Universidad Nacional de Educación a Distancia
Escuela Técnica Superior de Ingeniería Informática
Departamento de Lenguajes y Sistemas Informáticos

Collaboratively Authored Web Contents as Resources for Word Sense Disambiguation and Discovery.

TESIS DOCTORAL

M. Celina Santamaría Recio
Licenciada en Matemáticas
Licenciada en Ciencias Físicas
2010

UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA
Escuela Técnica Superior de Ingeniería Informática
Departamento de Lenguajes y Sistemas Informáticos

COLLABORATIVELY AUTHORED WEB CONTENTS AS RESOURCES FOR WORD SENSE DISAMBIGUATION AND DISCOVERY.

M. Celina Santamaría Recio
Licenciada en Matemáticas por la Universidad de Santiago de Compostela
Licenciada en Ciencias Físicas por la UNED

Directores:

Julio Gonzalo Arroyo

Profesor Titular de Universidad del Departamento de Lenguajes y Sistemas Informáticos
de la Universidad Nacional de Educación a Distancia

M. Felisa Verdejo Maillo

Catedrática de Universidad del Departamento de Lenguajes y Sistemas Informáticos
de la Universidad Nacional de Educación a Distancia

A mis padres, Benjamín y Celina;
a mis hijos, Antonio y Celina;
a Antonio.

Agradecimientos

Me sería imposible condensar en unas pocas líneas mi gratitud hacia tantas personas y por tantos motivos. Por otra parte, éste no parece el contexto más adecuado para exponer sentimientos ya que, los que son auténticos, deben protegerse.

Estoy en deuda con Julio Gonzalo, por sus siempre acertadas indicaciones, por su excepcional labor en la dirección de esta tesis, y sobre todo, por su incondicional apoyo a lo largo de todos estos años. Agradezco a Felisa Verdejo su valiosa supervisión, y la confianza que depositó en mí cuando empecé esta investigación. Ella me dió en primer lugar la oportunidad de realizarla. Agradezco a mi colega Javier Artiles su productiva colaboración en la parte final del trabajo.

Podría nombrar a muchos amigos; a todos les agradezco su cariño y comprensión.

Quiero expresar mi gratitud a mis padres, Benjamín y Celina, por su amor y dedicación; a Fermín y María Teresa, por ser para mí unos padres; a mis hermanos, María José, Mercedes y José Luis, compañeros y amigos en la vida y a María Teresa y Fermina, por ser hermanas para mí.

También todo mi gratitud hacia mis hijos, Antonio y Celina, por darme tantas facilidades, al ser los mejores hijos que podría haber imaginado, y mi mayor logro.

Mi más profundo agradecimiento a Antonio Yáñez por su continua participación en el desarrollo de esta tesis. Su inestimable ayuda, su generosa y eficaz colaboración y su capacidad para solucionar problemas han sido una de las claves de la investigación. Mi gratitud también por toda una vida.

Finalmente, agradezco a la Consellería de Educación y Ordenación Universitaria de la Xunta de Galicia el apoyo prestado para la realización de este trabajo, mediante la concesión de una licencia por estudios durante el curso académico 2008-2009.

Abstract

In this research we have addressed the use of collaboratively authored Web contents as a source of lexical information for Word Sense Disambiguation and Discovery. We have focused on two sources, the Open Directory Project (ODP) and Wikipedia, both collaboratively authored, although representatives of two different kinds of resource: ODP hierarchically organizes Web sites by domains, thus containing implicit lexical information, whereas Wikipedia is a large coverage, updated encyclopedic repository of explicit world knowledge associated to a lexicon. Compared with standard Lexical Databases (such as WordNet), such resources have a much larger coverage, a richer connection to world knowledge, and a much faster updating pace. Compared with the whole Web as a corpus, collaboratively authored Web contents are much cleaner and reliable. On the other hand, they are more structured and contain more explicit linguistic information than the full Web, although their size is several orders of magnitude smaller. The question, therefore, is how useful they can be for Natural Language Processing, and in particular to acquire information about word senses that can be used in practical applications.

Our research has consisted of two main tasks. In the first, we have tried to use ODP in order to enrich an existing lexical database (WordNet), making explicit connections between Web directories and word senses in WordNet, also exploiting such connections for Word Sense Disambiguation, sense clustering and discovery of new senses. In the second, we have studied whether Wikipedia can replace Wordnet as a sense inventory to organize Web search results for one word ambiguous queries.

Our main accomplishments are:

- We have described an algorithm that combines lexical information from WordNet with Web directories from the ODP to associate word senses with such directories. These associations can be used as rich characterizations to acquire sense-tagged corpora automatically, cluster topically related senses, and detect sense specializations. The algorithm is evaluated for the 29 nouns (147 senses) used in the Senseval 2 competition, obtaining 148 (word sense, Web directory) associations covering 88% of the domain-specific word senses in the test data with 86% accuracy. The richness of Web directories as sense

characterizations is evaluated in a supervised word sense disambiguation task using the Senseval 2 test suite.

The results indicate that, when the directory/word sense association is correct, the samples automatically acquired from the Web directories are nearly as valid for training as the original Senseval 2 training instances. The results support our hypothesis that Web directories are a rich source of lexical information, more structured and useful than the whole Web as a corpus.

- We have studied whether it is possible to use sense inventories to improve Web search results for one word queries in which ambiguity cannot be resolved and thus, the search engine should either promote diversity in search results or organize them according to the different query interpretations. To answer this question, we have compared two broad-coverage lexical resources of a different nature: Wordnet, as a de-facto standard used in Word Sense Disambiguation experiments and Wikipedia, as a large coverage, updated encyclopedic resource which may have a better coverage of relevant senses in Web pages.

  In this case, our results indicate that (i) Wikipedia has a much better coverage of search results, (ii) the relative distribution of senses in search results can be estimated using the internal graph structure of the Wikipedia and the relative number of visits received by each sense in Wikipedia, and (iii) associating Web pages to Wikipedia senses with simple and efficient algorithms, we can produce modified rankings that cover four times more Wikipedia senses than the original search engine rankings.

Along our research we have built, and made publicly available, two resources for the research community:

- A massive association of Web directories to WordNet senses which characterizes 24,558 nouns and 27,383 senses from WordNet with domain information from ODP.

- A testbed for experiments in search results diversity, consisting of (i) 40 highly ambiguous nouns, (ii) two alternative inventories of senses (derived from Wikipedia and WordNet respectively) together with useful lexical information for the senses, and (iii) a collection of 4000 documents, manually associated with the most appropriate senses in both inventories.

As an overall conclusion, we have shown that collaboratively authored Web contents are a very valuable source of lexical information, either for enriching linguistic resources, as we have done with ODP, or to replace them in specific applications as in our study of diversity using Wikipedia.

Resumen

En esta investigación, hemos abordado el uso de contenidos de la Red creados colaborativamente, considerándolos fuentes de información léxica, para realizar desambiguación y descubrimiento de sentidos. Nos hemos centrado en dos recursos, el Open Directory Project (ODP) y Wikipedia, ambos creados colaborativamente, aun cuando representan dos planteamientos diferentes: ODP organiza jerárquicamente sitios Web por dominios, y por tanto contiene información implícita sobre dichos temas, mientras que Wikipedia es un repositorio enciclopédico de conocimiento explícito asociado a un lexicón, de amplia cobertura y continuamente actualizado. En comparación con bases de datos léxicas estándar (como WordNet), estos recursos tienen una cobertura mucho más amplia, una conexión más rica con el conocimiento universal y un ritmo de actualización mucho más rápido. En comparación con la Web completa usada como corpus, los contenidos Web creados colaborativamente son mucho más limpios y fiables. Por otra parte, están más estructurados y contienen más información lingüística explícita que la Web en sí, pese a que su tamaño es varios órdenes de magnitud menor. La cuestión es, por tanto, hasta que punto pueden ser útiles en el Procesamiento del Lenguaje Natural, y en particular para adquirir información sobre sentidos utilizable en aplicaciones prácticas.

Nuestra investigación ha desarrollado dos tareas principales. En la primera, hemos intentado usar ODP para enriquecer una base de datos léxica ya existente (WordNet), estableciendo conexiones explícitas entre directorios de la Red y sentidos de WordNet, aprovechando además dichas conexiones para desambiguación de sentidos, agrupación de sentidos y descubrimiento de nuevos sentidos. En la segunda, hemos estudiado si Wikipedia puede reemplazar a WordNet como inventario de sentidos, para organizar resultados de búsqueda en la Red en el caso de consultas ambiguas formadas por una sola palabra.

Nuestros principales logros son:

- Hemos descrito un algoritmo que combina información léxica procedente de WordNet con directorios de la Red, para asociar sentidos de palabras con dichos directorios. Estas asociaciones pueden usarse como ricas caracterizaciones para adquirir automáticamente córpora etiquetados por sentidos, para

agrupar sentidos relacionados por temas, y para detectar especializaciones de sentidos. La evaluación del algoritmo para los 29 nombres (147 sentidos) usados en la competición Senseval 2 aporta 148 asociaciones de sentidos con directorios, cubriendo el 88% de los sentidos relativos a un dominio (tema) específico dentro del conjunto de evaluación, con un 86% de exactitud. La riqueza de los directorios de la Red en el papel de caracterizaciones de sentidos se evalúa en una tarea de desambiguación supervisada, usando el conjunto de test de Senseval 2.

Los resultados indican que cuando la asociación directorio/sentido es correcta, los ejemplos adquiridos automáticamente de los directorios de la Red son prácticamente tan válidos para entrenamiento como las frases de entrenamiento originales proporcionadas por Senseval 2. Estos resultados apoyan nuestra hipótesis de que los directorios de la Red son una rica fuente de información léxica, más estructurada y útil que la Red vista como un corpus.

- Hemos estudiado si es posible usar inventarios de sentidos para mejorar resultados de búsqueda en la Red, para consultas formadas por una sola palabra en las cuales la ambiguedad no puede resolverse, por lo que el motor de búsqueda debería promover la diversidad en dichos resultados, o bien organizarlos de acuerdo a diferentes interpretaciones de la consulta. Para responder a esta pregunta, hemos comparado dos fuentes léxicas de amplia cobertura y de diferente naturaleza: WordNet, el estándar de-facto en la experimentación sobre desambiguación de sentidos, y Wikipedia, generada colaborativamente y que por tanto puede ofrecer sentidos más cercanos a los que son relevantes en las páginas Web.

  En este caso, nuestros resultados muestran que (i) Wikipedia ofrece una cobertura mucho mayor para los resultados de búsqueda, (ii) la distribución de sentidos en los resultados de búsqueda puede estimarse usando la estructura interna (grafo) de Wikipedia y el número relativo de visitas a cada sentido en Wikipedia, y (iii) con la asociación de páginas Web con sentidos de Wikipedia mediante algoritmos simples y eficientes, podemos modificar el orden de presentación de los resultados cubriendo cuatro veces más sentidos de Wikipedia que con el orden original presentado por el motor de búsqueda.

Durante nuestra investigación, hemos construído dos recursos, los cuales hemos puesto a disposición de la comunidad investigadora:

- Una asociación masiva entre directorios de la Red y sentidos de WordNet que caracteriza 24,558 nombres y 27,383 sentidos de WordNet con información de dominio procedente de ODP.

- Un marco de experimentación (testbed) referido a diversidad en resultados de búsqueda, , consistente en (i) 40 nombres altamente ambiguos, (ii) dos inventarios de sentidos alternativos (derivados de Wikipedia y WordNet respectivamente) acompañados de útil información léxica para los sentidos, y (iii) una colección formada por 4000 documentos, manualmente asociados con los sentidos más adecuados en ambos directorios.

Como conclusión general, hemos mostrado que los contenidos de la Red creados colaborativamente son una valiosa fuente de información léxica, tanto para enriquecer recursos léxicos, como hemos hecho con ODP, como para reemplazarlos en aplicaciones específicas, como en nuestro estudio de la diversidad usando Wikipedia.

