Aufbereitung und Strukturierung von Information mittels automatischer Sprachverarbeitung

Dr. Wolfgang Stille, Nicolai Erbs, Dr. Torsten Zesch,
Prof. Dr. Iryna Gurevych, UKP Lab, TU Darmstadt[1],
Prof. Dr. Karsten Weihe, Algorithmik, TU Darmstadt[2]

Abstract. Unternehmenserfolg hängt wesentlich von der optimalen Nutzung des im Unternehmen vorhandenen Wissens ab. Gleichwohl wächst die Menge an unstrukturierter Information jährlich rapide an. Unternehmen und ihre Mitarbeiter sind folglich auf intensive Recherche und Aufbereitung von Wissen und deren verständliche und gut strukturierte Präsentation angewiesen, oftmals unter engen Zeitvorgaben. Insbesondere bei Entscheidungsprozessen haben Qualität, Struktur und Aufbereitung der zugrunde liegenden Information weitreichende Konsequenzen. Methoden der Sprachverarbeitung können die strukturelle Aufbereitung von Information unterstützend begleiten oder gar vollständig automatisieren.

1. Motivation

Alle fünf Jahre verzehnfacht sich die Menge digitaler Information [GCM08]. Insbesondere die Menge von Textdokumenten, z.B. in Form von Emails, Pressemitteilungen, Blogs, Foren, etc. wächst rapide an. Daher ist ein wachsender Anteil von Erwerbstätigen auf allen Ebenen in ihrer Arbeit mit intensiver Recherche und Präsentation ihres Ergebnisses befasst: Suche, Kategorisierung, Priorisierung, Strukturierung, Zusammenfassung und Aufbereitung von Information, häufig unter strengen Zeitvorgaben, bestimmen den Arbeitsalltag. Insbesondere bei Entscheidungsprozessen aller Art hat das Ergebnis dieser Prozesse weitreichende Konsequenzen.

[1] Ubiquitous Knowledge Processing Lab, Hochschulstr. 10, 64289 Darmstadt;
E-Mail: {stille,n_erbs,zesch,gurevych}@tk.informatik.tu-darmstadt.de
[2] Fachgebiet Algorithmik, Hochschulstr. 10, 64289 Darmstadt;
E-Mail: weihe@informatik.tu-darmstadt.de

Daraus ergeben sich Herausforderungen insbesondere in Unternehmen, deren wirtschaftlicher Erfolg zu einem sehr großen Teil von optimaler Nutzung des im Unternehmen angesammelten Wissens abhängt.

Die für Recherchen relevanten elektronischen Informationsquellen werden außerdem zunehmend komplexer und heterogener. Relevant sind nicht nur mehr herkömmliche Textdokumente, sondern beispielsweise auch Wissensdatenbanken, Protokolle, Emails, Foren und Blogs. Diese Quellen enthalten ebenfalls hochgradig relevantes Wissen, welches jedoch aufgrund ihres Umfangs, ihrer Komplexität und ihrer Heterogenität nur noch mittels Unterstützung durch intelligente Systeme zur Informationsaufbereitung und –strukturierung nutzbar gemacht werden kann.

Die zentrale Anforderung in diesem Zusammenhang ist folglich die Bereitstellung rechnergestützter Werkzeuge, welche die Aufbereitung von Informationen und ihre Strukturierung unterstützend begleiten oder gar vollständig automatisieren. Wir beschäftigen uns im Folgenden mit einer Vielzahl von Techniken der automatischen Sprachverarbeitung, welche bei diesen Aufgaben zur Anwendung kommen.

2. Fallbeispiel: Erstellung von Tischvorlagen

In vielen Situationen müssen Entscheidungsträger innerhalb eines knappen Zeithorizonts optimal mit relevanten, komprimierten und gut strukturierten Informationen versorgt werden. Darüber hinaus kann eine Anreicherung mit nicht textuellen Elementen hilfreich sein, um eine konstante Aufnahmefähigkeit der Zuhörer zu gewährleisten.

Auf der anderen Seite stehen die Organisatoren von Meetings, die sich mit einer stetig wachsenden Menge an Information in Unternehmen konfrontiert sehen, und oftmals in kurzer Zeit große Mengen von Informationen aufbereiten und strukturieren müssen. Bei diesen Prozessen sollen sie bestmöglich durch Software unterstützt werden.

