. (2002). *Mining the Web: Discovering Knowledge from Hypertext Data*. Morgan-Kauffman.

Chesley, P., Vincent, B., Xu, L., and Srihari, R. (2006). Using Verbs and Adjectives to Automatically Classify Blog Sentiment. Technical Report SS-06-03, AAAI Spring Symposium.

[2] `http://en.wikipedia.org/wiki/Wikipedia:Wikipedia_in_academic_studies`, `http://en.wikipedia.org/wiki/Wikipedia:Wikipedia_as_an_academic_source`, `http://meta.wikimedia.org/wiki/Wiki_Research_Bibliography`

Ebbertz, M. (2002). Web Languages. URL `http://www.netz-tipp.de/sprachen.html`.

Ferraresi, A., Zanchetta, E., Baroni, M., and Bernardini, S. (2008). Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google?*, Marrakech, Morocco.

Gabrilovich, E. and Markovitch, S. (2006). Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge. In *AAAI*, pages 1301–1306, Boston, MA.

Gruber, T. (2008). Collective knowledge systems: Where the Social Web meets the Semantic Web. *Web Semantics: Science, Services and Agents on the World Wide Web*, **6**(1), 4–13.

Keller, F. and Lapata, M. (2003). Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, **29**, 459–484.

Krötzsch, M., Vrandecic, D., and Völkel, M. (2005). Wikipedia and the Semantic Web – the missing links. In *Proceedings of First International Wikimedia Conference – Wikimania 2005*, Frankfurt, Germany.

Lapata, M. and Keller, F. (2005). Web-based models for natural language processing. *ACM Transactions on Speech and Language Processing*, **2**, 1–31.

Mühlhäuser, M. and Gurevych, I. (2007). *Handbook of Research on Ubiquitous Computing Technology for Real Time Enterprises*. IGI Global, Hershey PA, USA.

Müller, C. (2008). JWKTL: Java-based Wiktionary API. URL `http://www.ukp.tu-darmstadt.de/software/jwktl/`.

Müller, C. and Gurevych, I. (2008). Using Wikipedia and Wiktionary in domain-specific information retrieval. In F. Borri, A. Nardi, and C. Peters, editors, *Working Notes for the CLEF 2008 Workshop*.

Ruiz-Casado, M., Alfonseca, E., and Castells, P. (2005). Automatic Assignment of Wikipedia Encyclopedic Entries to WordNet Synsets. *Advances in Web Intelligence*, pages 380–386.

SIGWAC, A. (2008). The Special Interest Group of the Association for Computational Linguistics (ACL) on Web as Corpus. URL `http://www.sigwac.org.uk/`.

WAC (2008). The Web as Corpus Website. URL `http://webascorpus.sourceforge.net/`.

Xu, J. (2000). Multilingual search on the World Wide Web. In *Presentation to HICSS-33*, Maui, Hawaii.

Zesch, T. (2008). JWPL: Java-based Wikipedia API. URL `http://www.ukp.tu-darmstadt.de/software/jwpl/`.

Zesch, T., Gurevych, I., and Mühlhäuser, M. (2007). Analyzing and accessing Wikipedia as a lexical semantic resource. In *Data Structures for Linguistic Resources and Applications*, pages 197–205, Tuebingen, Germany. Gunter Narr, Tübingen.

Zesch, T., Müller, C., and Gurevych, I. (2008a). Extracting lexical semantic knowledge from wikipedia and wiktionary. In *Proceedings of the Conference on Language Resources and Evaluation (LREC), electronic proceedings*, Marrakech, Morocco.

Zesch, T., Müller, C., and Gurevych, I. (2008b). Using Wiktionary for computing semantic relatedness. In *Proceedings of Twenty-Third AAAI Conference on Artificial Intelligence*, pages 861–867, Chicago, Illinois.