

BACHELORARBEIT

Erweiterung des GERBIL-Frameworks zur Evaluation von Named-Entity-Linking Verfahren



Universität Potsdam
Institut für Informatik
Hasso-Plattner-Institut für Softwaresystemtechnik GmbH
Professur Internet-Technologien und Systeme

vorgelegt von

Henrik Jürges
Matr.-Nr. 751237

Aufgabenstellung und Betreuung:
Prof. Christoph Meinel
Dr. Harald Sack

Potsdam
29. April 2016

Kurzzusammenfassung

Named Entity Linking (NEL) ist ein Verfahren, um Text in natürlicher Sprache mit formalen Wissenbasen wie z. B. DBpedia zu verknüpfen. Antwortsystem, Suchmaschinen und Empfehlungssystem nutzen Verfahren zur automatisierten Textverarbeitung. Zur Verbesserung dieser Verfahren werden dabei Mehrdeutigkeiten von Texten in natürlicher Sprache aufgelöst und in formales Wissen überführt. Daneben ist vor allem die Evaluation von NEL-Verfahren ein aktueller Gegenstand der Forschung.

Diese Bachelorarbeit hat zum Ziel, bestehende NEL-Evaluationsverfahren zu verfeinern, um daraus differenzierte Aussagen über die Effektivität von NEL-Methoden und deren Evaluation treffen zu können. Dafür wird GERBIL, ein NEL-Evaluationsframework, analysiert und existierende Defizite aufgezeigt und klassifiziert. Für die Defizite werden Lösungen und Erweiterungen vorgeschlagen und implementiert.

Inhaltsverzeichnis

A	Abkürzungen und Akronyme	I
B	Präfixdefinitionen	II
C	Abbildungsverzeichnis	III
D	Tabellenverzeichnis	IV
1	Einleitung	1
1.1	Die Gliederung	1
1.2	Inhalt der Einleitung	1
2	Aufbau und Inhalt der wissenschaftlichen Arbeit	3
2.1	Verwandte Arbeiten und wissenschaftlicher Hintergrund (Related Work) .	3
2.2	Eigener (wissenschaftlicher) Beitrag	3
2.3	Evaluation des (wissenschaftlichen) Beitrags	4
2.4	Diskussion der Evaluationsergebnisse	4
2.5	Zusammenfassung und Ausblick	5
3	Nützliches und Wissenswertes zum Erstellen einer Seminararbeit	6
3.1	Allgemeine Hinweise	6
3.2	Weitere Formatierungshinweise	7
3.2.1	Abbildungen	7
3.2.2	Tabellen	8
3.2.3	Listings, Quellcode und Pseudocode	8
3.2.4	Mathematische Formeln	9
3.2.5	L ^A T _E X und andere Textverarbeitungssysteme	10
4	Inhaltliche Bestandteile der Seminararbeit	11
4.1	Gliederungspunkte	11
4.2	Inhalt der Gliederungspunkte	11
4.3	Umfang der Gliederungspunkte	12
4.4	Logischer Zusammenhang	12
5	Das Literaturverzeichnis und die korrekte Zitierweise	13
5.1	Was wird zitiert?	13
5.2	Bibtex das Zitiersystem von L ^A T _E X	13
5.3	Zitieren und das Internet	14
5.4	Zitieren und die Wikipedia	14
6	Zusammenfassung und Ausblick	15
E	Glossar	16

F	Abkürzungen und Akronyme	17
	Literaturverzeichnis	18

A Abkürzungen und Akronyme

A2KB	Annotate to knowledge base
ASCII	American Standard Code for Information Interchange
C2KB	Concept to knowledge base
D2KB	Disambiguate to knowledge base
FTP	File Transfer Protocol
HITS	Hypertext-Induced Topic Search
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
IATA	International Air Transport Association
IE	Informationsextraktion
IEC	International Electrotechnical Commission
IRI	Internationalized Resource Identifiers
ISO	International Organization for Standardization
LOD	Linked Open Data
NED	Named Entity Disambiguation
NEL	Named Entity Linking
NEN	Named Entity Normalisation
NER	Named Entity Recognition
NIL	Not in List
NLP	Natural Language Processing
RDF	Resource Description Framework
RDFS	Resource Description Framework Schema
OWL	OWL 2 Web Ontology Language
SPARQL	SPARQL Protocol and RDF Query Language
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
W3C	World Wide Web Community
WWW	World Wide Web
XML	Extensible Markup Language

B Präfixdefinitionen

Im Verlauf dieser Arbeit werden verschiedene Internationalized Resource Identifiers in abgekürzter Form benutzt. Im Folgenden sind diese aufgezählt:

dbo:	< http://dbpedia.org/ontology/ >
dbr:	< http://dbpedia.org/resource/ >
rdf:	< http://www.w3.org/1999/02/22-rdf-syntax-ns# >
rdfs:	< http://www.w3.org/2000/01/rdf-schema# >
owl:	< http://www.w3.org/2002/07/owl# >
foaf:	< http://xmlns.com/foaf/0.1/ >

Abbildungsverzeichnis

1	Das Umfeld des 3-SAT Problems	7
2	So sollte eine Abbildung nicht aussehen (JPG-Grafik, „copy and paste“)	8
3	Einbinden von Quellcode in die Arbeit	9