Contents

Contents 15

List of Figures 19

List of Tables 21

1 Introduction 23

2 State of the Art 29
2.1 Collaboratively Authored Web Contents 29
2.1.1 Open Directory Project 30
2.1.2 Wikipedia 33
2.1.3 Other Collaboratively Authored Resources 36
2.1.4 Resources Authored by Aggregation 37
2.1.5 Comparison with WordNet 38
2.2 Collaborative Authored Web Contents in NLP Techniques 43
2.2.1 Cross-Language Alignment 43
2.2.2 Measurements of Semantic Relatedness 44
2.2.3 Clustering and Classification 45
2.2.4 Question Answering 48
2.2.5 Summarization 49
2.3 Applications of Wikipedia to Word Sense Disambiguation and Discovery 50
2.3.1 Acquisition of Lexical Information 50
2.3.2 Acquisition of Sense-Tagged Corpora 52
2.3.3 Recognition and Semantic Disambiguation of Named Entities 54
2.4 Conclusions 56

3 Automatic Association of Web Directories to WordNet Senses 57
3.1 Introduction 57
3.2 Algorithm 60

3.2.1 Querying ODP Structure 60
3.2.2 Representing Retrieved Directory Descriptions 61
3.2.3 Representing WordNet Senses 61
3.2.4 Sense/Directory Comparisons 62
3.2.5 Candidate Sense/Directory Associations 62
3.2.6 Filters 62
3.2.7 Confidence Score 64
3.3 Evaluation 67
3.3.1 Coverage 67
3.3.2 Quality 69
3.4 Example Application: Automatic Acquisition of Sense-Tagged Corpora 72
3.5 Massive Processing of WordNet Nouns 77
3.6 Conclusions 78

4 Use of Wikipedia to Organize Web Search Results 79
4.1 Introduction 79
4.2 Test Bed 81
4.2.1 Set of Nouns 81
4.2.2 Sense Inventories 82
4.2.3 Information about Wikipedia Senses 83
4.2.4 Set of Documents 86
4.2.5 Manual Annotation 87
4.3 Coverage of Web Search Results: Wikipedia vs WordNet 89
4.4 Diversity in Google Search Results 90
4.5 Sense Frequency Estimators for Wikipedia 96
4.5.1 Estimators 96
4.5.2 Pairwise Frequency Correlations 101
4.5.3 Correlations between Frequency Estimators 105
4.5.4 Rank Correlations 106
4.5.5 Optimized Frequency Estimator for Wikipedia Senses . . 110
4.6 Association of Wikipedia Senses to Web Pages 110
4.6.1 VSM Approach 114
4.6.2 WSD Approach 114
4.6.3 WSD-Based Algorithm 115
4.6.4 Classification Results 125
4.6.5 Precision/Coverage Trade-off 127
4.6.6 Using Classification to Promote Diversity 131
4.7 Related Work 131
4.7.1 Diversity 132
4.7.2 Establishing Relative Frequencies for Senses 133

4.8 Conclusions 134

5 Conclusions 137

5.1 Main Contributions 137

5.1.1 Results 137

5.1.2 Resources 139

5.2 Further Work 140

Bibliography 143

A Complementary Tables and Figures 155

List of Figures

1.1 First Web search results for the query *jaguar*. 25

1.2 ODP search results for the query *jaguar*. 25

2.1 ODP Categories for the search *circuit* 31

2.2 Subcategories for the Category *Electronics and Electrical*. Horizontal lines denote sort priority 31

2.3 ODP Categories and Web sites for the search *circuit* (partial view) 32

2.4 Wikipedia Article for *Oasis (Band)*, partial view. 34

2.5 Detail of a Wikipedia article. Highlighted words link to their corresponding Wikipedia entry 34

2.6 Disambiguation Page for *jaguar*, partial view. 35

2.7 Wikipedia Categories for *Oasis (band)* 36

3.1 Contents of an ODP Web directory associated to *circuit 5* (racing circuit). 59

3.2 Results of the association algorithm for *circuit* 66

3.3 Examples of training material for *circuit* 73

4.1 Nouns selected for our test bed 82

4.2 Wikipedia senses and definitions for *camel* 85

4.3 WordNet senses for *camel* 86

4.4 Annotation process: classification of search results into the appropriates senses 88

4.5 Average number of senses per noun at different points in the search engine rank. 92

4.6 Ratio of Wikipedia/WordNet senses covered by search results . . . 94

4.7 Coverage of search results by Wikipedia/WordNet senses 95

4.8 Incoming links example: all these Wikipedia pages contain at least one link to the article for *amazon.com* 97

4.9 Visits to the Wikipedia article for *amazon.com* 98

4.10 Incoming Links: upper bars represent the average number of incoming links for senses which are represented in the search results, and lower bars for the others. 99
4.11 Visits in May 2009: upper bars represent the average for senses present in the search results, lower bars for the others. 100
4.12 Average of visits to senses in June 2009 102
4.13 Averages of visits in July 2009: unexpected increase of visits for the noun tesla 103
4.14 Google logo in tribute to Nikola Tesla, linking to the Wikipedia entry for *Nikola Tesla*. 104
4.15 Accesses to tesla in July 2009 104
4.16 Averages correlation between relative frequency estimators (visits and links) and actual relative frequencies in our test bed. 107
4.17 Average correlation between frequency ranks from our estimators (links and visits) and actual ranks in our test bed 109
4.18 Optimized frequency estimator: correlation with actual sense frequencies 111
4.19 WSD-based assignment of senses to documents in search results . 116
4.20 Example of sense representation: above, text from the Wikipedia entry for *Amazon Basin*; below, features extracted from the text . . 118
4.21 Training model: Acquisition of examples 119
4.22 Test Model: representation of the document 121
4.23 Document 014 for *amazon* (partial sight) 122
4.24 Document 014 for *amazon* (partial sight) 122
4.25 Example of document representation: document 014 for *amazon* . 123
4.26 WSD-based algorithm using frequencies information 125
4.27 Results of classification experiments using variations of TiMBL . 128
4.28 Results for the TiMBL-core+freq run compared to the baselines . 129
4.29 Precision/Coverage curve for VSM-GT+freq classification algorithm130

A.1 Textual information in the document 014 for Figure 4.23 167
A.2 Textual information in the document 014 for Figure 4.24 168

List of Tables

3.1 Coverage of nouns in the Senseval 2 test bed. 68
3.2 Precision over Senseval 2 nouns. 69
3.3 Confidence distribution. 71
3.4 Correlation between confidence and correctness 71
3.5 Relevance of the directories in the test set. 72
3.6 Training material obtained for the WSD experiment 74
3.7 Results of supervised WSD 75
3.8 Massive association of ODP directories to WordNet 1.7 nouns . . 77

4.1 Number of senses for the test nouns: Wikipedia vs WordNet . . . 84
4.2 Detected senses and coverage of search results: Wikipedia vs WordNet 89
4.3 Coverage of search results: Wikipedia vs. WordNet 90
4.4 Search results: detected senses presented by subsets and aggregated 91
4.5 Search results: averages of detected senses per noun 91
4.6 Search results: averages of ratios of detected senses 93
4.7 Search results: Coverage 96
4.8 Percentage of Wikipedia relevant senses which are represented in search results 98
4.9 Pairwise ordering equivalence probabilities (more frequency of links/visits $\Longleftrightarrow$ more frequency of documents) 104
4.10 Baseline precision by using the most frequent sense according to the combined-frequencies value 113
4.11 Classification Results 126
4.12 Enhancement of Search Results Diversity 131

A.1 Search results for the first 10 documents and for the Band nouns . 156
A.2 Search results for the first 25 documents and for the Band nouns . 157
A.3 Search results for the first 50 documents and for the Band nouns . 158
A.4 Search results for the first 75 documents and for the Band nouns . 159
A.5 Search results for the first 100 documents and for the Band nouns . 160

A.6 Search results for the first 10 documents and for Senseval nouns . 160
A.7 Search results for the first 25 documents and for Senseval nouns . 161
A.8 Search results for the first 50 documents and for Senseval nouns . 161
A.9 Search results for the first 75 documents and for Senseval nouns . 162
A.10 Search results for the first 100 documents and for Senseval nouns . 162
A.11 WSD experiments: TiMBL-core results 163
A.12 WSD experiments: TiMBL-inlinks results 164
A.13 WSD experiments: TiMBL-all results 165
A.14 WSD experiments: TiMBL-core+freq results 166

Chapter 1

Introduction

Word Senses in dictionaries and Word Senses in the Web

Is *jaguar* an ambiguous word? According to the Webster's Encyclopedic Unabridged Dictionary, the noun *jaguar* is a monosemous word, defined as *a large, ferocious, spotted feline*. We find an almost identical definition for the unique sense of *jaguar* in WordNet, which is perhaps the most important lexical database for Natural Language Processing in English. And this is probably the definition for jaguar that we learnt at school (what linguists sometimes call the *salient* meaning of jaguar).

But it is also possible to ask a Web search engine (which is perhaps more usual nowadays than using a dictionary). In this way, we are looking for the actual uses of the word *jaguar* in the largest corpus ever built. If we query Google, five of the six best ranked results relate to the *Jaguar* car brand, and only one refers to the feline (see Figure 1.1). This is not necessarily a defect of Webster's and WordNet lexicons: after all, the car brand is a named entity, an instance of a class (car brands) which does not need to be included in a dictionary. It is world knowledge rather than linguistic knowledge.

But let us suppose that we try to apply Natural Language Processing techniques to organize Google search results according to the possible meanings of the word "jaguar". As conventional dictionaries cannot capture the actual diversity of meanings and denotations of jaguar in the Web, our only chance is resorting to some kind of generic sense discovery or text clustering technique such as those used by meta-search engines such as Clusty (www.clusty.com). And of course such general techniques are not entirely satisfactory. For instance, Clusty is able to find a group labeled "cars" and a group labeled "Panthera onca", which is fine, but also displays, at the same level, groups labeled "Jacksonville" and "Treat the needs of each individual customer" (?). Results could be much more satisfactory if Clusty could rely on a lexical database with a complete coverage of the senses and denotations of every possible query word.

Collaboratively authored Web contents

A limitation of current lexical databases is that it is simply not feasible to include (and maintain) all relevant world entities in a dictionary. Or is it? Note that Google results in Figure 1.1 include pointers to actual encyclopedic definitions both for the feline and the car brand in a single reference source: Wikipedia. This is an encyclopedia with a distinctive feature: it is collaboratively edited by around 300,000 active *wikipedians* (from over 11,000,000 registered actually registered). The scale of the resource is, accordingly with the number of editors, incomparably larger than other conventional dictionaries and encyclopedias, with a comparable level of reliability.

Wikipedia is a distinctive example of collaboratively authored Web contents, a large-scale generalist resource (as opposed to domain-specific efforts such as IMDb.com, the Internet Movie DataBase). Perhaps the other prominent example is the Open Directory Project, which aims to organize Web contents at large in a collaboratively edited structure of web directories.

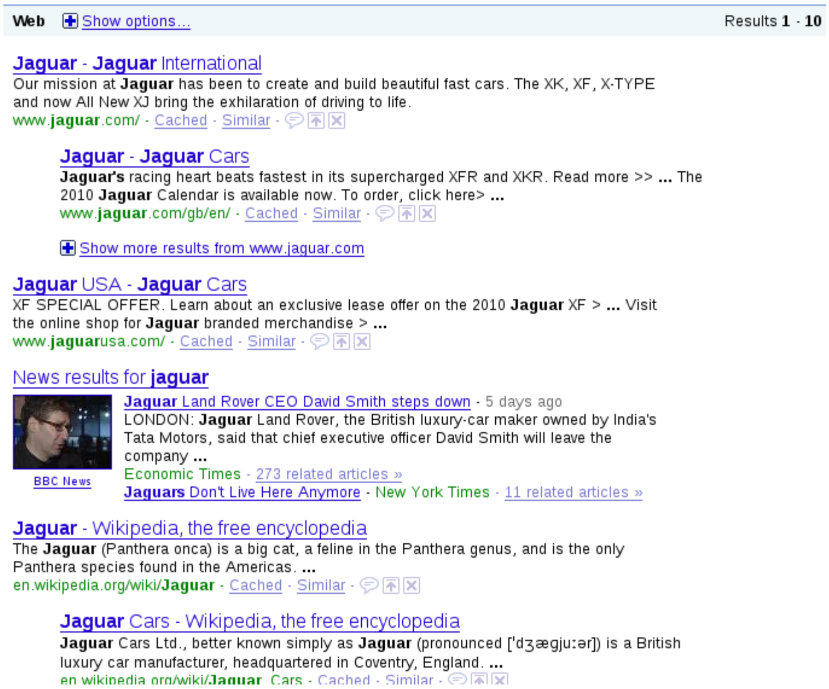
Research goals

The goal of our research is to investigate to what extent such kind of resources can complement, enrich or even replace conventional lexical databases. Our focus is on collaboratively edited contents (such as Wikipedia and ODP), as opposed to resources which are collaboratively built by aggregation, such as the folksonomies in Flickr or Delicious, where users aggregate information but do not explicitly collaborate or edit each other's contents. We initially discard aggregated sources not because they do not contain valuable lexical information, but because it is not explicit: it requires extensive data and text mining, and filtering massive amounts of noise, and therefore requires a completely different methodology.

Compared with standard Lexical Databases (such as WordNet), collaboratively edited Web contents have a much larger coverage, a richer connection to world knowledge, and a much faster updating pace. And compared with the whole Web as a corpus, collaboratively authored Web contents are supposed to be much cleaner and reliable. At the same time, they are more structured and contain more explicit linguistic information, although their size is several orders of magnitude smaller than the full Web. The question, therefore, is how useful they can be for Natural Language Processing, and in particular to acquire information about word senses that can be used in practical applications.

We have experimented with the two probably largest generalist Web sources of knowledge: the Open Directory Project and Wikipedia. The former is an implicit source of world and lexical knowledge, whereas the latter is an explicit encyclopedic resource. Our research consists of two main efforts:

- In our research with ODP, we have focused on how to enrich a lexical database (WordNet), studying how to map ODP web directories into word senses and how to exploit such connections to automatically acquire examples for supervised Word Sense Disambiguation, and to discover new senses and sense extensions.



