Semantic Support in Multilingual Text Retrieval

Dissertation

zur Erlangung des akademischen Grades

Doktoringenieur (Dr.-Ing.)

angenommen durch die Fakultät für Informatik
der Otto-von-Guericke-Universität Magdeburg

von M.A., Ernesto William De Luca,
geboren am 08. April 1976 in Turin, Italien

Gutachter: Prof. Dr. Andreas Nürnberger
Prof. Dr. Dietmar Rösner
Prof. Dr. Carla Marello

Magdeburg, den 16. Juni 2008

Berichte aus der Informatik

Ernesto William De Luca

Semantic Support in Multilingual Text Retrieval

Shaker Verlag
Aachen 2008

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de.

Zugl.: Magdeburg, Univ., Diss., 2008

Copyright Shaker Verlag 2008
All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the publishers.

Printed in Germany.

ISBN 978-3-8322-7489-4
ISSN 0945-0807

Shaker Verlag GmbH • P.O. BOX 101818 • D-52018 Aachen
Phone: 0049/2407/9596-0 • Telefax: 0049/2407/9596-9
Internet: www.shaker.de • e-mail: info@shaker.de

The future belongs to those who discover how to help users better understand the information that is provided. And the only way I know to do that is provide contexts to make the information meaningful.

(George A. Miller, 2001)

I want to talk about learning. But not the lifeless, sterile, futile, quickly forgotten stuff that is crammed in to the mind of the poor helpless individual tied into his seat by ironclad bonds of conformity!

I am talking about LEARNING - the insatiable curiosity that drives the adolescent boy to absorb everything he can see or hear or read about gasoline engines in order to improve the efficiency and speed of his 'cruiser'.

I am talking about the student who says, "I am discovering, drawing in from the outside, and making that which is drawn in a real part of me."

I am talking about any learning in which the experience of the learner progresses along this line: "No, no, that's not what I want"; "Wait! This is closer to what I am interested in, what I need"; "Ah, here it is! Now I am grasping and comprehending what I need and what I want to know!"

(Carl Rogers 1983: 18-19)

Abstract

Current search engines (e.g. Google, Yahoo) are used for retrieving relevant documents from the huge amount of data available and have become an essential tool for the majority of Web users. Standard search engines do not consider semantic information that can help in recognizing the relevance of a document with respect to the meaning of a query. Furthermore, most of them only support one language. However, users speaking several languages would, in principle, be able to switch to another language if the results of their original query are unsatisfactory.

This thesis provides a new *semantic-based search approach* to Web search. Unlike the "standard" search process, users are *redirected* to semantic concepts that could describe their query. This semantic-based *redirection* serves to support the search process by providing possible meanings of the query words and assigning the retrieved documents to the respective meanings.

In this thesis, *lexical resources* are used for providing the possible meanings of query words and the related translations in the languages the users are able to speak. The contexts of the query words are recognized using *Word Sense Disambiguation* approaches. Semantically similar documents are grouped together using *Document Categorization* techniques. Moreover, *Semantic* and *Multilingual Text Retrieval* methods are developed to support humans in filtering and retrieving relevant documents. Additionally, in order to assist users in the semantic-based multilingual search process, pre-processing approaches such as named-entity recognition, spell checking and stemming methods are applied to queries and documents. All the proposed methods are compared and evaluated. Several tools that implement combinations of these methods have been developed and discussed.

Kurzfassung

Herkömmliche Suchmaschinen (z.B. Google, Yahoo) dienen dem Auffinden relevanter Dokumente in der großen Menge von verfügbaren Daten und sind daher ein unentbehrliches Werkzeug für die Mehrheit der Internetnutzer geworden. Suchmaschinen bieten jedoch zur Zeit eine nur unzureichende Unterstützung, wenn es um die Behandlung von mehrdeutigen Anfragen geht und liefern für diese Fälle häufig eine Vielzahl unzutreffender Dokumente. Weiterhin unterstützen Suchmaschinen gewöhnlich nur eine einzelne Sprache bei der Bearbeitung einer Suchanfrage. Nutzer, die mehrere Sprachen sprechen, hätten jedoch im Prinzip die Möglichkeit, bei der Suche auf andere Sprachen auszuweichen, wenn die Ergebnisse in der ursprünglichen Sprache nicht in dem gewünschtem Umfang oder der benötigten Qualität geliefert werden.