Tabellenverzeichnis

1	Die Geschichte des World Wide Web	9
---	---	---

1 Einleitung

Mit den vorliegenden Hinweisen versuchen wir Ihnen einen Leitfaden zum Erstellen wissenschaftlicher Abschlussarbeiten, d.h. Seminar-, Bachelor-, Master-, etc. Arbeiten im Fach Informatik an die Hand zu geben. Wie Sie sicher schon beim Lesen wissenschaftlicher Arbeiten bemerkt haben werden, folgen diese meist einem einheitlichen Aufbau. Dies liegt nicht daran, dass die Autoren sich keine Mühe geben würden bzw. große Langweiler sind, denen eben nichts Neues einfallen würde. Nein, ein einheitlicher Aufbau erleichtert dem Leser – der meist nie besonders viel Zeit hat bzw. investieren möchte – die wesentlichen Beiträge der Arbeit schnell und effizient zu erfassen.

1.1 Die Gliederung

Aber wie gliedert man eine wissenschaftliche Arbeit? Meist kommt dabei das folgende einfache Schema zum Einsatz:

1. Einleitung
2. Verwandte Arbeiten und wissenschaftlicher Hintergrund (Related Work)
3. Eigener (wissenschaftlicher) Beitrag – dies können gerne mehrere Kapitel werden...
4. Evaluation des (wissenschaftlichen) Beitrags
5. Diskussion der Evaluationsergebnisse
6. Zusammenfassung und Ausblick

Auf die Eigenheiten der einzelnen Unterpunkte werden wir im Folgenden noch genauer eingehen. Beginnen wir einfach mit der Einleitung.

1.2 Inhalt der Einleitung

Die Einleitung soll den Leser zum **Thema hinführen**, die Arbeit in einen Gesamtzusammenhang einordnen und einen kurzen Überblick über den Inhalt der Arbeit geben. Dabei sind die folgenden Punkte besonders wichtig:

- Motivation des Themas – warum ist das Thema überhaupt von Bedeutung?
- Wie ordnet sich das Thema in einen größeren Gesamtzusammenhang (z.B. den Rahmen des Seminars bzw. in aktuelle Forschungsprobleme) ein?
- Darlegung der grundlegenden Fragestellungen: Worum geht es eigentlich? Was sind die zentralen Fragestellungen? Wie beabsichtigen wir diese zu lösen?
- Warum lohnt sich das Weiterlesen?

Wichtig ist, dass die Einleitung die **Dramaturgie** der Arbeit quasi wie einen „roten Faden“ sichtbar werden lässt.

Am Ende der Einleitung sollte ein **kurzer Überblick über den Inhalt** der einzelnen Kapitel folgen. Die vorliegende Arbeit könnte wie folgt skizziert werden:

Kapitel 2 gibt Hinweise zum Aufbau einer wissenschaftlichen Arbeit und wie deren Inhalte zu gestalten sind. Kapitel 3 gibt allgemeine Hinweise zur Formatierung von wissenschaftlichen Arbeiten. In Kapitel 4 werden die einzelnen inhaltlichen Bestandteile der wissenschaftlichen Arbeit dargestellt, worauf in Kapitel 5 wichtige Hinweise zur korrekten Zitierweise gegeben werden. Kapitel 6 beschließt die Arbeit mit einer Zusammenfassung der Ergebnisse und einem Ausblick auf die weitere Entwicklung des eigentlichen Themas.

2 Aufbau und Inhalt der wissenschaftlichen Arbeit

Im vorangegangenen Kapitel hatten wir bereits die Gliederung einer wissenschaftlichen Arbeit im Fachgebiet der Informatik kurz vorgestellt und erläutert, welche inhaltlichen Punkte in der „Einleitung“ behandelt werden sollten. Die folgenden Abschnitte skizzieren inhaltlich die übrigen der bereits genannten Gliederungspunkte.

2.1 Verwandte Arbeiten und wissenschaftlicher Hintergrund (Related Work)

Hier sind vor allem zwei inhaltliche Punkte zu berücksichtigen:

- **Notwendige Vorarbeiten und Grundlagen, die zum Verständnis der Arbeit notwendig sind**

Keine bzw. kaum eine Forschungsarbeit beginnt als „tabula rasa“, d.h. meist bauen wir auf vorhandenen Grundlagen bzw. Vorarbeiten auf. Die zum Verständnis der eigenen Forschungsarbeit notwendigen Grundlagen und Voraussetzungen müssen in diesem Kapitel skizziert bzw. zusammengefasst werden. Dabei sollte man vom durchschnittlichen Kenntnisstand eines Informatikers ausgehen, d.h. Allgemeinplätze und allzu Grundlegendes hat hier nichts zu suchen. Genauso soll hier nicht notwendigerweise eine kompletter Wissenschaftszweig in epischer Tiefe ausgebreitet werden, sondern lediglich die zum Verständnis notwendigen Teilbereiche in skizzenhafter Form und mit Angabe von Literaturhinweisen zusammengefasst werden.

- **Alternative Ansätze und Forschungsarbeiten zum Thema**

Besonders wichtig ist es, spezielle Vorarbeiten und alternative Forschungsansätze zum behandelten Thema darzulegen. Dieser Abschnitt ist der von Ihnen durchgeführten Literaturrecherche gewidmet. Welche Arbeiten zum aktuellen Thema gibt es? Wie sind andere Wissenschaftler an das Thema herangegangen? Hatten Sie Erfolg?