Web ➕ Show options... Results **1** - **10**

Jaguar - **Jaguar** International
Our mission at **Jaguar** has been to create and build beautiful fast cars. The XK, XF, X-TYPE and now All New XJ bring the exhilaration of driving to life.
www.**jaguar**.com/ - Cached - Similar

Jaguar - **Jaguar** Cars
Jaguar's racing heart beats fastest in its supercharged XFR and XKR. Read more >> **...** The 2010 **Jaguar** Calendar is available now. To order, click here> **...**
www.**jaguar**.com/gb/en/ - Cached - Similar

➕ Show more results from www.jaguar.com

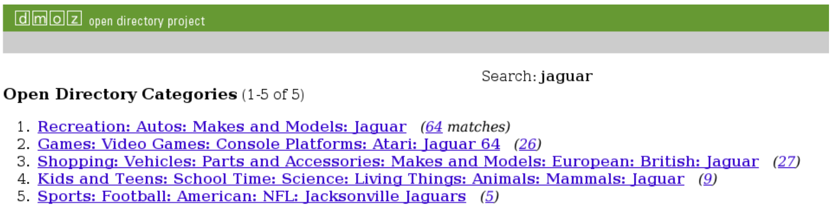
Jaguar USA - **Jaguar** Cars
XF SPECIAL OFFER. Learn about an exclusive lease offer on the 2010 **Jaguar** XF > **...** Visit the online shop for **Jaguar** branded merchandise > **...**
www.**jaguar**usa.com/ - Cached - Similar

News results for **jaguar**
Jaguar Land Rover CEO David Smith steps down - 5 days ago
LONDON: **Jaguar** Land Rover, the British luxury-car maker owned by India's Tata Motors, said that chief executive officer David Smith will leave the company **...**
Economic Times - 273 related articles »
Jaguars Don't Live Here Anymore - New York Times - 11 related articles »
BBC News

Jaguar - Wikipedia, the free encyclopedia
The **Jaguar** (Panthera onca) is a big cat, a feline in the Panthera genus, and is the only Panthera species found in the Americas. **...**
en.wikipedia.org/wiki/**Jaguar** - Cached - Similar

Jaguar Cars - Wikipedia, the free encyclopedia
Jaguar Cars Ltd., better known simply as **Jaguar** (pronounced ['dʒægjuːər]) is a British luxury car manufacturer, headquartered in Coventry, England. **...**
en.wikipedia.org/wiki/**Jaguar**_Cars - Cached - Similar

Figure 1.1: First Web search results for the query *jaguar*.



dmoz open directory project

Search: **jaguar**

Open Directory Categories (1-5 of 5)

1. Recreation: Autos: Makes and Models: Jaguar *(64 matches)*
2. Games: Video Games: Console Platforms: Atari: Jaguar 64 *(26)*
3. Shopping: Vehicles: Parts and Accessories: Makes and Models: European: British: Jaguar *(27)*
4. Kids and Teens: School Time: Science: Living Things: Animals: Mammals: Jaguar *(9)*
5. Sports: Football: American: NFL: Jacksonville Jaguars *(5)*

Figure 1.2: ODP search results for the query *jaguar*.

- In our research with Wikipedia, we have investigated whether a collaboratively authored Web resource (Wikipedia) can effectively replace a conventional lexical database (WordNet) for an Information Access problem: the organization of Web search results for ambiguous queries (such as *jaguar* in our initial example).

Research with ODP

In our work with ODP, we have exploited the fact that web directories provide rich domain information and world knowledge, which is something missing in WordNet. Figure 1.2 shows an example of web directories found in ODP for the query *jaguar*.

Our first goal has been to design an algorithm that combines lexical information from WordNet with Web directories from the ODP to associate word senses with such directories. The algorithm has been evaluated for the 29 nouns (147 senses) used in the Senseval 2 competition, obtaining 148 (word sense, Web directory) associations covering 88% of the domain-specific word senses in the test data with 86% accuracy.

Then we have used these associations as rich characterizations to acquire sense-tagged corpora automatically and detect sense specializations. The richness of Web directories as sense characterizations has been evaluated in a supervised word sense disambiguation task using the Senseval 2 test suite. Our results indicate that, when the directory/word sense association is correct, the samples automatically acquired from the Web directories are nearly as valid for training as the original Senseval 2 training instances. This supports our initial hypothesis that Web directories are a rich source of lexical information, smaller but more reliable than the whole Web as a corpus.

Finally, we have applied our association algorithm to a relevant set of WordNet nouns, obtaining a massive association of Web directories to WordNet senses which characterizes a substantial amount of nouns and senses from WordNet with domain information from ODP.

Research with Wikipedia

In our work with Wikipedia, we have studied whether it is possible to use sense inventories to improve Web search results for one word queries, in which ambiguity cannot be resolved and the search engine should promote diversity in search results or organize them according to the different query interpretations (as in our *jaguar* example).

To answer this question, we have built a test set to compare Wikipedia and WordNet coverage of word senses in search results for ambiguous, one word queries. The testbed consists of (i) 40 highly ambiguous nouns, (ii) two alternative inventories of senses (derived from Wikipedia and WordNet respectively) together with useful lexical information for the senses, and (iii) a collection of 4000 documents (100 documents per noun as retrieved by the Google search engine), manually associated with the most appropriate senses in both inventories.

In this case, our results indicate that (i) Wikipedia has a much better coverage of Web search results, (ii) the relative distribution of senses in search results can be estimated using the internal graph structure of the Wikipedia and the relative number of visits received by each sense in Wikipedia, and (iii) associating Web pages to Wikipedia senses with simple and efficient algorithms, we can produce modified rankings that cover four times more Wikipedia senses than the original search engine rankings.

In our research with Wikipedia and WordNet, we have considered two ways of using collaboratively authored Web contents: enrichment and replacement of WordNet as a de-facto standard lexical database. We have approached the problem of Word Sense Disambiguation in two different ways: a canonical disambiguation, in which word occurrences are disambiguated in context; and an Information

Access version, in which documents are assigned to an appropriate sense for the query word that retrieves them. In both cases, Web contents are used for sense discovery, and in particular to recognize and handle named entities denoted by lexical items with related or unrelated dictionary senses. In the experiment with ODP, this leads to the enrichment of WordNet with sense extensions. In our experiments with Wikipedia, we directly replace WordNet definitions - of little use for web search, as the results of our research prove - with Wikipedia entries, that incorporate world knowledge which is essential in a Web search problem.

Thesis outline

The rest of this doctoral dissertation is organized as follows:

- In Chapter 2, we discuss the State of the Art in relation to our research goals. We start with a description of collaboratively authored Web contents, and how these are differentiated with collaborative resources built by aggregation (such as folksonomies) Then we review how these resources have been used in a variety of Natural Language Processing techniques. Finally, we focus on their applications to Word Sense Disambiguation and Discovery.

- Chapter 3 describes our experiments with ODP Web directories and WordNet. We start by proposing an algorithm to associate Web directories with WordNet senses and discovery new (hyponym) senses. Then we evaluate the algorithm for coverage, precision and quality on the Senseval-2 lexical sample testbed, and finally we process all suitable word senses to build an enriched version of the WordNet lexical database.

- Chapter 4 is devoted to our comparison of WordNet and Wikipedia as sense inventories to organize Web search results. We start by describing the design of our testbed; then we compare how WordNet and Wikipedia cover the meanings in search results, and we also use them to estimate Web search results diversity. Then we test two types of information for Wikipedia entries (incoming internal links and number of visits) as estimators for sense frequencies in search results. Finally, we apply two types of techniques (one based on Word Sense Disambiguation and another one based on the Vector Space Model) to associate Web pages in search results with Wikipedia senses, and we show how the results can be effectively used to promote search results diversity.

- Chapter 5 summarizes the main conclusions of our work and discusses future work.

Chapter 2

State of the Art

This chapter is a review of previous research on Word Sense Disambiguation and Discovery using collaboratively authored Web resources. We start by describing this type of Web content, paying special attention to the Open Directory Project (ODP) and Wikipedia, which are the resources used in our research. Then we will summarize the use of ODP and Wikipedia for Natural Language Processing (NLP) applications in general. Finally we focus on their use for Word Sense Disambiguation and Discovery.

2.1 Collaboratively Authored Web Contents

In our research, we have preferred collaboratively authored contents because in them knowledge is explicit, or at least topically organized, and they are more reliable, as the continuous supervision of users generally prevents the information from being manipulated. Among them, we have selected ODP and Wikipedia because they are possibly the largest non-specialized lexical resources in the Web which are available for direct download. Altogether they give us the opportunity of experimenting with explicit (Wikipedia) and implicit (ODP) lexical information.

According to how they are generated and maintained, social resources can be classified in two groups: collaboratively authored and authored by aggregation. The former imply a process of continuous updating and improvement of the stored knowledge by the contributors, who are allowed to add new information or correct the data they do not agree with. In contents generated by aggregation, on the other hand, each person makes independent contributions, and structure emerges as a result of aggregation. The typical structure of contents generated by aggregation is a folksonomy. Although we have focused on collaboratively Web resources, for comparison purposes we will also briefly describe some resources generated by aggregation. Finally, we will also include a description of WordNet as the de facto standard lexical database with semantic relations. WordNet will be used in our research as a lexicon to be enriched by Web contents, and also as a reference to evaluate Web contents comparatively with language engineering resources.

2.1.1 Open Directory Project

The Open Directory Project (ODP)[1] is the largest, most comprehensive human-edited directory of the Web. It is generated and developed by a huge, global community of volunteer editors. Its use is totally free, as it was founded in the spirit of the Open Source movement. RDF dumps of ODP data are available in http://rdf.dmoz.org. For our research, we downloaded the file structure.rdf.u8, which provides the hierarchical structure, the category names and their descriptions.

The main goal of ODP is to provide the means for the Internet to organize itself. Each contributor is required to organize a small part of the Web, and to select the relevant contents in an particular topic, sharing his results with the rest of users. Spreading this concept to thousands of contributors, the final result is a catalog of the Web. The ODP is considered as some of the most important editor/contributor projects of the 20th century. It is currently formed by more than 600.000 categories, which classify over 4.000.000 Web sites.

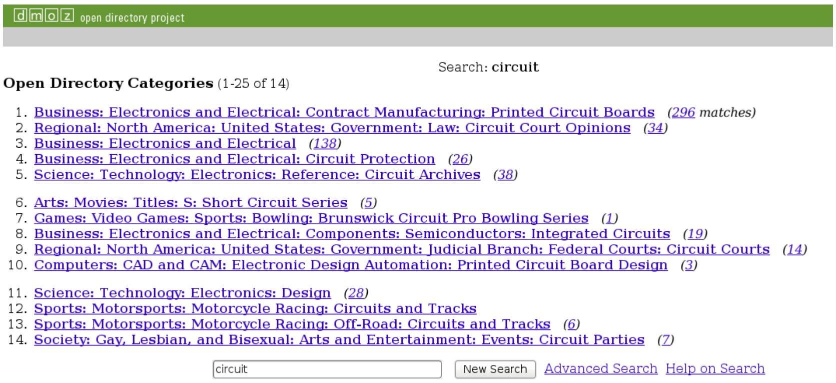
ODP is a knowledge base in which concepts are organized in a hierarchy with multiple inheritance. The following elements are particularly relevant when considering ODP as a lexical resource:

Categories ODP categories are the nodes of the directed acyclic graph structure (see Figure 2.1). They are used to classify Web pages, under a common concept. From a node, there is access to more specific categories (subcategories) (see Figure 2.2), and also to more general ones. We consider them as topical descriptions.

Descriptions of Categories ODP provides to a high amount of categories with short descriptions. In such descriptions, it is explained the topic to which they refer, or the kind of Web sites that should be catalogued in the category.

URLs ODP also provides with the URLs of the Web sites classified under each category. This sites can be crawled in order to acquire more information about the category (see Figure 2.3). Note that the URLs listed in a given category do not belong to its child nodes; for instance, the node for "Physics" will list web pages about Physics in general, but pages about Theoretical Physics will be listed in the appropriate subnode.