Ausgangsposition unseres Szenarios ist folglich ein relativ unstrukturiertes Einzeldokument, bzw. eine thematisch kohärente Sammlung von Dokumenten oder gar Informationsschnipseln aus diversen Quellen innerhalb des Unternehmens.

Ziel ist es, ein knappes, gut strukturiertes und übersichtliches Dokument in Form einer Tischvorlage mit automatischer Unterstützung zu generieren. Mit Hilfe von Verfahren aus der Sprachverarbeitung werden Querverweise eingefügt, wichtige Begriffe markiert und das Dokument sowohl hierarchisch in Form einer Gliederung als auch linear in Form einer semantischen Absatzformatierung strukturiert. Im Falle sehr umfangreicher

Rohdaten können diese in einem Vorverarbeitungsschritt automatisch zusammengefasst werden und redundantes Material entfernt werden. Auf Wunsch kann das resultierende Dokument noch mit Bildmaterial und Concept Maps angereichert werden, um die Zuhörer bzw. Leser optimal mit multimodaler Information zu versorgen.

3. Wissensaufbereitung mittels Sprachverarbeitung

Im Folgenden werden die verwendeten Methoden zur Aufbereitung und Strukturierung von in Textform vorliegender Information im Einzelnen vorgestellt. Sie beruhen auf dem „Darmstadt Knowledge Processing Software Repository“ (DKPro[3]) [EG09], welches eine Sammlung von modularen, skalierbaren und robusten Sprachverarbeitungswerkzeugen basierend auf dem Apache UIMA Standard [FL04] darstellt.

Das vorgestellte Framework ist flexibel einsetzbar, unabhängig von Art und Umfang der ihm vorliegenden Textinformationen. Die Eingabe kann auf verschiedene Arten erzeugt werden: manuell, z.B. in Form eines fertigen, aber relativ unstrukturierten Dokuments oder einer Sammlung von thematisch relevanten Dokumenten. Ebenso ist eine (semi-)automatisierte Generierung einer Menge thematisch relevanter und kohärenter Information aus Dokumentensammlungen, Wissensdatenbanken, Emails, Blogs, etc. mit Methoden des Information Retrieval denkbar. Diese Informationen in digitaler Textform stellen das Rohmaterial dar, von dem wir im Folgenden ausgehen.

3.1 Kompression des Rohmaterials

Je nach Umfang des Rohmaterials muss dieses komprimiert werden, und zwar nach Möglichkeit verlustfrei, also ohne nennenswerte Einbußen in seinem Informationsgehalt. Dies kann dadurch erreicht werden, dass redundante Informationen – wie sie im Normalfall bei der Verwendung unterschiedlicher Datenquellen zu einem Thema zustande kommen - entfernt werden. Dies geschieht mit Algorithmen zum Auffinden von Duplikaten bzw. sehr ähnlichen Textpassagen [GM07]. Eine weitere Kompression erreicht man durch die Anwendung von Verfahren zur automatischen Zusammenfassung von Texten [BES97]. Als Nebenprodukt der Duplikatentfernung und Zusammenfassung kann hierbei auf Quellen

[3] Open-source Projektseite: http://code.google.com/p/dkpro-core-asl

verwiesen werden, die sehr ähnliche oder weiterführende Informationen enthalten („Weiterführende Literatur“).

3.2 Textsegmentierung

Textsegmentierung bezeichnet die Aufteilung eines Textes in thematisch zusammenhängende Abschnitte. Mit Hilfe unterschiedlicher Verfahren zur Erkennung lexikalischer Kohäsion können Texte vollautomatisch gegliedert werden [Hea97]. Neben bekannten Verfahren, die lexikalische Ketten oder Clustering verwenden, wurden am UKP Lab vielversprechende Methoden entwickelt, die auf semantischen Graphen basieren [Mar10].

Neben der Unterteilung in Absätze mittels linearer Textsegmentierung wird mit Hilfe hierarchischer Segmentierung eine Baumstruktur erzeugt, die mehrere Granularitätsebenen (Kapitel, Hauptabschnitte, Unterabschnitte, Absätze, ...) beinhaltet.