Wichtig ist, dass Sie jede der vorgestellten Arbeiten

- korrekt zitieren (Bibliografie),
- kurz die wichtigsten Ergebnisse bzw. Strategien skizzieren und
- diese (kurz und knapp) in Zusammenhang mit ihrer eigenen Arbeit stellen.

Wie unterscheidet sich der eigene Ansatz von den vorgestellten Arbeiten? Warum ist der eigene Ansatz eventuell erfolgsversprechender?

2.2 Eigener (wissenschaftlicher) Beitrag

Hier haben Sie die Freiheit, Ihren eigenen Arbeiten angemessen viel Raum zur Verfügung zu stellen. Achten Sie dabei auf einen logischen Aufbau der Darstellung, d.h. Grundlegendes zuerst.

- Wie sind Sie vorgegangen?
- Wo gibt es Probleme?
- Wie werden diese gelöst?
- Schreiben Sie in verständlicher Weise und drücken Sie sich dabei jeweils möglichst präzise, d.h. unmissverständlich aus (vgl. Kap. 3)
- Verwenden Sie Abbildungen, Tabellen und Beispiele.
- Setzen Sie kein Wissen als implizit vorhanden voraus, sondern sprechen Sie explizit alle Probleme und wichtigen Fakten an.

Bedenken Sie dabei stets, dass ein Leser nicht dasselbe Wissen besitzen kann wie Sie und das Sie ihm deshalb ihre Ergebnisse erklären müssen.

2.3 Evaluation des (wissenschaftlichen) Beitrags

Natur- bzw. ingenieurwissenschaftliche Forschung erzielt oft quantitative Ergebnisse, deren Qualität objektiv beurteilt werden muss. Dies erfolgt üblicherweise mit Hilfe einer speziellen Evaluation, d.h. die Qualität der erzielten Ergebnisse muss mit den Ergebnissen anderer Arbeiten auf objektive Weise verglichen werden können.

- Oft existieren zu diesem Zweck Benchmarks, die aber auch selbst, angepasst an die eigene, spezielle Problemstellung zusammengestellt werden können.
- Wird ein Benchmark bzw. ein Evaluationsverfahren selbst erstellt, sollten die Autoren diesen öffentlich zur Verfügung stellen, damit die erzielten Ergebnisse nachvollziehbar werden.

Merke: Was man nicht nachvollziehen kann, wird angezweifelt.

- Wird ein existierender Benchmark verwendet, muss dieser korrekt zitiert werden und eigene Ergebnisse mit bereits bekannten Ergebnissen in Relation gestellt werden.

2.4 Diskussion der Evaluationsergebnisse

Wurde eine Evaluation der erzielten Ergebnisse durchgeführt, müssen diese diskutiert werden. Dabei sollten (falls jeweils zutreffend) folgende Fragen beantwortet werden:

- Warum ist das eigene Ergebnis besser/schlechter als das zum Vergleich herangezogene?
- Sind die erzielten Ergebnisse objektiv oder gibt es Gründe, diese eventuell in Zweifel zu ziehen?
- Welche Vorbedingungen müssten verändert werden, um bessere Ergebnisse bzw. eine objektivere Evaluation zu erzielen?
- Falls die Evaluation für unseren speziellen Fall nicht aussagekräftig (genug) ist, wie sollte diese verändert werden?

2.5 Zusammenfassung und Ausblick

In diesem Abschnitt sollten die erzielten Ergebnisse noch einmal kurz zusammengefasst werden und ein Ausblick auf weiterführende Forschungs- und Entwicklungsarbeiten gegeben werden (vgl. Kap. 6).

3 Nützliches und Wissenswertes zum Erstellen einer Seminararbeit

3.1 Allgemeine Hinweise

Ganz allgemein handelt es sich bei der Seminararbeit bereits um eine **wissenschaftliche Arbeit**. Begehen Sie nicht den Fehler und sehen Sie diese als wortwörtliche Wiedergabe Ihres Seminarvortrags an, sondern beachten Sie stets die folgenden Punkte:

- Die sprachliche Darstellung sollte dem Rahmen angepasst sein und stets auf einer sachlichen Argumentationsebene rangieren.
- Der Schreibstil sollte unpersönlich gehalten werden. Vermeiden Sie Sätze, die die Wörter „wir“, „uns“, „Sie“ usw. enthalten.¹
- Vermeiden Sie (Bandwurm-)Sätze, die sich über mehr als 2–3 Zeilen ziehen. Sie erhöhen damit signifikant die Lesbarkeit Ihrer Arbeit.
- Vermeiden Sie sprachliche Komplexität, d.h. beschränken Sie die Anzahl der notwendigen Nebensätze auf ein Mindestmaß.
- Drücken Sie sich einfach und präzise aus. Vermeiden Sie eine „geschraubte“ Ausdrucksweise.
- Achten Sie auf temporale Konsistenz in der Verwendung von Präsens oder Präteritum bei Verben.
- Vermeiden Sie Worthülsen und unnötige Redewendungen ohne signifikanten Inhalt.
- Legen Sie Ihren Standpunkt stets mit der angemessenen Objektivität dar, auch wenn es um die Bewertung von Vor- oder Nachteilen des jeweiligen Themenganges geht.
- Vermeiden Sie unnötige Anglizismen (z.B. „connecten, downloaden, backupen“). Existiert in diesem Zusammenhang bereits eine deutsche Redewendung, dann benutzen Sie diese (z.B. „öffentlicher Schlüssel“ statt „public key“). Verwenden Sie englische Begriffe, so passen Sie diese entsprechend den deutschen Rechtschreibregeln bez. Flexion, Silbentrennung, Getrennt- und Zusammenschreibung an.
- Wenn Sie Begriffe einführen bzw. verwenden, deren Bedeutung nicht unmittelbar auch einem Nichtfachmann geläufig ist, sollten diese stets erläutert werden. Die Fachbegriffe sollten unmittelbar im Text oder im Text **und** dem Glossar erklärt werden. In natur- und ingenieurwissenschaftlichen Arbeiten ist es eher unüblich, Erklärungen in Fußnoten zu platzieren.
- Achten Sie auf einen **logischen Aufbau** Ihrer Arbeit, sowie ihrer einzelnen Unterkapitel und Gliederungspunkte.

¹Eine Ausnahme von dieser Regel stellen persönliche Kommentare, Bewertungen oder Urteile von Sachverhalten dar. Hier sollte klar werden, dass es sich um Ihre eigene Leistung handelt.

- **Dringende Empfehlung:** Lassen Sie Ihre Ausarbeitung am besten von einem „Nichtfachmann“/einer „Nichtfachfrau“, also nicht von einem Informatiker/einer Informatikerin Korrekturlesen. Auf diese Weise werden Schwachstellen in Ihrer Argumentation und andere logische Mängel am deutlichsten.
- Verwenden Sie konsequent die **neue deutsche Rechtschreibung**. Lassen Sie Ihren Text von einem Rechtschreibprogramm prüfen. Allerdings kann dieses nur die korrekte Schreibweise von Einzelwörtern und nicht die korrekte Verwendung von Einzelwörtern im Satzzusammenhang überprüfen, insbesondere wenn es um grammatikalische Fehler oder Kommasetzung geht. Verlassen Sie sich daher nicht alleine auf das Programm.
- In deutschen Texten werden anstelle der englischen "Quotes" die Anführungsstriche in Form von „Gänsefüßchen“ geschrieben. In dem Textsatzsystem \LaTeX stehen hierfür z.B. die Befehle `\glqq` und `\grqq` zur Verfügung.
Achtung: Dazu muss das Paket `ngerman` in den Header eingebunden werden.

3.2 Weitere Formatierungshinweise

3.2.1 Abbildungen

Abbildungen sind für das Erläutern und Verdeutlichen von komplexen Strukturen und Abläufen, wie sie üblicherweise in wissenschaftlichen Arbeiten beschrieben werden, unverzichtbar. Die Abbildungen in Ihrer Arbeit sind stets zu **nummerieren** und wie folgt zu setzen (siehe Abb. 1).

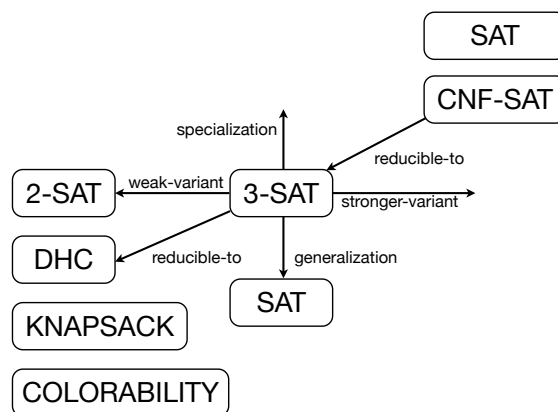


Abbildung 1: Das Umfeld des 3-SAT Problems

Jede Abbildung soll neben einer Abbildungsnummer eine Bildunterschrift („Das Umfeld des 3-SAT Problems“) besitzen, die die Grafik näher erläutert. Weiterhin ist darauf zu achten, dass zu jeder Abbildung eine Bezugnahme im Text aufgenommen wird, d.h. an geeigneter Stelle sollte ein Verweis der Form (siehe Abb. ...) erfolgen.

Beachten Sie bitte, dass es sich bei der vorliegenden Grafik um eine .eps-Datei handelt (Encapsulated Postscript), die mit Hilfe des \LaTeX -Pakets `epsfig` eingebunden wurde.

Die Abbildung kann natürlich auch mit Hilfe anderer Pakete eingebunden werden. Achten Sie dabei aber stets auf eine für den Druck geeignete Bildqualität!

Achtung:

Eine Abbildung sollte nicht einfach per „copy and paste“ aus dem WWW in Ihre Ausarbeitung übernommen werden. Einerseits leidet im Allgemeinen bei einer solchen Vorgehensweise die Qualität bei einer entsprechenden Vergrößerung für die Druckaufbereitung stark darunter (siehe Abb. 2), andererseits muss sichergestellt werden, dass der Urheber der Originalgrafik damit einverstanden ist, dass die Grafik in die Seminararbeit aufgenommen wird. Der sicherste Weg ist deshalb, die Grafik mit einem geeigneten Grafikprogramm *selbst* zu erzeugen und anschließend in den Text einzubinden.