[1] http://www.dmoz.org/

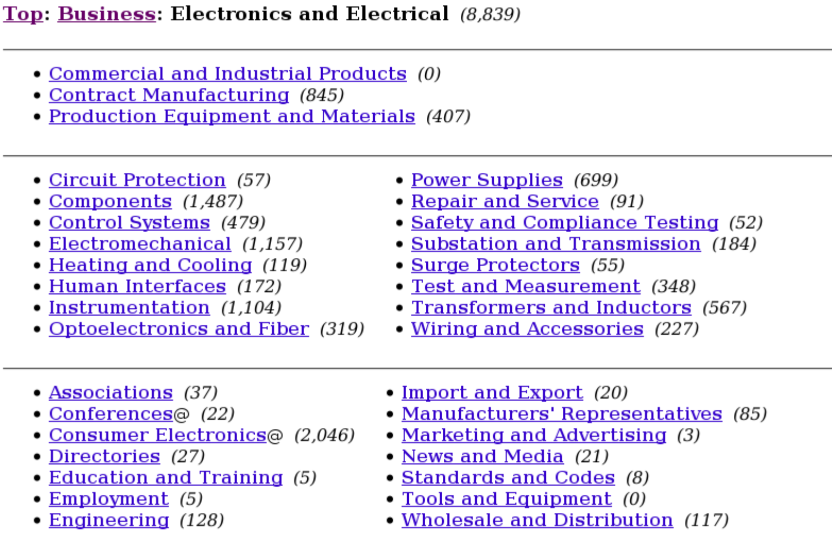


Search: **circuit**

Open Directory Categories (1-25 of 14)

1. Business: Electronics and Electrical: Contract Manufacturing: Printed Circuit Boards *(296 matches)*
2. Regional: North America: United States: Government: Law: Circuit Court Opinions *(34)*
3. Business: Electronics and Electrical *(138)*
4. Business: Electronics and Electrical: Circuit Protection *(26)*
5. Science: Technology: Electronics: Reference: Circuit Archives *(38)*
6. Arts: Movies: Titles: S: Short Circuit Series *(5)*
7. Games: Video Games: Sports: Bowling: Brunswick Circuit Pro Bowling Series *(1)*
8. Business: Electronics and Electrical: Components: Semiconductors: Integrated Circuits *(19)*
9. Regional: North America: United States: Government: Judicial Branch: Federal Courts: Circuit Courts *(14)*
10. Computers: CAD and CAM: Electronic Design Automation: Printed Circuit Board Design *(3)*
11. Science: Technology: Electronics: Design *(28)*
12. Sports: Motorsports: Motorcycle Racing: Circuits and Tracks
13. Sports: Motorsports: Motorcycle Racing: Off-Road: Circuits and Tracks *(6)*
14. Society: Gay, Lesbian, and Bisexual: Arts and Entertainment: Events: Circuit Parties *(7)*

Figure 2.1: ODP Categories for the search *circuit*

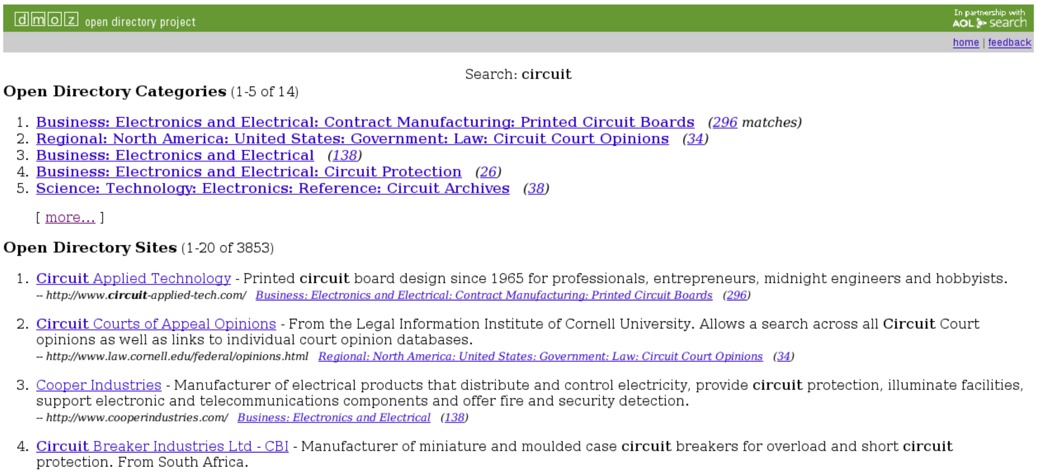

Top: Business: Electronics and Electrical *(8,839)*

- Commercial and Industrial Products *(0)*
- Contract Manufacturing *(845)*
- Production Equipment and Materials *(407)*

- Circuit Protection *(57)*
- Components *(1,487)*
- Control Systems *(479)*
- Electromechanical *(1,157)*
- Heating and Cooling *(119)*
- Human Interfaces *(172)*
- Instrumentation *(1,104)*
- Optoelectronics and Fiber *(319)*
- Power Supplies *(699)*
- Repair and Service *(91)*
- Safety and Compliance Testing *(52)*
- Substation and Transmission *(184)*
- Surge Protectors *(55)*
- Test and Measurement *(348)*
- Transformers and Inductors *(567)*
- Wiring and Accessories *(227)*

- Associations *(37)*
- Conferences@ *(22)*
- Consumer Electronics@ *(2,046)*
- Directories *(27)*
- Education and Training *(5)*
- Employment *(5)*
- Engineering *(128)*
- Import and Export *(20)*
- Manufacturers' Representatives *(85)*
- Marketing and Advertising *(3)*
- News and Media *(21)*
- Standards and Codes *(8)*
- Tools and Equipment *(0)*
- Wholesale and Distribution *(117)*

Figure 2.2: Subcategories for the Category *Electronics and Electrical*. Horizontal lines denote sort priority



Search: **circuit**

Open Directory Categories (1-5 of 14)

1. Business: Electronics and Electrical: Contract Manufacturing: Printed Circuit Boards *(296 matches)*
2. Regional: North America: United States: Government: Law: Circuit Court Opinions *(34)*
3. Business: Electronics and Electrical *(138)*
4. Business: Electronics and Electrical: Circuit Protection *(26)*
5. Science: Technology: Electronics: Reference: Circuit Archives *(38)*

[more...]

Open Directory Sites (1-20 of 3853)

1. **Circuit** Applied Technology - Printed **circuit** board design since 1965 for professionals, entrepreneurs, midnight engineers and hobbyists.
   *-- http://www.**circuit**-applied-tech.com/ Business: Electronics and Electrical: Contract Manufacturing: Printed Circuit Boards (296)*
2. **Circuit** Courts of Appeal Opinions - From the Legal Information Institute of Cornell University. Allows a search across all **Circuit** Court opinions as well as links to individual court opinion databases.
   -- http://www.law.cornell.edu/federal/opinions.html Regional: North America: United States: Government: Law: Circuit Court Opinions (34)
3. Cooper Industries - Manufacturer of electrical products that distribute and control electricity, provide **circuit** protection, illuminate facilities, support electronic and telecommunications components and offer fire and security detection.
   -- http://www.cooperindustries.com/ Business: Electronics and Electrical (138)
4. **Circuit** Breaker Industries Ltd - CBI - Manufacturer of miniature and moulded case **circuit** breakers for overload and short **circuit** protection. From South Africa.

Figure 2.3: ODP Categories and Web sites for the search *circuit* (partial view)

Descriptions of URLs Each URL under a category is enriched with a short description of its contents. This information is very useful in order to acquire textual information very related to the Web site and thus, to the category.

@links @links are used to link from one category to another that could theoretically be a subcategory of the first. For instance, *Top/Science/Technology/Electronics/Conferenc* is located in the list of subcategories of *Top/Business/Electronics and Electrical* with the name of *Conferences@*.

Related Categories Links They establish relations between related categories that do not have a parent-child relationship.

Sort Priority This is used as an alternative to creating an additional layer of subcategories, listing in top positions those most visited or popular categories. It is also used to split the different facets (or perspectives) of a densely populated category. An example of sort priority is shown in Figure 2.2, where the horizontal lines separate the clusters.

First level ODP categories (hanging directly from the *Top* category) are: *Arts, Business, Computers, Games, Health, Home, Kids and Teens, News, Recreation, Reference, Science, Shopping, Society, Sports, World* and a hidden category *Adult*. In our research we discarded the categories (i) *Adult*, which is mainly related to porn stars and titles of adult movies, (ii) *World*, which aggregates the contents in languages other than English, (iii) *News*, because it organizes topics by media and (iv) *Reference*, because it tends to re-direct to reference material (such as dictionaries).

As a weakness of ODP as lexical resource, we must note that (i) the coverage of topics is not complete: indeed, we should only expect good coverage for domain-specific topics, (ii) the coverage is not evenly distributed, and is biased by the different editorial teams.

2.1.2 Wikipedia

Wikipedia[2] defines itself as a multilingual, web-based, free-content encyclopedia project based on an openly-editable model. Wikipedia articles provide links to guide the user to related pages with additional information.

Wikipedia is written collaboratively by thousands of volunteers from around the world. Each contributor can add an article to Wikipedia, or make changes to an existing one, in order to improve or correct its contents, following specific policies and guidelines to ensure the quality of the stored information.

Since its creation, in 2001, Wikipedia has quickly developed, and on February 2010, there are more than 85,000 active contributors, more than 14,000,000 articles in more than 260 languages, and more than 3,000,000 articles in the English version. Every day, hundreds of thousands of visitors from around the world collectively make thousands of edits and create thousands of new articles to augment the knowledge in Wikipedia, which has become the largest encyclopedia ever known.

Besides the online encyclopedia, Wikipedia offers free copies of the available contents, for research or other purposes, which can be downloaded as database dumps. Such dumps are in the form of wikitext source and meta-data embedded in XML, making the relevant information (hyperlinks, categories, etc) easy to find by detecting the appropriate tags.

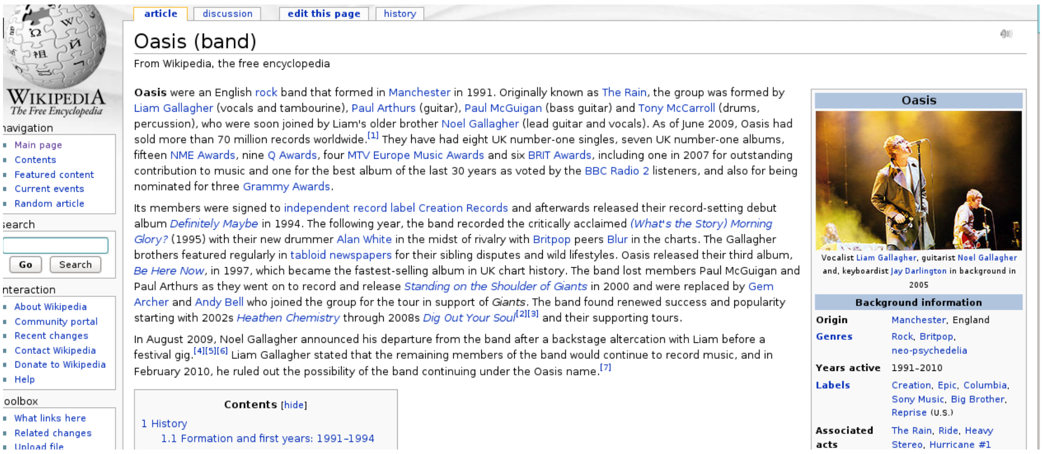
These outstanding features make Wikipedia an exceptional source for acquiring lexical information. This possibility has been used in several ways in recent years, and it will likely be even more exploited in the future, because of its continuous evolution.

The most relevant elements of Wikipedia, considered as a lexical resource, are the following:

Articles The basic entry in Wikipedia is the article, a piece of hypertext describing a concept, which is uniquely identified by its title. These articles are connected by means of links or categories, generating a complex structure (see Figure 2.4).

Links Many concepts mentioned in an article are connected to its corresponding description (Wikipedia article) using a link or a piped link, providing an easy access to related knowledge. The links in Wikipedia typically show a topical

[2] http://wikipedia.org/



Oasis (band)

From Wikipedia, the free encyclopedia

Oasis were an English rock band that formed in Manchester in 1991. Originally known as The Rain, the group was formed by Liam Gallagher (vocals and tambourine), Paul Arthurs (guitar), Paul McGuigan (bass guitar) and Tony McCarroll (drums, percussion), who were soon joined by Liam's older brother Noel Gallagher (lead guitar and vocals). As of June 2009, Oasis had sold more than 70 million records worldwide.[1] They have had eight UK number-one singles, seven UK number-one albums, fifteen NME Awards, nine Q Awards, four MTV Europe Music Awards and six BRIT Awards, including one in 2007 for outstanding contribution to music and one for the best album of the last 30 years as voted by the BBC Radio 2 listeners, and also for being nominated for three Grammy Awards.

Its members were signed to independent record label Creation Records and afterwards released their record-setting debut album *Definitely Maybe* in 1994. The following year, the band recorded the critically acclaimed *(What's the Story) Morning Glory?* (1995) with their new drummer Alan White in the midst of rivalry with Britpop peers Blur in the charts. The Gallagher brothers featured regularly in tabloid newspapers for their sibling disputes and wild lifestyles. Oasis released their third album, *Be Here Now*, in 1997, which became the fastest-selling album in UK chart history. The band lost members Paul McGuigan and Paul Arthurs as they went on to record and release *Standing on the Shoulder of Giants* in 2000 and were replaced by Gem Archer and Andy Bell who joined the group for the tour in support of *Giants*. The band found renewed success and popularity starting with 2002s *Heathen Chemistry* through 2008s *Dig Out Your Soul*[2][3] and their supporting tours.

In August 2009, Noel Gallagher announced his departure from the band after a backstage altercation with Liam before a festival gig.[4][5][6] Liam Gallagher stated that the remaining members of the band would continue to record music, and in February 2010, he ruled out the possibility of the band continuing under the Oasis name.[7]

Contents [hide]

1 History

1.1 Formation and first years: 1991–1994

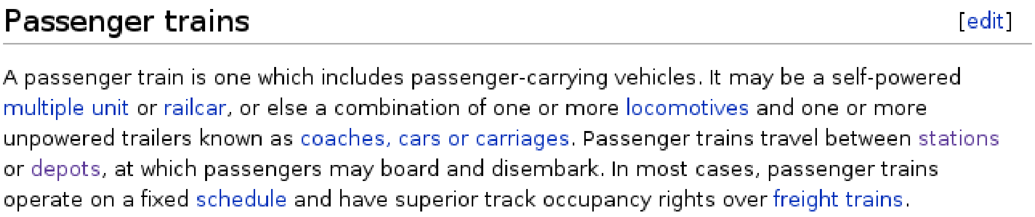
Oasis

Vocalist Liam Gallagher, guitarist Noel Gallagher and, keyboardist Jay Darlington in background in 2005

Background information	
Origin	Manchester, England
Genres	Rock, Britpop, neo-psychedelia
Years active	1991–2010
Labels	Creation, Epic, Columbia, Sony Music, Big Brother, Reprise (U.S.)
Associated acts	The Rain, Ride, Heavy Stereo, Hurricane #1

Figure 2.4: Wikipedia Article for *Oasis (Band)*, partial view.

association between the concepts that are described by the articles. External links are also provided, connecting to Web pages not belonging to Wikipedia (see Figure 2.5).