3.3 Generierung von Überschriften und Inhaltsverzeichnissen

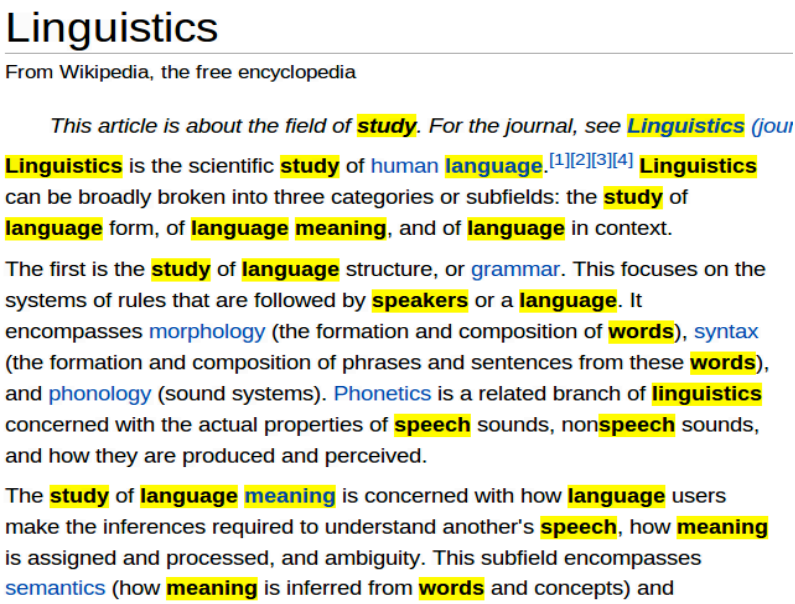
Auf Basis der Methoden zur Textsegmentierung und Erkennung von Schlüsselphrasen können zur weiteren Strukturierung des Dokuments abschnittsweise Überschriften erzeugt werden [BDB07]. Hierzu werden die Schlüsselphrasen innerhalb der einzelnen Abschnitte analysiert und bewertet. Die hierarchische Textsegmentierung zusammen mit den erzeugten Überschriften stellt das Grundgerüst für ein Inhaltsverzeichnis dar.

3.4 Erkennung und Hervorhebung von Schlüsselphrasen

Das Hervorheben von Schlüsselphrasen in einem Text dient dem schnelleren Auffassen der Inhalte. Besonders wichtige Wörter oder Phrasen werden mit Hilfe von lexikalischen oder statistischen Verfahren der Sprachverarbeitung oder mit Methoden des maschinellen Lernens identifiziert [WPF99]. Abb. 1 illustriert eine solche Hervorhebung von Schlüsselphrasen anhand eines Beispieltextes aus Wikipedia.

3.5 Erzeugung von Querverweisen

Für eine Verlinkung innerhalb von Dokumenten müssen Anker und mögliche Ziele, auf welche die Anker verweisen, identifiziert werden. In Dokumenten, die bereits Querverweise enthalten, können die vorhandenen Anker bewertet werden und weitere Anker auf Basis dieser Bewertung iden-



Linguistics

From Wikipedia, the free encyclopedia

This article is about the field of ***study****. For the journal, see* ***Linguistics*** *(jou*

Linguistics is the scientific **study** of human **language**.[1][2][3][4] **Linguistics** can be broadly broken into three categories or subfields: the **study** of **language** form, of **language** **meaning**, and of **language** in context.

The first is the **study** of **language** structure, or grammar. This focuses on the systems of rules that are followed by **speakers** or a **language**. It encompasses morphology (the formation and composition of **words**), syntax (the formation and composition of phrases and sentences from these **words**), and phonology (sound systems). Phonetics is a related branch of **linguistics** concerned with the actual properties of **speech** sounds, non**speech** sounds, and how they are produced and perceived.

The **study** of **language** **meaning** is concerned with how **language** users make the inferences required to understand another's **speech**, how **meaning** is assigned and processed, and ambiguity. This subfield encompasses semantics (how **meaning** is inferred from **words** and concepts) and

Abb. 1: Vollautomatische Hervorhebung von Schlüsselphrasen

tifiziert werden. Ebenso werden Ziele für die identifizierten Anker gesucht. Falls keinerlei Querverweise in Dokumenten vorhanden sind, werden Nominalphrasen als potentielle Anker verwendet und anhand ihres Informationsgehaltes gewichtet. Ziele werden über die semantische Ähnlichkeit zwischen dem Anker, dem Ursprungstexts, und dem potentiellen Ziel gewählt [HBZ99]. Dies ist insbesondere hilfreich im Falle mehrdeutiger Anker. Abb. 2 illustriert die semi-automatische Verlinkung eines Beispieltextes aus Wikipedia. Hier erhält der Benutzer für jeden Anker einen Vorschlag mehrerer möglicher Ziele, von denen er eines auswählen kann.