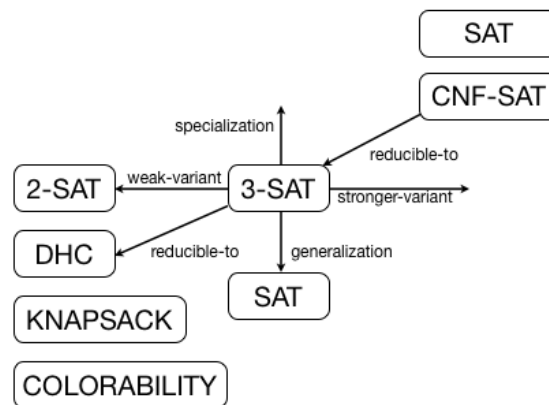


Abbildung 2: So sollte eine Abbildung nicht aussehen (JPG-Grafik, „copy and paste“)

Verwenden Sie eine Abbildung, die in dieser Form bzw. in einer sehr ähnlichen Form bereits veröffentlicht wurde, müssen Sie dies durch eine bibliografische Referenz deutlich machen, d.h. Sie müssen wie bei einem Zitat die Fundstelle im Text als Literaturangabe aufnehmen.

3.2.2 Tabellen

Für Tabellen gilt dasselbe wie für Abbildungen. Setzen Sie Tabellen nie in den Fließtext, sondern in die entsprechende \LaTeX -Tabellenumgebung und versehen Sie diese ebenfalls mit einer *Tabellennummer* und einer *Tabellenüberschrift* (siehe Tabelle 1).

Grundsätzlich sollte man bei dem Setzen einer Tabelle sparsam mit grafischen Elementen – wie Hilfslinien oder bunte Tabellenunterlegungen – umgehen, da diese die Lesbarkeit sehr stark beeinträchtigen können.

3.2.3 Listings, Quellcode und Pseudocode

Sollten Sie Quellcode eines Programmes mit in Ihre Arbeit aufnehmen oder einen Algorithmus mit Hilfe von Pseudocode veranschaulichen, sind diese im Sinne einer Abbildung

Tabelle 1: Die Geschichte des World Wide Web

1945	Vannevar Bush beschreibt MEMEX: das erste Hypertextsystem.
1965	Ted Nelson prägt als erster das Wort Hypertext auf der ACM-Jahreskonferenz.
1968	Doug Engelbart entwickelt ein Hypertext-basiertes Prototypensystem NLS und erfindet zu diesem Zweck die Maus als Eingabegerät.
1980	Tim Berners Lee schreibt ein erste Notizbuch-Programm mit Hypertext-links.
1989	Tim Berners Lee verfasst ein erstes Memorandum zu seinem Hypertext-Dokumentenverwaltungssystem am Kernforschungszentrum CERN.
1990	Zusammen mit Robert Cailliau entwickelt Tim Berners Lee den ersten WWW-Server und WWW-Browser: die Geburtsstunde des WorldWide-Web.

(siehe Abb. 3) zu behandeln. Ebenso wie eine Abbildung sind Listings, Quellcode oder Pseudocode jeweils mit einer laufenden Nummer und einer Bildunterschrift zu versehen.

```

for( i = 0; i < 10; i++ )
{
    for( j = 0; j < 10; j++ )
    {
        // calculate  $a_{ij}$ 
        a[i][j] = b[j][i];
    }
}

```

Abbildung 3: Einbinden von Quellcode in die Arbeit

Sie können Listings z.B. mit Hilfe des \LaTeX -Pakets **listings** einbinden. Insbesondere, wenn längere Quellcode-Passagen in den Text eingebunden werden sollen, empfiehlt sich diese Vorgehensweise, da das Paket automatisch Seitenumbrüche korrekt formatiert und auch die Verwendung von internen Zeilennummern ermöglicht.

3.2.4 Mathematische Formeln

\LaTeX bietet sich insbesondere als Textverarbeitungssystem an, wenn es um das korrekte Formatieren mathematischer Ausdrücke und Formeln geht. Versehen Sie alle Formeln, die Sie in Ihrer Arbeit verwenden – insbesondere diejenigen, auf die Sie später noch Bezug nehmen – mit einer entsprechenden Nummerierung.

$$\frac{\sum_{n>0} z^n}{\prod_{1\leq k\leq n}(1-q^k)} \tag{1}$$

3.2.5 L^AT_EX und andere Textverarbeitungssysteme

Wenn Sie das Textsatzsystem L^AT_EX für die Erstellung der Ausarbeitung verwenden, dann benutzen Sie das vorliegende L^AT_EX-Musterdokument als Layoutvorlage für Ihre Arbeit. Sollten Sie ein anderes Textverarbeitungsprogramm als L^AT_EX verwenden, dann halten Sie sich bitte an die in diesem Beispieldokument verwendeten Bemaßungen und Formatierungen. Sie können sich z.B. dieses Beispieldokument ausdrucken und die entsprechenden Maßangaben, wie

- linker, rechter Rand
- Abstand oben, unten
- etc.

abmessen und in Ihrem eigenen Textverarbeitungsprogramm verwenden.