Diese Dissertation stellt eine neue Art der Suche vor, die im Unterschied zu den momentan existierenden Suchmaschinen semantische Informationen berücksichtigt. Die Suche unterscheidet sich von dem Standardsuchprozess darin, dass Nutzer zu semantischen Konzepten weitergeleitet werden, um die Bedeutungen ihrer Anfrage besser zu erkennen. Diese semantische Weiterleitung unterstützt den Suchprozess, indem unterschiedliche Bedeutungen der Suchbegriffe zur Auswahl angeboten und die gefundenen Dokumente den entsprechenden Bedeutungen zugeordnet werden.

Ziel der Arbeit ist die Entwicklung eines Verfahrens, welches Nutzer beim Suchprozess unterstützt und dabei die jeweiligen von ihnen beherrschten Sprachen und die unterschiedlichen Bedeutungen des Suchbegriffes anhand von *lexikalischen Ressourcen* berücksichtigt. Um die Kontexte der in der Anfrage enthaltenen Wörter zu erkennen, werden *Word Sense Disambiguation*-Methoden eingesetzt. Verschiedene überwachte und unüberwachte Lernverfahren werden zur Gruppierung und Kategorisierung semantisch ähnlicher Dokumente eingesetzt, die interaktiv nach Bedeutung gefiltert werden können. Um Nutzer bei der Filterung und Suche nach relevanten Dokumenten anhand von Semantik und mehrsprachigem Hintergrundwissen zu unterstützen, werden *Semantic* und *Multilingual Text Retrieval*-Methoden entwickelt. Hierzu werden ergänzend Vorverarbeitungsschritte wie mehrsprachige Rechtschreibkorrektur, mehrsprachiges Stemming und Named-Entity-Recognition implementiert. Alle vorgestellten Verfahren werden verglichen und evaluiert. Abschließend werden die im Rahmen dieser Arbeit entwickelten Werkzeuge ausführlich diskutiert.

Contents

Abstract v

Kurzfassung vii

1 Introduction 1

I Fundamentals 7

2 Language Communication 11
2.1 Interpersonal Communication . . . 11
2.2 Human-Computer Interaction . . . 13
2.3 Word Meaning . . . 15
2.3.1 Semantic Prototypes and Primitives . . . 16
2.3.2 Word Senses and Concepts . . . 18
2.3.3 Lexical Gaps and Compound Words . . . 20
2.4 Conclusions . . . 21

3 Language Resources 23
3.1 Written Language Corpora . . . 23
3.2 Lexical Resources . . . 24
3.2.1 WordNet . . . 27
3.2.2 EuroWordNet (EWN) . . . 27
3.2.3 MultiWordNet . . . 30
3.2.4 WordNet Criticism . . . 32
3.3 Conclusions . . . 34

4 Text Retrieval 37
4.1 Short History . . . 39
4.2 Query and Document Preprocessing . . . 41
4.2.1 Spelling Correction . . . 42
4.2.2 Stemming . . . 43
4.2.3 Named-Entity Recognition . . . 44
4.2.4 Query Expansion . . . 45
4.2.5 Text Indexing . . . 46
4.3 Retrieval Models . . . 46
4.3.1 Vector Space Model . . . 47

4.3.2 Document Encoding . . . 48
4.4 Performance Measures . . . 49
4.5 Classification and Clustering . . . 50
4.6 Conclusions . . . 54

5 Word Sense Disambiguation 57
5.1 Short History . . . 57
5.2 Word Sense Granularity . . . 58
5.3 Word Sense Clustering . . . 58
5.4 Word Sense Disambiguation Approaches . . . 59
5.4.1 WSD Corpus-based Approaches . . . 60
5.4.2 WSD Knowledge-based Approaches . . . 61
5.4.3 Evaluating WSD . . . 65
5.5 Word Sense Disambiguation and Information Retrieval . . . 67
5.6 Conclusions . . . 69

II Multilingual Text Retrieval 71

6 Motivation and Challenges 75
6.1 Needs for Multilingual Text Retrieval . . . 76
6.2 Query and Document Translation . . . 78
6.3 Interactive Multilingual Information Retrieval . . . 79
6.4 Conclusions . . . 81