3.6 Erstellung von Concept Maps

Im Gegensatz zu Mind Maps, die als Unterstützung beim Brainstorming verwendet werden, sind Concept Maps ein Mittel zur graphischen Darstellung von Wissensstrukturen und damit gut zur Darstellung der inhaltlichen Bezüge eines Dokumentes geeignet. Concept Maps unterstützen insbesondere Wahrnehmung und Verständnis von komplexen Zusammenhängen in Texten, indem sie Relationen zwischen Entitäten gra-

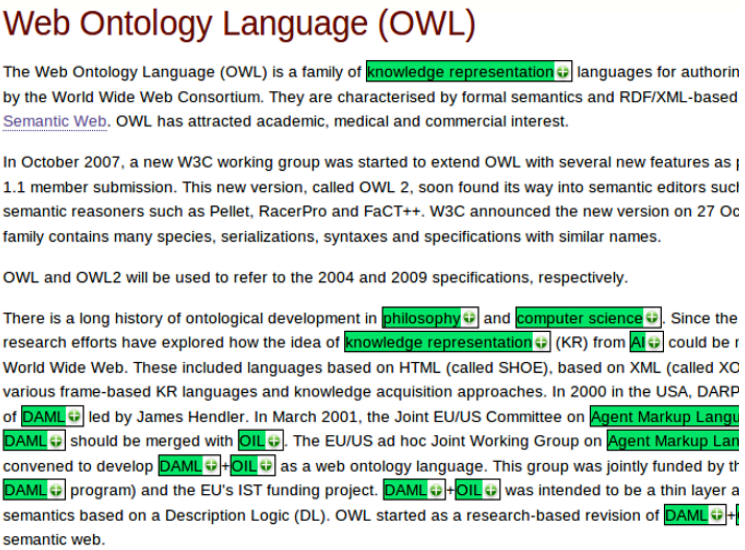


Abb. 2: Semi-automatische Erzeugung von Querverweisen

phisch illustrieren [ACB07]. Diese werden mit Hilfe von Verfahren der Sprachtechnologie aus einzelnen Abschnitten eines Dokuments extrahiert. Der Einsatz von Optimierungsalgorithmen aus dem Bereich des Graph Drawing [TBD88] garantiert die optimale Lesbarkeit der generierten Concept Maps.

3.7 Einfügen von geeignetem Bildmaterial

Oftmals befindet sich im Intranet von Unternehmen auch umfangreiches Bildmaterial, welches zur visuellen Anreicherung und somit zur besseren Präsentation von Informationen und Auflockerung eines Dokuments verwendet werden kann. Zu diesem Zweck wird eine Gliederung des Textes (z.B. mittels Methoden der Textsegmentierung) erzeugt, und für jeden Abschnitt wird semantisch passendes Bildmaterial eingefügt [SRC 10].

4. Zusammenfassung und Ausblick

Methoden der automatischen Sprachverarbeitung eröffnen vielfältige und umfangreiche Möglichkeiten zur Aufbereitung und Strukturierung von Wissen. Insbesondere helfen Sie dabei, die Informationsflut in Unternehmen in den Griff zu bekommen und aus umfangreichen und unstruk-