Denken Sie bitte daran, dass neben einer gedruckten Version Ihrer Seminausarbeitung auch eine **pdf**-Datei abzugeben ist. Diese können Sie zusammen mit Ihrer Präsentation (pdf-Datei + evtl. ppt-Datei) via E-Mail zusenden.

4 Inhaltliche Bestandteile der Seminararbeit

4.1 Gliederungspunkte

Die Seminaarausarbeitung sollte **ohne** die Standardseiten wie

- Titelseite
- Kurzzusammenfassung
- Inhaltsverzeichnis
- Glossar
- Abkürzungsverzeichnis
- Literaturverzeichnis
- Index (Sachindex)

tatsächlich **20 Seiten** umfassen! Für Diplom-/Masterarbeiten gelten entsprechend 60–80 Seiten.

4.2 Inhalt der Gliederungspunkte

Unterteilen Sie den eigentlichen Text Ihrer Arbeit in logische, inhaltlich aufeinander aufbauende Gliederungspunkte. Stellen Sie sicher, dass der Inhalt jedes Gliederungspunktes auch mit dessen einleitenden Sätzen übereinstimmt. Achten Sie bei der Erstellung der einzelnen Gliederungspunkte auf den sprichwörtlichen „roten Faden“, der sich durch die Arbeit ziehen sollte. Reihen Sie nicht nur einzelne Fakten hintereinander, sondern bringen Sie diese in einen logischen Zusammenhang. Dies gilt auf allen Gliederungsebenen, d.h. sowohl für den Gesamtaufbau der Arbeit wie auch für die einzelnen Unterkapitel.

Hüten Sie sich vor Plagiaten! Dem Vorwurf des Plagiats setzt man sich auch dann aus, wenn man einer anderen Arbeit zu dicht folgt und seine eigene Arbeit zu sehr an eine andere Arbeit anlehnt. Die Suchmaschine Google und das WWW bieten einen reichen Schatz an studentischen Arbeiten zu den verschiedensten Themen. Aber bedenken Sie:

- Ihr Dozent ist ebenfalls in der Lage, einen Browser zu bedienen.
- Was Sie im WWW finden, kann auch Ihr Dozent finden.
- Ihr Dozent hat in der Regel einen besseren Überblick über bereits bestehende Arbeiten zum Thema als Sie.
- Ihr Dozent hat bereits vor Ihnen eine WWW-Recherche zum Thema durchgeführt.

Grundsätzlich können Sie fremde Quellen immer zu Rate ziehen und diese korrekt zitieren. Eine komplette Arbeit einfach abzuschreiben bringt allerdings auch Ihnen persönlich weder einen Erkenntnisgewinn noch Erfahrungen im Erstellen einer wissenschaftlichen Arbeit.

4.3 Umfang der Gliederungspunkte

Die einzelnen Unterkapitel sollten entweder selbsterklärend sein bzw. sollte sich deren Zusammenhang aus den bereits vorangegangenen Kapiteln erschließen. Ist dies nicht der Fall, müssen Sie eventuell die einzelnen Kapitel umorganisieren bzw. zusätzliche Erklärungen einfügen.

4.4 Logischer Zusammenhang

Generell gilt auch hier: Lesen Sie Ihre Arbeit am Ende komplett in einem Stück durch. Wenn Sie glauben, Ihre Arbeit sei logisch konsistent und vollständig, dann lassen Sie diese von einer unbeteiligten Person (am besten einem Nichtfachmann/einer Nichtfachfrau) noch einmal durchlesen. Diese wird Sie auf eventuell bestehende logische Unzulänglichkeiten hinweisen.

5 Das Literaturverzeichnis und die korrekte Zitierweise

5.1 Was wird zitiert?

Jede Behauptung tatsächlicher Art, d.h. stets wenn Sie konkrete Werte oder Aussagen wiedergeben, gilt solange als Behauptung, bis Sie diese auch belegen können. Ein Beleg besteht entweder in einer korrekten Herleitung, wie z.B. einem mathematischen Beweis, oder aber in einer Angabe der Fundstelle (Literatur oder WWW), aus der die besagte Behauptung gewonnen wurde (= bibliografische Referenz). Erwähnen Sie in Ihrer Arbeit Internetdienste, Programme, Sprachen, Internetstandards oder bestimmte Werkzeuge, dann belegen Sie diese beim ersten Vorkommen in Ihrem Text mit einer bibliografischen Referenz. Im Allgemeinen wird immer stets an der Stelle zitiert, die es zu belegen gilt [6].

Achten Sie darauf, dass der Zitierhinweis stets Bestandteil des Satzes ist, d.h. der Punkt kommt erst dahinter. Die Quellenangabe sollte auf das am Ende der Ausarbeitung vorhandene Literaturverzeichnis verweisen. Gegebenenfalls kann man für diesen Zweck *zusätzlich* Fußnoten verwenden.

5.2 Bibtex das Zitiersystem von L^AT_EX

BibTeX ist ein Programm zur Erstellung von Literaturangaben und -verzeichnissen in TeX- oder L^AT_EX-Dokumenten.