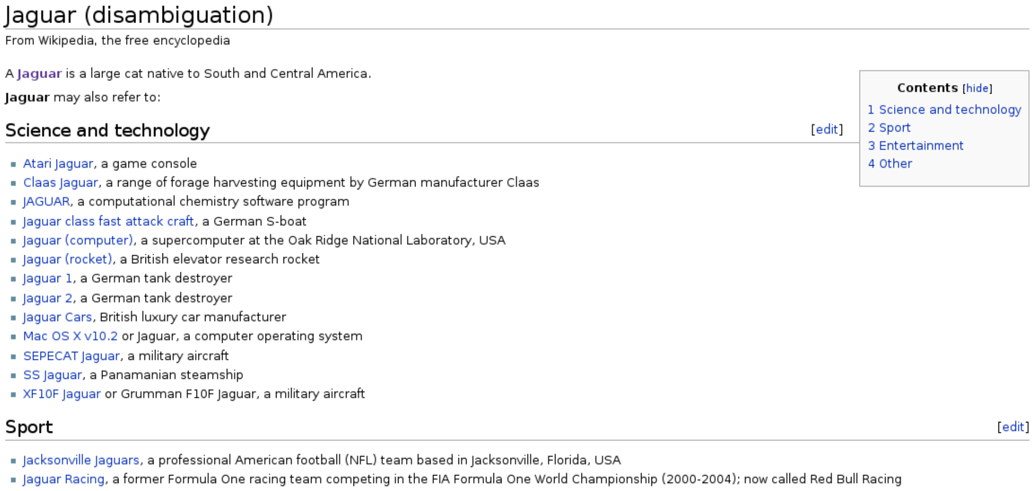
Passenger trains [edit]

A passenger train is one which includes passenger-carrying vehicles. It may be a self-powered multiple unit or railcar, or else a combination of one or more locomotives and one or more unpowered trailers known as coaches, cars or carriages. Passenger trains travel between stations or depots, at which passengers may board and disembark. In most cases, passenger trains operate on a fixed schedule and have superior track occupancy rights over freight trains.

Figure 2.5: Detail of a Wikipedia article. Highlighted words link to their corresponding Wikipedia entry

Redirect Pages Frequently, a Wikipedia article can receive more than a title; thus, a redirect page exists for each alternative name that can be used to refer to a concept in Wikipedia. For such names, the Wikipedia page consists of a link to the article that actually contains the description of the concept. These redirected titles can be seen as synonyms.

Disambiguation Pages An interesting structure is formed by the so-called disambiguation pages, specially created for ambiguous words. Such pages offer a list of possible meanings of the required word (typically a noun), with links to the articles describing these meanings. Disambiguation pages do not have a prescribed structure, and present the different options in an unordered or semi-ordered list (see Figure 2.6). The English Wikipedia currently contains more than 100,000 disambiguation pages.



Jaguar (disambiguation)

From Wikipedia, the free encyclopedia

A **Jaguar** is a large cat native to South and Central America.

Jaguar may also refer to:

Contents [hide]

1 Science and technology
2 Sport
3 Entertainment
4 Other

Science and technology [edit]

- Atari Jaguar, a game console
- Claas Jaguar, a range of forage harvesting equipment by German manufacturer Claas
- JAGUAR, a computational chemistry software program
- Jaguar class fast attack craft, a German S-boat
- Jaguar (computer), a supercomputer at the Oak Ridge National Laboratory, USA
- Jaguar (rocket), a British elevator research rocket
- Jaguar 1, a German tank destroyer
- Jaguar 2, a German tank destroyer
- Jaguar Cars, British luxury car manufacturer
- Mac OS X v10.2 or Jaguar, a computer operating system
- SEPECAT Jaguar, a military aircraft
- SS Jaguar, a Panamanian steamship
- XF10F Jaguar or Grumman F10F Jaguar, a military aircraft

Sport [edit]

- Jacksonville Jaguars, a professional American football (NFL) team based in Jacksonville, Florida, USA
- Jaguar Racing, a former Formula One racing team competing in the FIA Formula One World Championship (2000-2004); now called Red Bull Racing

Figure 2.6: Disambiguation Page for *jaguar*, partial view.

Categories Categories provide a topical structure for Wikipedia articles, by placing each article in at least one category representing a topic. These categories are inter-related by connecting them to one or more parent categories. Wikipedia categories can not be considered a taxonomy, it is rather a directed, mostly acyclic graph, in which different schemes of categorization for topics coexist. An article can be classified in some of these main types of categories: (i) topic categories (articles relating to a particular topic), (ii) list categories (articles on subjects in a particular class) (iii) list-and-topic categories (combinations of the two above types), (iv) intermediate categories (used to organize large classes of subcategories, such as *Category:Albums by artist*), (v) universal categories (used to provide a complete list of articles which are otherwise normally divided into subcategories), (vi) project categories (used mainly by Wikipedia editors for project management purposes, rather than for browsing), and (vii) stub categories, in which incomplete articles are catalogued. For instance, the article *oasis (band)* falls in the categories presented in Figure 2.7.

This complex structure makes inadequate to consider all the categories as full meaning topical descriptors (as we do with ODP categories), since we do not have a list of successive specifications of a general concept, but a list of classifiers, some of them good descriptors, and some others which do not define accurately the target sense. For instance, *Ivor Novello Award winners* and *Musical Quintets* do not provide the same value as categories for the

Figure 2.7: Wikipedia Categories for *Oasis (band)*

.

```
Musical groups established in 1991, 1990s music groups,
2000s music groups, 2010s music groups,
Musical groups disestablished in 2010,
Musical groups from Manchester, Musical quintets,
English rock music groups, Britpop musical groups,
Creation Records artists, BRIT Award winners,
Ivor Novello Award winners,
MTV Europe Music Awards winners, Oasis (band),
```

Oasis (Band) but they are in the same level of parenthood with it. Currently, there are near 400,000 categories in the English Wikipedia.

Multilingual Links A number of concepts are described in different languages through multilingual Wikipedia. The full contents of multilingual articles may or may not be direct translations of a given article. Nevertheless, different pages describing independently the same concept should be closely related. Another interesting feature is that for a given article, translations of its title into other languages for which the underlying concept is described, are offered as hyperlinks.

2.1.3 Other Collaboratively Authored Resources

Other high-quality collaboratively authored resources in the Web are essentially domain specific. A salient example is the Internet Movie Database (IMDb)[3], an online collaboratively generated database, and one of the largest repositories of information related to movies, actors, television shows, production crew personnel, video games, and even fictional characters featured in visual entertainment media. The database includes filmographies for all people identified in listed titles, providing detailed biographies, photographies, and also plot summaries, memorable quotes, awards, reviews, box office performance, filming locations, technical specs, promotional content, trivia, and links to official and other websites, among other information.

The vast amount of specialized information existing in IMDb could be exploited as a lexical resource for NLP tasks involving named entities in the specific domain of entertainment, as well as for multilingual research, considering the multilingual

[3] http://www.imdb.com

versions of IMDb. This and other collaboratively authored resources as allmusic[4] are very valuable as knowledge databases, but with less potential than ODP or Wikipedia, as their current coverage is clearly lower, and they are devoted to specific-domain contents.

2.1.4 Resources Authored by Aggregation

Knowledge is implicit in these resources, and some mining is required to extract it. Although the contents authored by aggregation have potential as sources of lexical information, we have decided to focus on collaboratively edited contents as cleaner sources for our goals. For completeness, we briefly summarize below how aggregated sources have been used for Natural Language Processing.

Social Bookmarks Services

Semantic Web is an approach in which Web resources are created not only to be used by humans, but also to be understood and processed by machines. In order to make this possible, such resources contain different types of metadata. The most common method to generate metadata is to firstly define an ontology and then use the ontology to add semantic markups for Web resources; This model is known as top-down approach ([Zhang et al., 2006]). The release and popularity of the social bookmarks services as Delicious or Flickr have been the starting point for a new, bottom-up approach to semantic annotation, in which the informal categorization of the folksonomies is used to derive emergent semantics (semantic agreement derived from these large-scale resources).

Delicious

Two major examples of folksonomies are found in Delicious and Flickr. Delicious[5] is a social bookmarking Web service in which Web bookmarks are discovered, stored and shared. Delicious is based on a non-hierarchical classification system in which users can tag each of their bookmarks with freely selected index terms (generating a kind of folksonomy). A combined view of all bookmarks with a given tag is available. Its collective nature makes it possible to access to bookmarks added by similar-minded users.

Flickr

Flickr[6] is an image and video hosting website, Web services suite, and online community. In addition to being a popular website for users to share and embed personal photographs, the service is widely used by bloggers to host images that they embed in blogs and social media. As the hosted images are usually associated to textual information, Flickr can be considered as a source of lexical data.

[4] www.allmusic.com

[5] http://delicious.com/

[6] http://www.flickr.com/

In [Zhang et al., 2006], global semantics are statistically inferred from the folksonomies to semantically annotate the Web resources. The global semantic model also disambiguate the tags and group synonymous tags together.

Another approach to analyze and give more structure to folksonomies consists of discovering knowledge that is already implicitly present in them by the way in which different users assign tags to resources ([Schmitz et al., 2006]). In this work, association rule mining is proposed as a method for projecting a folksonomy onto a two-dimensional structure. This method is evaluated on two selected projections of Delicious, showing promising results that can be used to enrich ontologies and even to find emergent semantics by converging use of the same vocabulary.

In [Schmitz, 2006], a method for inducing an ontology from the Flickr tag vocabulary using a subsumption-based model is introduced. Comparing this work with similar models ([Sanderson, 1999] and [Clough et al., 2005]), the average of relations is not reported in the first one, is of 105 in the second and of more than 1200 in this work. Precision is, respectively, 23%, 15% and 51%.

2.1.5 Comparison with WordNet

Lexical Knowledge Bases such as WordNet are an essential element in NLP research. The manual creation of such resources involves a formidable effort, considering that, in order to be efficient, these resources should have a large coverage and establish relations subtle enough to be used both in open or in specific domains. In fact, the shortage of appropriate knowledge bases has been an impeding factor in the field of NLP, which is known as the knowledge acquisition bottleneck.

WordNet [Miller et al., 1990] is widely recognized as the most used lexical resource in NLP applications. It consists of a large-scale lexical database for the English language, developed by the Cognitive Science Laboratory of Princeton University. In the most recent version up to date (WordNet 3.0), it contains 155,287 words organized in 117,659 synsets (sets of synonym terms) for a total of 206,941 word-sense pairs, distributed in four syntactic categories: nouns, verbs, adjectives and verbs. A synset is usually described with a gloss or definition.

WordNet shares some features with monolingual dictionaries: its glosses and examples are similar to word definitions. However, WordNet offers much richer information than common dictionaries or sense inventories. Based on psycholinguistic premises, one of the WordNet implicit assumptions is that lexicalized concepts can be organized by semantic relations. Such network of semantic relations becomes a key value in order to be useful for NLP tasks.

The design of WordNet is based on *synsets*, or sets of lexicalized expressions associated with the same concept. This sets of synonyms, (equivalent to senses) are linked to each other by means of semantic relations.

The number of relations considered by WordNet is limited. Besides synonymy, which is implicit in the concept of *synset*, there is another relation that plays an essential role in the WordNet structure: hypernymy/hyponymy. This relation gives a hierarchical structure to WordNet synsets. For nouns and verbs, every synset, apart from those located on the top of the ontology has at least one hypernym. In addition, there are other relations such as meronymy and antonymy, which are only applied to a subset of the database.

The hierarchic organization of concepts makes WordNet close to ontologies although, in contrast with inference ontologies, WordNet is mainly focused on lexical knowledge, i.e. it is focused on the representations of lexicalized units related to concepts. That explains the lack of artificial levels, sometimes present in generalistic ontologies to allow certain inferences.

Since its release, WordNet was regarded as a very useful resource, being quickly exploited for various NLP tasks as in [Voorhees, 1993], which represents one of the first attempts of using the WordNet lexical information for IR. In this research, *hoods* formed by hypernyms of given *synsets* were used, as thesauri categories in thesauri approaches. Some authors have proposed methods to calculate semantic distances between senses, by using the WordNet hierarchy ([Agirre et al., 1994], [Sussna, 1993]). In [Agirre and Rigau, 1996] and [Agirre and Rigau, 1997], Word Sense Disambiguation relied on *semantic density*, an extension of the conceptual distance introduced by [Rada et al., 1989] and [Sussna, 1993]. [Fernández-Amorós et al., 2001] enriched and generalized this method by adding new parameters and semantic relations.[Resnik, 1995a] and [Resnik, 1995b] explored the semantic similarity between senses (very close to conceptual distance), throughout WordNet hierarchy. In the following sections, we will see other works for which WordNet plays a major role.