7 Multilingual Information Retrieval Systems 83
7.1 Machine Translation . . . 84
7.2 Corpus-based Approach . . . 86
7.3 Knowledge-based Approach . . . 88
7.3.1 Concept-based Information Retrieval . . . 88
7.3.2 The Dictionary-based Retrieval . . . 91
7.3.3 Thesaurus- and Controlled vocabulary-based Approach . . . 93
7.3.4 Combination of Dictionary-based and Corpus-based Approaches . . . 95
7.4 Interactive Multilingual Information Retrieval Systems . . . 96
7.5 Discussion . . . 100

III The Multilingual Sense Folder Approach 103

8 Sense Folder Disambiguation 107
8.1 Query and Document Pre-Processing . . . 110
8.1.1 Spelling Correction (Multispell) . . . 110
8.1.2 Stemming (Snowball) . . . 113
8.1.3 Named-Entity Recognition (Stanford NER) . . . 113
8.2 The Search Engine Framework (CARSA) . . . 113
8.3 Semantic Prototype Engine . . . 116

8.4 Sense Folder Classification . . . 118
8.4.1 Indexing the Results Set . . . 119
8.4.2 Disambiguating the Meaning of the Search Terms . . . 119
8.4.3 Multilinguality in Sense Folder Classification . . . 120
8.5 Sense Folder Clustering Methods . . . 121
8.5.1 k-Means Clustering . . . 121
8.5.2 Expectation-Maximization Clustering . . . 122
8.5.3 Density-based Clustering . . . 123
8.6 Merging Word Senses and Merging Methods . . . 124
8.6.1 Merging Word Senses . . . 125
8.6.2 Merging Methods . . . 126
8.7 Conclusions . . . 127

9 Evaluation **129**
9.1 Evaluation of Hand-Tagged Data . . . 129
9.2 BankSearch Data Set Collection Evaluation . . . 132
9.2.1 "rule" and "operation" BankSearch Subsets . . . 133
9.2.2 Linguistic Parameters and Manual Merging Methods . . . 136
9.3 Sense Folder Corpus Evaluation . . . 140
9.3.1 Corpora Analysis . . . 141
9.3.1.1 Senseval and Semcor Data Sets . . . 142
9.3.1.2 IR-SemCor Corpus . . . 143
9.3.1.3 Building the Sense Folder Corpus . . . 143
9.3.2 Baselines . . . 144
9.3.3 Evaluation of the Linguistic Parameters . . . 149
9.3.4 Merging Methods Evaluation . . . 158
9.3.5 Concluding Remarks . . . 163

10 Multilingual User Interfaces **167**
10.1 Multilingual Sense Folder User Interface . . . 167
10.1.1 Interacting with the Multilingual User Interface . . . 168
10.1.2 Search Configuration . . . 170
10.2 The Lexical Restructuring Tool (LexiRes) . . . 172
10.2.1 Related Work . . . 174
10.2.2 User Interface and Interaction . . . 174
10.3 MultiLexExplorer: Combining Multilingual Web Search with Multilingual Lexical Resources . . . 176
10.3.1 Language Learning Software . . . 178
10.3.2 Navigating Multilingual Information . . . 179
10.3.3 Related Work: Tools for Visualization and Navigation of Linguistic Ontologies . . . 180
10.3.4 System Architecture . . . 181
10.4 A Mobile User Interface . . . 185

11 Concluding Remarks **189**

11.1 Open Problems 189
11.2 RDF/OWL EuroWordNet Representation 191

Appendix **193**

A Sense Folders Creation **195**
A.1 Overview 195
A.2 Example for term "network" and its linguistic relations 196

B Merging Evaluation **199**
B.1 Overview 199
B.2 Best Performing Merging Methods and Thresholds 201
B.3 Other Tested Merging Methods and Thresholds 209
B.4 Combination of Best Performing Merging Methods and Thresholds 221

List of Figures

2.1 Encoding-Decoding Process [147]. 11
2.2 Human-Computer Interaction Overview. 14
2.3 Dimensions of choice in a semantic formalism [286]. 17
2.4 A Classification of Translation Equivalents [24]. 21

3.1 Example of the WordNet hierarchy for a given term A. 28
3.2 EuroWordNet Architecture (see [274]). 29
3.3 Example of the MultiWordNet hierarchy for the term "operation" 33