Um ein Literaturverzeichnis zu erstellen, werden aus einem L^AT_EX-Dokument alle Zitatverweise herausgesucht und über eine Literatur-Datenbank dem entsprechenden Werk zugeordnet. Bei der Literaturdatenbank handelt es sich um eine Textdatei (*.bib-Datei), in der alle bekannten Angaben über ein Werk (Buch, Wissenschaftliche Publikation, Webseite etc.) in einer bestimmten Syntax notiert werden.

Die zitierten Werke werden sortiert und durch eine entsprechende Anweisung im LaTeX-Dokument aufgelistet. Die Formatierung dieser Literaturliste ist variabel. Der im Dokument eingestellte BibTeX-Stil (engl. *style*) bestimmt, welche Angaben in welcher Formatierung dargestellt werden.

BibTeX ist in der Lage, auch mit sehr großen Literaturbeständen sowie mit sehr großen Dokumenten problemlos zusammenzuarbeiten. BibTeX hat sich daher im wissenschaftlichen Umfeld schon seit Jahren als offenes Standardformat für Literaturangaben etabliert. Das folgende Beispiel (entnommen aus einer BibTeX-Datei)

```
@article{lin1973,
  author = {Shen Lin and Brian W. Kernighan},
  title = {An Effective Algorithm for the Travelling-Salesman Problem},
  journal = {Operations Research},
  volume = {21},
  year = {1973},
  pages = {498--516},
}
```

wird durch den BibTeX-Stil *alphadin* in diese Ausgabe in der Literaturliste (engl. *bibliography*) überführt:

[LK73] Lin, Shen; Kernighan, Brian W.: An Effective Algorithm for the Travelling-Salesman Problem. In: Operations Research 21 (1973), S. 498–516

Der Befehl `\cite{lin1973}` innerhalb eines LaTeX-Dokuments wird durch die in der BibTeX-Datei mit dieser ID angegebene Referenz, im Beispiel '[LK73]', ersetzt.

Neben dem BibTeX-Stil *alphadin*, gibt es den Stil *plain*, bei dem der Schlüssel lediglich aus Ziffern besteht, z.B. [12]. Daneben gibt es verschiedene Varianten dieser Stile, die sich hauptsächlich in der Darstellung der Literaturliste unterscheiden und oft spezifisch für verschiedene wissenschaftliche Verlage, Konferenzen und Zeitschriften sind (vgl. [1]). Wer nicht zitiert hat, aber eine Quelle nennen will, tut dies durch `\nocite{lin1973}`.

5.3 Zitieren und das Internet

Auch wichtige Quellen, die nur im Internet publiziert wurden, müssen zitiert werden. Unterscheiden Sie bitte dabei, ob es sich lediglich um eine Web-Präsenz, wie z.B. ein Web-Portal oder eine Übersichtsseite handelt, deren Inhalt sich mit der Zeit verändern kann. Dies kann z.B. der Fall sein, wenn Sie die Suchmaschine Google² erwähnen und dazu den URL als Referenz angeben. In diesem Fall empfiehlt es sich, die URL als Fußnote anzugeben.

Andererseits können Sie auch auf ein Web-Dokument verweisen, dessen Inhalt für sich selbst und der sich wahrscheinlich nicht so schnell wieder verändern wird. Dann müssen Sie den URL des Dokuments in die Bibliographie aufnehmen. Zur korrekten Formatierung und Silbentrennung von URLs verwenden Sie das L^AT_EX-Paket `URL`; dieses sorgt für ein korrektes Umbrechen am Zeilenende. Beachten Sie hier zu jedem URL auch das Datum anzugeben, an dem sie den URL zuletzt erfolgreich zugegriffen haben, da Sie sich auf eine ganz bestimmte Version dieses Dokuments beziehen, dessen Inhalt sich eventuell mit der Zeit verändern könnte.

5.4 Zitieren und die Wikipedia

Das Zitieren der Online-Enzyklopädie Wikipedia³ wird aktuell noch kontrovers diskutiert. Schuld daran ist die mangelnde Persistenz der Inhalte, d.h. im Prinzip kann jeder Benutzer den Inhalt eines Wikipedia-Artikels willkürlich verändern, so dass dieser nicht als gesicherte Referenz herangezogen werden kann. Auch wenn in der Wikipedia mittlerweile ein hohes Maß an Selbstkontrolle vorherrscht, sollte bei der wissenschaftlichen Bibliografie Wert auf das Prinzip der Nachvollziehbarkeit gelegt werden. Verwenden Sie daher bitte möglichst stets gesicherte, d.h. regulär publizierte Quellenangaben. Dies ist insbesondere dann ratsam, wenn Sie Grundlagenarbeiten und Nachschlagewerke zitieren.

²<http://www.google.com/>

³<http://www.wikipedia.org/>

6 Zusammenfassung und Ausblick

Das Kapitel „Zusammenfassung und Ausblick“ soll die gewonnenen Ergebnisse und Erkenntnisse Ihrer Arbeit knapp zusammenzufassen. Stellen Sie dabei eindeutig klar, was wichtig ist und was nicht. Dazu zählt auch, dass Sie einen Ausblick auf die Weiterentwicklung innerhalb des von Ihnen bearbeiteten Themengebiets geben können.

- Was haben wir erreicht?
- Was sind die nächsten Schritte?
- Wie können die gewonnenen Ergebnisse angewendet werden?
- Warum sind die gewonnenen Ergebnisse bedeutend?