Comparing WordNet with the main collaboratively authored Web contents (ODP and Wikipedia), WordNet should be more consistent, as it has been explicitly designed for NLP purposes. As for their hierarchical structure, the synset hierarchy in WordNet is clearly determined by linguistic criteria, whereas for ODP relations are less strict and depend on the editors' discretional criteria. In Wikipedia, categories form a complex graph, in which articles can be organized in more than a way.

WordNet also provides more linguistic information, which is essential for NLP tasks. Which is then the point of using ODP and Wikipedia? The main reasons are (i) coverage: as we have seen, WordNet contains 155,287 words, 117,659 synsets and 206,941 word-sense pairs whereas ODP is formed by more than 600,000 categories, classifying more than 4,000,000 Web sites, and Wikipedia by more than 3,000,000 articles catalogued in near 400,000 categories (considering only the English Wikipedia); (ii) the different nature of their sense distinctions, developed with an emphasis on usable world knowledge rather than on linguistic principles, which makes them complementary to standard lexical resources; and (iii) their exhaustive updating and maintenance, which makes them a perfect tool to handle the Web as the largest textual repository.

WordNet Extensions

In this Section, we briefly review some of the most relevant extensions built on WordNet. In [Cuadros and Rigau, 2008], a benchmarking of some of these resources is performed in order to study their relative quality.

The MCR

The Multilingual Central Repository (MCR)[7] is a multilingual lexical knowledge base that integrates and distributes all the knowledge acquired in the MEANING project [Rigau et al., 2002]. The aim of this project is to collect and process language information from the Web, in order to create multilingual lexical knowledge bases that could be employed to support open domain word sense disambiguation, and also provide a common conceptual infrastructure, being the MCR its major achievement.

The MCR is based on the model proposed by the EuroWordNet project ([Vossen, 1998]). EuroWordNet is a multilingual lexical database, with wordnets for several European languages, following the structure of the original WordNet. The MCR can be considered as a sense inventory for nouns, verbs, adjectives and adverbs for Basque, Catalan, English, Italian and Spanish languages, which are the five wordnets involved in the MEANING project. These wordnets are associated to an Inter-Lingual-Index (ILI), based on WordNet 1.6, that interlinks similar words from different languages. The current version of MRC ([Atserias et al., 2004]) is completed with more information: it integrates five local wordnets (five versions of the English WordNet), the eXtended WordNet ([Mihalcea and Moldovan, 2001]), which improves WordNet with a semantic annotation of the glosses, an upgraded version of the EuroWordNet Top Ontology ([Álvez et al., 2008]), MultiWordNet domains ([Magnini and Cavaglia, 2000]), the Suggested Upper Merged Ontology (SUMO) ([Niles and Pease, 2001]) and finally large collections of semantic preferences, acquired both from SemCor (using a model that learns selectional preferences for classes of verbs, ([Agirre and Martinez, 2001]) and integrated in WordNet ([Agirre and Martinez, 2002]), and from BNC ([McCarthy, 2001]). Hundreds of thousand of new semantic relations and properties have been automatically acquired by means of a process of porting all this knowledge to the different wordnets.

Other international projects include Balkanet ([Tufis et al., 2004]) or EuroTerm ([Stamou et al., 2002]), among several developments around the Global WordNet Association[8].

[7] http://adimen.si.ehu.es/cgi-bin/wei5/public/wei.consult.perl

[8] http://www.globalwordnet.org/

On the other hand, some WordNet extensions use the Web as a resource of lexical information, as in the enrichment of WordNet with topic signatures that provide domain information and in the association of sense-tagged corpora to WordNet senses. Being close to our research topic, we discuss them in more detail below.

Acquisition of Domain Information for WordNet Senses In its initial versions, WordNet did not include topical or domain information, which is very valuable for sense disambiguation and for many other purposes. One approach to the problem of adding domain information to WordNet senses is presented in [Agirre et al., 2000]. In this work, the Web is considered as a resource to enrich WordNet senses, by associating the so-called topic signatures to them. Topic signatures are vectors consisting of words contextually related to the sense, together with a measure of the strength of the association sense-word for each word. In order to create the topic signature for a given sense of a word, (i) a query is submitted to a search engine (Altavista), formed by extracting from WordNet positive information, (required terms in the query) about the sense, and negative information, (negated terms) also acquired from WordNet, about the rest of senses of the word. (ii) The retrieved documents are processed to mine the words that will be used to create the context, giving more weight to the more frequent words. The words and their weights, in decreasing order of weights, form the topic signature for each word sense. The usefulness of topic signatures is proved for WSD tasks. Other possible application of such signatures is the clustering of WordNet senses, assuming that close senses will have close topic signatures. Two relevant collections of topic signatures are the TSWEB7 ([Agirre and de Lacalle, 2004]), in which all WordNet nominal senses receive a topic signature automatically extracted from the Web, and the TSSEM ([Cuadros et al., 2005]), which have been acquired from the SemCor. In the first case, the topic signatures for the senses were produced by (i) submitting certain monosemic terms, near to each sense in WordNet, as queries to Google, (ii) processing the results and (iii) weighting them with the TFIDF formula. In the second case, topic signatures are produced by by selecting the subcorpus related to each sense (sentences containing the sense) from the Corpus SemCor, and weighting, as in the previous case, with the TFIDF formula.

Acquisition of Sense-Tagged Corpora The automatic acquisition of sense-tagged corpora is the most obvious use of the Web to improve WSD. Such sense-tagged corpora are essential in the performance of supervised WSD algorithms. The following approaches enrich WordNet senses with this information.

In [Mihalcea and Moldovan, 1999], the method of monosemous relatives ([Leacock et al., 1998]) is adapted with the aim of extracting training sentences Web Searching from the Web. In [Agirre and Martínez, 2000], this strategy is performed in order to train a supervised WSD system, with discouraging results. The conclusion was that, although the obtained examples were correct, the acquisition from the Web of training material could present some structural problems such as biased acquisition.

In [Agirre and Martinez, 2004], another Web corpus is built, by improving the monosemous-relative technique with additional filters, focusing on the bias question by making a comparison among several bias options. In this work it is shown that the monosemous relatives tech- nique can be used to extract examples for all nouns in WordNet.

Bootstrapping

[Mihalcea, 2003] enhances the method described in [Mihalcea and Moldovan, 1999] with a bootstrapping approach based on [Yarowsky, 1995]. In this approach, a set of seeds (initial small set of tagged samples) is extracted from SemCor, WordNet, and the Web (using the monosemous relatives technique described in [Mihalcea and Moldovan, 1999]). These samples are used to mine the Web for documents. After a disambiguation process performed by means of the [Mihalcea and Moldovan, 2000] algorithm, new seeds are generated and submitted to the Web for a new search. The sense-tagged corpus generated with this approach (GenCor) was tested in Senseval-2, achieving one of the best scores both in the lexical sample and in the all-words tasks. In [Mihalcea, 2002a], a supervised system obtains comparable results being trained either with the Web corpus generated via the bootstrapping method or with manually-tagged data.

In summary, the acquisition of lexical knowledge could be regarded as one of the main tasks in NLP research, since this information is essential in the performance and evaluation of the majority of systems. However, the process implies a high cost and effort, and the results, as good as they may be, always admit some improvement. One way to alleviate the shortage of lexical resources consists of enriching some existing knowledge bases, and particularly WordNet, which is the most salient one, by extracting relevant information from other available resources. One of the aims of this research is to extract lexical information from the Web, which has become the main resource for mining information, but rather than exploiting the full Web, we have focused on collaborative authored contents (ODP and Wikipedia), because they are cleaner, more reproducible and more reliable about the quality of the stored information.

As it has been mentioned, the MCR provides conceptual information which is used to enrich WordNet. This is the same line that we follow in Chapter 2, in which we develop a method for enriching WordNet senses with domain information extracted from the ODP directories. Indeed, our method of classifying new terms using Web directories (see Chapter 3) was examined as an option of acquiring new senses in the design of experiments for the MEANING project.

2.2 Collaborative Authored Web Contents in NLP Techniques

In this Section, we review the use of collaboratively authored Web contents, and specifically ODP and Wikipedia for general NLP applications. We will show that the use of these resources (and more precisely of Wikipedia) is a state-of-the-art tendency, widely applied in multiple NLP tasks. Although ODP has received less attention than Wikipedia, it is still a very interesting resource that can be even more useful than Wikipedia, depending on the considered task. Therefore, we have focused on both resources in our research.

2.2.1 Cross-Language Alignment

Multilingual Wikipedia is probably the best existing resource for obtaining parallel and comparable corpora, as it covers more than 250 languages and entries for the same topic in different languages are closely related or even are direct translation in some cases.

[Adafre and de Rijke, 2006] proposes two methods for identifying similar sentences written in Dutch and English. In the first approach, a Dutch Wikipedia page is translated into English using an online MT system (Babelfish of Altavista). The text pairs are produced by splitting both texts into sentences or chunks, linking the corresponding chunks to form pairs, computing the similarity measure of the pairs and filtering the results. In the second method, a bilingual lexicon is generated from Wikipedia using the link structure; then, each sentence is represented in both languages by the set of hyperlinks it contains. Finally, similarity measures are computed and results are filtered. These two methods are evaluated in 30 Wikipedia articles randomly selected. On average, the MT based approach returns 26% correct sentences and the bilingual lexicon based approach returns 45% and, on average, the MT approach has three times more coverage than the bilingual lexicon approach. Adafre

[Yasuda and Sumita, 2008] describes a bootstrapping method based on Wikipedia multilingual articles in which statistical machine translation (SMT) and a sentence-aligned corpus between Japanese and English are generated. To align the Japanese and English sentences, first the Wikipedia articles in Japanese are translated into English by using a MT system. Then, sentence similarity is calculated by using a MT evaluation metric. Finally, sentences are aligned by using the similarities calculated in the previous step. These results are used to train the system, in a bootstrapping process. The evaluation results show that 10% of Japanese sentences are correctly translated, with high alignment quality. Yasuda

[Ye et al., 2009] describes a method for building a multilingual association Ye dictionary by associating multilingual words and concepts together in a graph. The approach consists of exploiting the Wikipedia links to associate multilingual words and concepts together and generate the graphs. Experimental results show that using the multilingual association dictionary to conduct the tasks of filtering and expanding in the English-Chinese CLIR experiments offers a better retrieval performance than using the LDC EC2.0 dictionary.

As a general conclusion, Wikipedia has proved to be a useful resource for acquiring comparable and parallel corpora, not only at a word level with the generation of dictionaries, but also at the sentence and even document levels, by the alignment of multilingual Wikipedia articles related to the same concept. The main weakness observed in these works is the lack of coverage, which will likely be improved as the multilingual branches of Wikipedia grow.

2.2.2 Measurements of Semantic Relatedness

Wikipedia has also been used in computing semantic relatedness methods. Articles, categories and hyperlinks have been exploited in different approaches, showing the potential of Wikipedia as a resource of lexical knowledge.

Gabrilovich [Gabrilovich and Markovitch, 2007] proposes the Explicit Semantic Analysis (ESA), a method for representing texts in a vectorial space of concepts extracted from Wikipedia, and measuring the semantic relatedness of two given texts. The meaning of the text is represented as a weighted vector, and relatedness of words in the first place and relatedness of texts in a second stage are computed. The method is evaluated both on text categorization and on measuring the semantic relatedness between texts with significant reported improvement over previous work. The same method is performed using ODP with worse results.

Strube In [Strube and Ponzetto, 2006], the Wikipedia folksonomy is used for computing semantic relatedness measures. In order to measure the relatedness of a pair of words, the system starts by retrieving the Wikipedia pages to which the words refer, then extracts the categories to which the pages belong, and given the set of paths found between the category pairs, measures are computed by selecting the shortest path for path-based measures and the path which maximizes information content for information content based measures. The relatedness measures are evaluated on three datasets: [Miller and Charles, 1991] list of 30 noun pairs (M&C), the 65 word synonymy list from [Rubenstein and Goodenough, 1965] of which M&C is a subset, and finally the Word Similarity-353 Test Collection [Finkelstein et al., 2002], comparing performances of Wikipedia and WordNet. The authors found that both perform better than a Google baseline, WordNet performs better for small datasets, and Wikipedia outperforms WordNet on 353-TC. Additionally, the authors apply the measures of relatedness to a representative NLP task: an extension of a machine learning based co-reference resolver. This system uses relatedness scores as features for classifying referring expressions as denoting the same discourse entities. Mining the WordNet taxonomy and the Wikipedia encyclopedic knowledge base, as well as including semantic parsing information, they induce semantic features for co-reference learning [Ponzetto and Strube, 2006]. The results indicate that both Wikipedia and WordNet provide semantically relevant features for co-reference resolution. Indeed, the optimal system configurations always include features from both WordNet and Wikipedia, suggesting that they work as complementary knowledge sources. This result is of particular interest for our work, as it is one of the few examples of direct comparison between Wikipedia and WordNet, and the conclusion is that both sources can be considered as complementary.