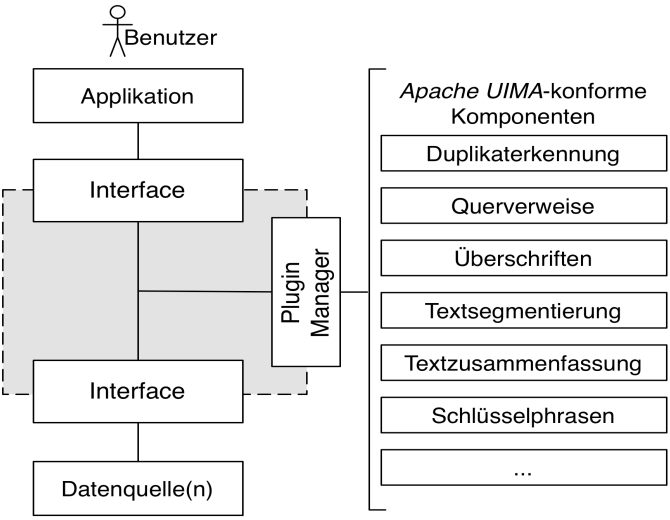


Abb. 3: Modulare Architektur prototypischer Systeme

turierten Quellen strukturierte und gut lesbare Informationen zu gewinnen.

Hier existieren bereits prototypische Systeme, welche durch das modulare Konzept vergleichsweise leicht in bestehende Umgebungen eingebettet werden können (siehe Abb.3). Wir illustrieren dies, indem wir auf der KnowTech 2011 ein System vorführen, welches exemplarisch einige der genannten Technologien in Wikis integriert.[4]

Im Bereich der Aufbereitung und Strukturierung von Information mittels automatischer Sprachverarbeitung existieren noch viele interessante Fragestellungen, deren Bearbeitung eine genaue Analyse der im Kontext real ablaufenden Arbeitsprozesse bedarf. Dank des modularen Konzepts der Kernkomponenten zur Sprachverarbeitung sind hier der Kombination und Integration verschiedener Komponenten je nach Anwendungsszenario kaum Grenzen gesetzt.

Literatur

[ACB07] Ahmad, F.; de la Chica, Sebastian; Butcher, K.; Sumner, T.; Martin, J.: Towards Automatic Conceptual Personalization Tools, Conference on Digital Libraries, 2007, S. 452-461.

[4] http://www.ukp.tu-darmstadt.de/research/current-projects/wikulu/

[BDB07] Branavan, S.; Deshpande, P.; Barzilay. R.: Generating a Table-of-Contents. Association for Computational Linguistics, Band 45, 2007, S. 544-551.

[BES97] Barzilay, R.; Elhadad, M: Using Lexical Chains for Text Summarization. ACL Workshop on Intelligent Scalable Text Summarization, 1997, S. 10–17.

[EG09] Eckart de Castilho, R.; Gurevych, I.: DKPro-UGD: A Flexible Data-Cleansing Approach to Processing User-Generated Discourse, First French-Speaking Meeting around the Framework Apache UIMA, 2009.

[FL04] Ferrucci, D.; Lally, A.: UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment, Natural Language Engineering, Band 10/3-4, 2004, S. 327-348.

[GCM08] Gantz, J.; Chute, C.; Manfrediz, A.; Minton, S.; Reinsel, D.; Schlichting, W.; Toncheva, A.: The Diverse and Exploding Digital Universe, IDC White Paper, 2008.

[GM07] Gabrilovich, E.; Markovitch, S.: Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis, 20th International Joint Conference on Artificial Intelligence, 2006, S. 1606-1611.

[HBZ09] Hoffart, J.; Bär, D; Zesch, T.; Gurevych, I.: Discovering Links Using Semantic Relatedness. INEX 2009 Workshop, 2009, S. 314-325.

[Hea97] Hearst, M.: TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages, Computational Linguistics, Band 23/1, 1997, S. 33-64.

[Mar10] Martin, M.: Applying Graph Algorithms to Text Segmentation, Bachelor Thesis, Darmstadt University of Technology, 2010.

[SRC 10] Schwarz, K.; Rojtberg, P.; Caspar, J.; Gurevych, I.; Goesele, M.; Lensch, H.: Text-to-Video: Story Illustration from Online Photo Collections, Knowledge-Based and Intelligent Information and Engineering Systems, 2010, S. 402-409.

[TBD88] Tamassia, R.; Batini, C. ; Di Battista, G.: Automatic graph drawing and readability of diagrams. IEEE Transactions on Systems, Man & Cybernetics, 18(1), 1988, S. 61-79.

[WPF99] Witten, I.; Paynter, G.;Frank, E.;Gutwin, C.; Nevill-Manning, C.: KEA: Practical Automatic Keyphrase Extraction, Fourth ACM Conference on Digital Libraries, 1999, S. 254-255.