E Glossar

Hier sollten die wichtigsten Schlüsselbegriffe, die in Ihrer Ausarbeitung thematisiert werden, knapp – also wie in einem Lexikon – gesammelt an einer Stelle zentral erläutert werden. Das Glossar sollte im Umfang nicht mehr als eine Seite Ihrer Ausarbeitung einnehmen. Sie müssen also allgemein bekannte Begriffe, wie z.B. „Internet“ nicht unbedingt hier mit aufnehmen.

Die Erläuterung eines Begriffes im Glossar schließt nicht aus, dass Sie diesen Begriff bei seiner Einführung im Text nicht auch bereits erklärt haben.

Browser: Ein spezielles Programm, mit dem man über das WWW Zugang zu WWW-Servern erlangen und von diesem angeforderte Dokumente anzeigen kann.

Client: Bezeichnet ein Programm, dass einen Server kontaktiert und von diesem Informationen anfordert. Der im WWW eingesetzte Browser ist in diesem Sinne ein Client. Aber es gibt auch andere Clients im WWW, die WWW-Server kontaktieren und Informationen von diesen herunterladen, wie z.B. Suchmaschinen oder Agenten.

HTML: Hypertext Markup Language; das einheitliche Dokumentenformat für Hypermedia-Dokumente im WWW. Dokumente, die im WWW übertragen und vom Browser dargestellt werden sollen, sind in HTML kodiert.

HTTP: Hypertext Transfer Protocol; das Protokoll, das die Kommunikation von Browsern und WWW-Servern im WWW regelt. Fordert ein Browser ein Dokument vom WWW-Server an oder beantwortet der WWW-Server eine Anfrage, muss diese Anfrage den Konventionen des HTTP-Protokolls gehorchen.

Netzanwendung: Ein Anwendungsprogramm, dessen Ablauf den Zugriff auf Ressourcen einschließt, die nicht lokal auf dem ausführenden Rechner liegen, sondern auf einem entfernten Rechner über das Netzwerk zugegriffen werden.

Server: Bezeichnet einen Prozess, der von Clients kontaktiert wird, um diesen Informationen zurück zu liefern. Oft wird auch der Rechner, auf dem ein Server-Prozess abläuft, als Server bezeichnet.

F Abkürzungen und Akronyme

Hier sollten **alle** von Ihnen im Text verwendeten Abkürzungen in einem Verzeichnis zusammengestellt werden. Falls im Text keine Abkürzungen benutzt werden, brauchen Sie natürlich auch kein Abkürzungsverzeichnis zu erstellen.

4CIF	4 fach Common Intermediate Format
AAC	Advanced Audio Coding
AAL	ATM Adaption Layer
ABR	Available Bit Rate
AC	Audio Code
ACK	Acknowledgement
ADM	Add Drop Multiplexer
ADSL	Asymmetric Digital Subscriber Line
AH	Authentication Header
AIFF	Audio Interchange File Format
AM	Amplituden-Modulation
ANSI	American National Standards Institute
API	Application Programming Interface
ARP	Address Resolution Protocol
W3C	World Wide Web Community
WWW	World Wide Web

Literatur

- [1] Bibtex. Bibtex and bibliography styles, 2010. [Online; Stand 22. Juni 2010].
- [2] M. Hausenblas. Exploiting Linked Data For Building Web Applications. *IEEE Internet Computing*, 2(5):34–42, 2009.
- [3] Wayne Hodgins and Erik Duval. Draft standard for learning technology - Learning Object Metadata - ISO/IEC 11404. Technical Report IEEE P1484.12.2/D1, ISO/IEC, 2002.
- [4] Wessel Kraaij and Wilfried Post. Task based evaluation of exploratory search systems. In *Proc. of First Workshop on Evaluating Exploratory Search Systems: SIGIR 2006*. ACM Press, August 2006.
- [5] Shen Lin and Brian W. Kernighan. An effective algorithm for the travelling-salesman problem. *Operations Research*, 21:498–516, 1973.
- [6] Gary Marchionini. Exploratory search: from finding to understanding. *Commun. ACM*, 49(4):41–46, 2006.
- [7] Jörg Waitelonis and Harald Sack. Augmenting video search with linked open data. In *Proc. of Int. Conf. on Semantic Systems 2009*, September 2009.
- [8] Wikipedia. Bibtex — wikipedia, die freie enzyklopädie, 2010. [Online; Stand 22. Juni 2010].
- [9] Max L. Wilson and m.c. schraefel. Bridging the gap: Using ir models for evaluating exploratory search interfaces. In *SIGCHI 2007 Workshop on Exploratory Search and HCI*, 2007.

Selbstständigkeitserklärung

Hiermit erkläre ich, Henrik Jürges, an Eides statt, dass ich die vorliegende Arbeit „Erweiterung des GERBIL-Frameworks zur Evaluation von Named-Entity-Linking Verfahren“ selbstständig angefertigt, nicht anderweitig zu Prüfungszwecken vorgelegt und keine anderen als die angegebenen Hilfsmittel verwendet habe. Sämtliche wissentlich verwendete Textausschnitte, Zitate oder Inhalte anderer Verfasser wurden ausdrücklich als solche gekennzeichnet und im Literaturverzeichnis aufgeführt.

Potsdam, den 29. April 2016

Henrik Jürges