The Wikipedia Link Vector Model or WLVM ([Milne, 2007]) measures semantic relatedness by using only the hyperlink structure of Wikipedia, not considering the full textual content. This technique has a low processing cost, but accuracy is also unsatisfactory. [Witten and Milne, 2008] discusses an improved method (WLM) with much better results. Comparing the measures of semantic relatedness in correlation with human judgements and considering the WordSimilarity-353 collection as test set, the respective accuracies are 0.35 for WordNet and, on the other hand, 0.49 for Wikirelate!, 0.69 for WLM and 0.75 for ESA, all three performing with Wikipedia. A conclusion from these results is that Wikipedia is a better resource than WordNet for semantic relatedness measuring. Milne

In summary, Wikipedia has been widely and successfully used to compute semantic relatedness. It should be noted that the compared performance of WordNet and Wikipedia has been also studied, and it is shown that Wikipedia and WordNet can be regarded as complementary resources for this task, although in isolation Wikipedia gives better results.

2.2.3 Clustering and Classification

Web search results clustering is a topic that receives an increasing attention from the research community. In this Section, we focus on the clustering and classifying research performed by using collaboratively authored contents, a topic close to our work, as in Chapter 4 we investigate how an inventory of senses extracted from Wikipedia can help providing a cleaner clustering of search results by classifying each results page as belonging to one (or more) senses of the query term.

In the context of clustering, [Carmel et al., 2009] employs Wikipedia to enhance automatic cluster labeling. In this work, Wikipedia is used to improve cluster labeling by extracting candidate labels from it. As starting point, the documents, represented as weighted-term vectors, are indexed by generating a search index, also providing useful statistic values. Then, important terms are extracted by calculating the set of terms that maximizes the Jensen-Shannon Divergence (JSD) distance between the cluster C and the entire collection, and finally, candidate Carmel labels are extracted by two different strategies: (i) extracting labels directly from the clustered documents content, and (ii) extracting labels from Wikipedia by generating a search index from a Wikipedia dump and executing a query to the Wikipedia index, based on the list of important terms. The result of this query is a list of documents sorted by their similarity score to the query. For each document, both its title and the set of categories associated with the document are considered as candidate cluster labels. Finally, candidate labels are evaluated by several automatic heuristics. The scores of all heuristics are then aggregated and the labels with the highest aggregated scores are returned. The clustering algorithm is not predetermined in this system, being the impact of the clusters coherence one of the aspects evaluated in this paper.

The data sets selected for evaluating the system are the 20 News Groups (20NG) data collection, and a set made by randomly selecting 100 different categories from the ODP hierarchy, and then, for each category, randomly selecting up to 100 documents, resulting in a collection size of about 10,000 documents. In both collections, the categories were manually labeled for evaluation purposes. As evaluation criterion, a proposed label for a given cluster is considered correct if it is identical, an inflection, or a WordNet synonym of the cluster's correct label. Given a collection of clusters, and the parameter k that indicates the number of required cluster labels, the system proposes up to k labels for each cluster. The feature selection method, the number of important terms for querying Wikipedia, the number of Wikipedia results to be used for label extraction, and the heuristic used for candidate evaluation are considered as system parameters. Then, for each possible configuration of such parameters, system performance is evaluated by using two different measures.

The aspects evaluated are: the effectiveness of using Wikipedia to enhance cluster labeling, the effect of the number of important terms that are used to query Wikipedia, the number of top scored results from which candidate labels are extracted, the effect of applying different heuristics for evaluating and the effect of the clusters coherency on the labeling process. The results show that the Wikipedia labels agree with manual labels associated by humans to a cluster, much more than with significant terms that are extracted directly from the text, and that for more than 85% of the clusters in the test collection, the manual label (or an inflection, or a synonym of it) appears in the top five recommended labels.

Banerjee proposes a method in which short texts representations are enriched with additional elements extracted from Wikipedia, using such enhanced representation to improve the clustering of these short texts. The authors extract a labeled dataset from Google News and represent the texts in two different ways: in the first one, each article is represented by a vector of weighted terms appearing in the article and in the title, giving more importance to the title. In the second, the term frequency vector of the above method is enhanced by using the title and the snippet of the news article as two separate queries to Wikipedia, retrieving the more relevant Wikipedia articles, and using this information to complete the vector. Comparing these two representations in six different clustering algorithms, the representation using Wikipedia obtains better results, reaching 89.6% accuracy for one of the clustering methods.

[Gabrilovich and Markovitch, 2005] propose a method for text categorization, Gabrilovich
based on large repositories of knowledge. In this work, the knowledge base is ODP, using not only the hierarchy, but also the contents of the URLs stored in the categories, in order to extract the knowledge. Based on this knowledge, a feature generator builds new features that enrich ODP. The resulting space of features is then used to classify texts. Using support vector machines as learning algorithm to build text categorizers, and precision-recall Break-Even Point (BEP) to measure text categorization performance, the authors report a baseline of 87.7% (Micro BEP) and 60.2% (Macro BEP) being their results 88.0% and 61.4% respectively. Another interesting fact is that during the feature generation, a contextual analysis is performed, producing implicit word sense disambiguation.

provide evidence suggesting that Wikipedia articles and the category and article Syed
link graphs can successfully describe common concepts in a set of documents. This information is useful in the annotation and categorization of documents. In this work, Wikipedia is directly employed to predict concepts that characterize a set of documents. The authors see this task related to text categorization (but not identical, as there may exist documents with concepts in common, but not belonging to the same category), and also similar to computing semantic relatedness between concepts (but focusing on predicting concepts in common between documents). In this research, three different methods are implemented: (i) Article Text: the test document or set of related documents are used as search query to a Lucene Wikipedia index. After retrieving top N matching Wikipedia articles (based on cosine similarity) for each document in the set, their Wikipedia categories are extracted and scored with two different scoring schemes; (ii) Text and Categories with Spreading Activation: the Wikipedia category links network is also used for prediction of related concepts. The top N Wikipedia categories predicted as a result of method one and scoring scheme one work as the initial set of activated nodes in the category links graph and (iii) Text and Links with Spreading Activation: the top N matching Wikipedia articles (based on the second score) to each test document are considered as the initial set of activated nodes in the article links graph. The evaluation results show that precision, average precision and recall improve at higher average similarity thresholds. A comparison of the different methods using the F-measure metric shows that the method using spreading activation with two pulses (SA2) almost always performs better than other methods at different average similarity thresholds for predicting categories or super-concepts of the test documents, whereas including the article links information is useful for predicting

more specialized concepts.

These results show that Wikipedia is also useful to describe common concepts for a set of documents.

2.2.4 Question Answering

The use of collaborative authored resources, and especially Wikipedia as a resource and as a target for question answering systems is a growing line of research; there has even been a QA task directly defined for the online encyclopedia in CLEF[9]. In CLEF 2009, GikiCLEF was one of the exercises proposed in one of the eight main tracks (the Multilingual Question Answering (QA@CLEF)). The task was focused on open list questions over Wikipedia that required geographic reasoning, complex information extraction, and cross-lingual processing, for collections in Bulgarian, Dutch, English, German, Italian, Norwegian, Portuguese, Romanian and Spanish.

Ahn As for generic QA, [Ahn et al., 2004] identifies the question's topic and finds the corresponding article in Wikipedia. Then, the expected class of the answer is detected and finally, matching answers are located by analyzing the article. This work describes the participation in the TREC 2004 Question Answering track. In this participation, Wikipedia is used both as a source in various stages of a previous system, with poor results, and to identify facts that are potentially important for the user, comparing facts extracted from a target collection with the information from Wikipedia. The latter shows substantial improvements over the baseline.

Buscaldi In [Buscaldi and Rosso, 2006], Wikipedia is not used as a resource where to find answers, but it is rather employed for providing validations to the answers given by a previous system. Also, the categories of the Spanish Wikipedia are exploited to determine a set of appropriate patterns for the expected answer, showing that Wikipedia is indeed a useful resource for question answering tasks. The results reach an improvement of 4.5% in the recall gain, for the all type of questions group. [Kaisser, 2008] presents the online demo of the QuALiM Question Answering system, in which answers are supplemented with relevant passages from Wikipedia. The possible answers for a query, obtained from search engines are supported by information extracted from Wikipedia, improving the search results.

Xu [Xu et al., 2009] analyzes the use of Wikipedia in pseudo-relevance feedback for query expansion. In this work, pseudo-relevance information is generated using Wikipedia, and three options for query expansion, in which this information is exploited, are proposed and compared. The evaluation on four TREC test collections shows that retrieval performance of three types of queries (entity queries, ambiguous queries and broader queries), can be improved. Indeed, the proposed method outperforms the baseline relevance model in terms of precision and robustness.

[9] www.clef-campaign.org

[Li et al., 2007] suggest that standard pseudo-relevance feedback might not be useful for short queries, and propose a method to expand the queries by using Wikipedia as a external corpus. Query expansion starts by obtaining Wikipedia articles related to the query, and clustering the documents by using the categories information associated to them. Then, a rank is established, according to which, the most populated clusters become more represented. Finally, some terms of the best ranked articles are selected in order to expand the query. Although the results are not very encouraging, the authors see room for substantial improvement. This work is closely related to our work with Wikipedia; nevertheless, the method is not exploited to enhance search results diversity, but is rather focused on query expansion. Li

Overall, Wikipedia has been used not only as a direct resource for finding answers, but also - with even better results- as a resource for providing validation or enrichment of the answers and for query expansion, showing that it is very valuable for question answering tasks.

2.2.5 Summarization

The unlimited amount of online information suggests that some form of automatic summarization is essential in order to make it manageable.

[Nelken and Yamangil, 2008] propose a method for using the Wikipedia information to bootstrap text summarization systems. This information is the Wikipedia article revision history, which consists of the iteratively generated refinements of a Wikipedia article, available in the successive available snapshots of the encyclopedia. By comparing different versions of the same document, and repeating the process over a large collection of documents, the authors collect users' editorial choices. These data are then employed for three different tasks: (i) automated text correction (lexical level), (ii) sentence compression (sentence level) and (iii) assuming that the temporal persistence of a sentence throughout the revision history is a good indicator of its importance, the authors use Wikipedia revision data for training text summarization systems (document level). The compression rates for the Knight and Marcu method (KM), manually produced summaries and the system are, respectively, 72.91%, 53.33% and 67.38%, which represents an increase in compression over KM. The grammaticality rate is also better and the importance (the value of the retained information) decreases slightly. The method is also evaluated on the summarization of two Wikipedia articles, with promising results. Nelken

[Nastase, 2008] uses Wikipedia for topic expansion: hyperlinks in Wikipedia articles are employed to expand key words and key phrases extracted from a query, for summarization purposes. This method is applied on a large graph that covers the entire document collection for one topic. The nodes of this graph are words or named entities in the texts and the grammatical relations (edges) are links. The Nastase procedure consists of: (i) To find words/NEs related to the topic, an activation signal starting from the topic words and their expansions is spread. Starting from these nodes, the signal is propagated by assigning a weight to each edge and each node traversed, based on the signal strength (higher weights are closer to topic). Once the graph is initialized in this way, a PageRank algorithm is applied to boost the top ranked nodes; (ii) from this graph, the subgraph that covers connections between all open class words/NEs in the topic or expanded topic query is extracted. Each edge in the extracted subgraph corresponds to a grammatical relation in a sentence; Therefore, all sentences represented in the subgraph are collected and re-ranked; (iii) in order to form the summary, the ranked list of sentences (starting with the highest rank) is traversed, adding sentences to the summary. A comparison among the summaries produced with no topic expansion, WordNet expansion and Wikipedia expansion, respectively, shows that expanding a topic only with Wikipedia hyperlinks gives the best results. An interesting fact is that combining both Wikipedia and WordNet expansions does not improve performance.

In summary, Wikipedia has also been successfully used for summarization purposes, and it has found to be superior than WordNet rather than complementary. It is interesting to note that in [Nelken and Yamangil, 2008], a feature of Wikipedia is newly exploited: the successive refinements of an article made by contributors.

2.3 Applications of Wikipedia to Word Sense Disambiguation and Discovery

In this Section, we focus on the research most closely related to our work: the use of Wikipedia for Word Sense Disambiguation and Discovery.

2.3.1 Acquisition of Lexical Information

Wikipedia has been widely employed as a source of lexical information. We will distinguish three lines of research: acquisition of textual material, acquisition of relations, and creation of taxonomies.

Acquisition of textual material

[Ruiz-Casado et al., 2005] enrich ontologies with encyclopedic knowledge, by associating entries of English Wikipedia with concepts of WordNet reporting an accuracy in disambiguating the sense of the Wikipedia entries of 83.89%. The enrichment of WordNet senses with lexical material is one of our main aims in this thesis, but we have focused on ODP. As we show in Chapter 4, considering only the Wikipedia senses that can be mapped to WordNet is a limited perspective.

[Gabay et al., 2008] describes a method to generate a partially tagged corpus using Wikipedia hyperlinks. This corpus contains information about the correct segmentation of 523,599 non-consecutive words in 363,090 sentences and is used to create a corpus of Modern Hebrew.

Acquisition of Relations

[Nastase and Strube, 2008] propose decoding the information present in the Wikipedia folksonomy by extracting instances of relations, relation types and class attributes, to finally propagate this acquired knowledge throughout the category network. A evaluation of the resulting relations is performed against other source (ResearchCyc), with poor results, because of the small number of relations present in both sources; A second evaluation, performed manually over four sets of 250 relations, results in precision ranging from 84% to 98%, depending on the considered relation.