4.1 Text Retrieval System. 38
4.2 Text Retrieval Chapter Overview. 40

5.1 Word Sense Disambiguation Overview 61

7.1 Multilingual Text Retrieval Overview (adapted from [188]). 83

8.1 Overview of the Sense Folder Approach. 108
8.2 Sense Folder Annotation. 110
8.3 Corrections for a misspelled word (MultiSpell). 111
8.4 Suggestions generated with the spell checker (MultiSpell). 112
8.5 Suggestions generated with stemming methods (Snowball). 114
8.6 Suggestions generated with the Stanford named-entity recognizer. 115
8.7 Overview of the CARSA retrieval system [12]. 116
8.8 Example of linguistic relations for a given term A. 117
8.9 Implementation of the k-Means algorithm 122

10.1 Multilingual User Interface - "lingua" example 169
10.2 Multilingual User Interface - "speech" example 171
10.3 Pre- and Post-Processing Configuration 173
10.4 Sense Folder Configuration 173
10.5 Example of the word "bank" - manual merging functions - in the LexiRes Editor. 175
10.6 Example of the word "bank" - SynSet translations - in the LexiRes Editor. 177
10.7 Screenshot of the MultiLexExplorer. 182
10.8 MultiLexExplorer Architecture. 184
10.9 Mobile User Interface: a) Search window, and b) "Sense Folder" result list, c) "Intelligent Bookmarks". 186

List of Tables

3.1 Comparison of Lexical Resources 35

4.1 Confusion matrix 50

8.1 WordNet noun collocation of the term “range” 120
8.2 Rule example for disambiguation 124
8.3 Bank example for disambiguation 125

9.1 Confusion matrix of system evaluation (searching for “network”) 130
9.2 Confusion matrix of system evaluation (searching for “range”) 130
9.3 Subset of a classified result set for keyword “range” 131
9.4 BankSearch Corpus categories 133
9.5 “rule” restructured SynSets for clustering 133
9.6 Confusion matrix for the term “rule” (Sense Folder Classification) 134
9.7 Confusion matrix for the term “rule” (k-Means Clustering) 134
9.8 Re-classification matrix for the term “rule” 134
9.9 Comparison of WordNet SynSets (Operation) and restructured SynSets for clustering 136
9.10 Confusion matrix for the term “operation” (Sense Folder Classification) 137
9.11 Confusion matrix for the term “operation” (k-Means Clustering) 137
9.12 Re-classification matrix for the term “operation” 137
9.13 Confusion matrix for the term “operation” (Sense Folder Classification) 137
9.14 Confusion matrix for the term “operation” (k-Means Clustering) 137
9.15 Re-classification matrix for the term “operation” 138
9.16 Comparison of WordNet SynSets (Rule) and restructured SynSets for clustering 138
9.17 Comparison of WordNet SynSets (Operation) and restructured SynSets for clustering 139
9.18 Evaluation of the combination of linguistic parameters and manual merging methods. 140
9.19 Sense Folder Corpus Overview 144
9.20 Sense Folders for the term “argument” 146
9.21 Fine-Grained Classification Baselines 148
9.22 Coarse-Grained Classification “Random” Baselines 148
9.23 Coarse-Grained Classification “First Sense” Baselines 148
9.24 Coarse-Grained Classification “Most Frequent Sense” Baselines 149
9.25 Components of Fine-Grained Evaluation Procedures 150

9.26 Accuracy achieved with Fine-Grained Classification (`Tf+NoMerge+NoStem`) 153
9.27 Accuracy achieved with Fine-Grained Classification (`Tf+NoMerge+Stem`) 154
9.28 Accuracy achieved with Fine-Grained Classification (`TfxIdf+NoMerge+NoStem`) 156
9.29 Accuracy achieved with Fine-Grained Classification (`TfxIdf+NoMerge+Stem`) 157
9.30 Components of Coarse-Grained Evaluation Procedures 159
9.31 Accuracy achieved with Coarse-Grained Classification (`TfxIdf+Merge+NoStem`) 165
9.32 Accuracy achieved with Coarse-Grained Classification (`Tf+Merge+Stem`) 165
9.33 Accuracy achieved with Multi-Stage Coarse-Grained Classification (`TfxIdf+Merge+NoStem`) 166
9.34 Accuracy achieved with Multi-Stage Coarse-Grained Classification (`Tf+Merge+Stem`) 166