[Yan et al., 2009] shows a method for extracting relations by combining dependency patterns from dependency analysis of texts in Wikipedia, and patterns generated from redundant information related to the Web. Given a set of Wikipedia articles, the method outputs a list of concept pairs for each article with a relation label assigned to each concept pair, in four main steps: (i) Wikipedia articles are preprocessed to obtain concept pairs, each of which with an associated sentence; (ii) for each concept pair, context information is retrieved from the Web, and ranked relational terms and surface patterns are generated; (iii) for each concept pair, dependency patterns from corresponding sentences in Wikipedia articles are generated as well; (iv) concept pairs are clustered according to their context.

In an evaluation on two selected categories of Wikipedia, the best results arise from a combination of dependency patterns and surface patterns; Precision and coverage are 75.63 and 23.94% for the first category and 76.87 and 19.61% for the second one. The main weakness of this method is, therefore, the short coverage.

Creation of Taxonomies

[Ponzetto and Strube, 2007] derive from the Wikipedia folksonomy a large scale taxonomy. As evaluation, the amount of is-a relations correctly extracted by comparing with ResearchCyc is computed. The average F-measure is 88%, with a baseline of 85%. Also, by means of the is-a relations produced, it is possible to compute the semantic similarity in this work (previously, in [Ponzetto and Strube, 2006], the use of the Wikipedia categorization as a conceptual network to compute the semantic relatedness of words is proposed, but semantic similarity can not be addressed because of the lack of is-a relations). In [Kassner et al., 2008], a process of acquiring a large, domain independent taxonomy for German, and adaptable to other languages is described. This work is based on the taxonomy generated in [Ponzetto and Strube, 2007].

[Milne et al., 2006] builds a domain-specific thesaurus (for agriculture), by automatically mining Wikipedia, with very good coverage of concepts and semantic relations, in comparison with a classic thesaurus. This work could be classified both in the first and the second group as well.

2.3.2 Acquisition of Sense-Tagged Corpora

Mihalcea

In the field of Natural Language Processing, there have been successful attempts to connect Wikipedia entries to WordNet senses: [Ruiz-Casado et al., 2005] reports an algorithm that provides an accuracy of 84%. [Mihalcea, 2007] uses internal Wikipedia hyperlinks to derive sense-tagged examples. But instead of using Wikipedia directly as sense inventory, Mihalcea then manually maps Wikipedia senses into WordNet senses (claiming that, at the time of writing the paper, Wikipedia does not consistently report ambiguity in disambiguation pages) and shows that a WSD system based on acquired sense-tagged examples reaches an accuracy well beyond an (informed) most frequent sense heuristic. The sense tagged corpus is produced in these steps:

1. Extracting links. All the paragraphs in Wikipedia that contain an occurrence of the ambiguous word as part of a link or a piped link are extracted, explicitly avoiding named entities by considering only those word occurrences that are spelled with a lower case.

2. Collecting labels. All the possible labels for the given ambiguous word are collected, by extracting the leftmost component of the links.

3. Labeling WordNet senses. The labels are manually mapped to their corresponding WordNet sense, creating a sense tagged corpus. This step is performed independently by two annotators, to ensure its correctness.

Once the sense-annotated examples are generated, a WSD system is implemented in three stages:

1. Preprocessing step, in which the text is tokenized and annotated with part-of-speech tags and collocations are identified

2. Extraction of features from the context considering a local context, in a similar way as proposed in [Lee and Ng, 2002].

3. Classifying step: the features are integrated in a Naive Bayes classifier, which was selected for its performance in previous work [Lee and Ng, 2002].

To evaluate the usefulness of the sense annotations acquired from Wikipedia, a word sense disambiguation experiment on 30 of the ambiguous words used during the Senseval-2 and Senseval-3 evaluations is performed, focusing on nouns. Note that this is the set of nouns used in our experiments in Chapter 4.

Two baselines are considered: selecting the most frequent sense, and a baseline that implements the corpus-based version of the Lesk algorithm ([Kilgarriff and Rosenzweig, 2000]). The results, using ten-fold cross-validation are 72.58% for the first baseline, 78.02% for the second and 84.65% for the WSD system. An analysis of the learning rate with respect to the amount of available data, is performed, by applying ten fold cross-validation using 10%, 20%, ..., 100% of the data and averaging the results. It is shown a continuously growing accuracy with increasingly larger amounts of data.

Finally, sense coverage is studied, by measuring the correlation between the relative sense frequencies of all the words in both Wikipedia and Senseval datasets. Using the Pearson (r) correlation factor, an overall correlation of r = 0.51 between the sense distributions in the Wikipedia corpus and the Senseval corpus is found, indicating a medium correlation what suggests that it would be interesting to propose new sense inventories.

In both works, [Ruiz-Casado et al., 2005] and [Mihalcea, 2007], the relation between WordNet and Wikipedia senses is studied, but in the first one a close connection is established, whereas in the second a clear difference is reported. These results suggest the interest of further studying such topic.

Wikify

Along the same line, the Wikify system ([Mihalcea and Csomai, 2007]) identifies important concepts in given documents, linking them to the corresponding Wikipedia pages and then enriching online documents with references to semantically related information. Given a text or hypertext document, text wikification consists of simulating the production of hyperlinks in Wikipedia, by recognizing the important words in the text and linking them to an appropriate Wikipedia article. This process implies two different tasks: keyword extraction, and link disambiguation. The authors propose a system in which (i) the hypertext is converted into plain text, (ii) the keywords are identified in a keyword extraction module, (iii) two WSD algorithms representing opposite strategies (knowledge-based and statistical) are implemented, using the annotations (links in a the corresponding Wikipedia page) for the keywords in order to resolve link ambiguities and to enrich the annotations with the corresponding Wikipedia article and (iv) the hypertext enriched with the new links is re-generated. An evaluation is performed for both tasks, keyword extraction and WSD, for a set of 85 Wikipedia pages containing 7,286 linked concepts. The final disambiguation proves to be competitive with other state-of-the-art WSD results.

In this work, Wikipedia is considered as a sense inventory, but there is no comparison with other inventories such as WordNet, and it is only used for disambiguation purposes. In our research, however, we analyze Wikipedia senses from different perspectives. The Wikify system has been exploited, among others, for educational purposes ([Csomai and Mihalcea, 2007]).

In [Medelyan et al., 2008] Wikipedia is successfully used for topical indexing. The strategy used is similar to the one presented in [Mihalcea and Csomai, 2007]. Compared to our research, the most relevant difference is that the disambiguation approach relies on both the relatedness to context and the centrality of each sense, measured by the relative frequency of a sense being used as a link. This frequency is an internal measure of the relevance of a sense, similarly to what we do in (see Section 4.5).

2.3.3 Recognition and Semantic Disambiguation of Named Entities

In [Mihalcea, 2007], proper names are avoided, but they are the focus of interest of Cuzerzan and Bunescu, which present alternative approaches to the problem of recognition and disambiguation of named entities.

Cucerzan Named entities are mentioned in texts in an ambiguous way (a *surface form*). [Cucerzan, 2007] presents a large-scale system for the recognition and semantic disambiguation of named entities. We summarize the main features of the proposed method, by describing first the strategies applied to collect the required information, second the disambiguation process and finally the evaluation of the system.

In a previous stage, the surface forms and their associated named entities together with tags and contextual information are collected, by applying several strategies:

1. The entity surface forms are extracted from the titles of entity pages, the titles of redirecting pages, the disambiguation pages, and the references to entity pages in other Wikipedia articles.

2. For each named entity associated with a surface form, category tags are extracted from the *List pages* and the category structure of Wikipedia. Also, contextual data for the entities are extracted from both the information present in its Wikipedia page and in the other articles that explicitly refer to that entity

In order to disambiguate surface nouns in a document, first the document is processed: by using capitalization rules and statistics, the named entities and their boundaries are identified. Then, the named entities that correspond to the same surface form are assigned, by means of probabilistic methods, to one of the four labels: Person, Location, Organization, and Miscellaneous.

The contextual and category information extracted from Wikipedia is used to disambiguate the entities in the text. The disambiguation of a surface form process employs a vector space model, in which a vectorial representation of the processed document which aggregates the Wikipedia contexts that occur in the document and its category tags is compared with the vectorial representations of the Wikipedia candidate entities, also formed by categories and contexts. The disambiguation is performed by selecting the assignment of entities to surface forms that maximize the similarity between the document vector and the entity vectors. Finally, hyperlinks to the appropriate pages in Wikipedia are created.

The system is evaluated on a set of 350 Wikipedia articles and on a set of 20 news stories, with a disambiguation baseline implemented by taking the entity page or redirect page titled exactly as the surface form when possible, or, otherwise, the Wikipedia most frequently mentioned entity associated to such form. For the articles, the evaluation consists of comparing the hyperlinks assigned by the system with the links provided by Wikipedia contributors. After discarding exceptions, the accuracy is 86.2% for the baseline system and 88.3% for the proposed system. The difference in accuracy between the two systems is significant at $p = 0.01$, if only actually ambiguous surface forms are considered.

For the news stories, the disambiguation is defined as correct when the assigned Wikipedia article is the best possible. The proposed system obtains an accuracy of 91.4%, versus a 51.7% baseline (significant at $p = 0.01$), which is a remarkable difference.

In this work, Wikipedia is used as a source of lexical information useful to disambiguate named entities, which is very related to our research. The accuracy of the system is not directly comparable to ours, because (i) this work is focused only on named entities, (ii) the sets used for evaluation are different in both cases, and we evaluate on a more unrestricted set (Web search results).

Bunescu

In [Bunescu and Pasca, 2006], another method for detecting and disambiguating named entities is presented. The authors propose to organize all named entities from Wikipedia into a dictionary structure where each string entry is mapped to the set of entities that can be represented by this entry in Wikipedia. To reach this goal, first, the named entities are detected by applying some heuristics steps and, second, the dictionary is constructed as follows: for each named entity, its title name, its redirect names and its disambiguation names are all incorporated as entries in the dictionary and then, each entry string is mapped to the set of entities that may denote in Wikipedia.

In the next stage, using the proposed dictionary and the hyperlinks from Wikipedia articles, a dataset of disambiguated occurrences of proper names is generated. In fact, each link contains the title name of an entity, and also the proper name used to refer to it. Using this information, if an occurrence of a proper name inside a Wikipedia article is related to more than one entity by means of the dictionary, then for each named entity, a new element is added to the disambiguation dataset, consisting of the entity title, a context of words extracted from the corresponding article for the entity

Once the dictionary and the disambiguation dataset are generated, follows the disambiguation process: documents are represented in a vector space; for that purpose, a vocabulary V is created by reading all Wikipedia articles and recording, for each word stem w, its document frequency in Wikipedia. Then, each Wikipedia article is represented as a vector with a weighted component for each word in V. The named entity corresponding to a proper name will be the one with the highest score, given a certain scoring function on proper names and named entities. Thus, different disambiguation methods will depend on the definition of the scoring function. The authors implement two options in order to disambiguate proper names: (i) a function based on the cosine similarity between the context of the entities and the text of the article containing the proper name and (ii) a taxonomy kernel generated by using the information contained in the disambiguation dataset and trained with a support vector machine.

These two methods are evaluated under four different scenarios. The reported results show that the taxonomy kernel highly outperforms the cosine similarity in the first three scenarios, with no significant improvement in the last one.

The motivation of this work is to improve the effectiveness of search engines when the query is ambiguous, by grouping search results according to the corresponding sense. As we will see in Chapter 4, this way of dealing with ambiguity is alternative to our proposal of promoting diversity, and thus, strongly related to our research, although the work presented in [Bunescu and Pasca, 2006] is focused on named entities, whereas we study all kind of nouns.

2.4 Conclusions

As we have shown, the usefulness of collaboratively authored Web contents for NLP, and in particular of Wikipedia, has been broadly proved, making a strong case for its use in our research. Although ODP has received less attention, its potential as a source of domain information is obvious, and therefore we attempt to link it with WordNet in Chapter 3, and then exploit the information implicit in the directory/word sense associations obtained.

On the other hand, Wikipedia has captured the interest of the research community, considering the amount of research in which it has been successfully used in recent years. Nevertheless, to the best of our knowledge, Wikipedia has never been applied (and, at the same time, compared with WordNet) as a sense inventory to improve Web Search, which is the topic that we address in Chapter 4.

Chapter 3

Automatic Association of Web Directories to WordNet Senses

3.1 Introduction

As we have seen in Chapter 2, extracting lexical information from the Web is a very productive line of research and also an attractive possibility for acquiring lexical information and corpora. But a common problem to Web applications is how to detect and filter out all the noisy material, and how to characterize the rest [Kilgarriff, 2001c]. Our starting hypothesis is that Web directories (e.g. Yahoo, Altavista or Google directories, the Open Directory Project -ODP-, etc.), in which documents are mostly manually classified into hierarchical topical clusters, are an optimal source to acquire lexical information; their size is not comparable to the full Web, but they are still enormous sources of semi-structured, semi-filtered information waiting to be mined.

In this Chapter, we will describe an algorithm for assigning Web directories (from ODP) as characterizations for word senses in WordNet 1.7 senses[1].

For instance, let us consider the noun *circuit*, which has six senses in WordNet 1.7. These senses are grouped in *synsets* together with their synonym terms, and linked to broader (more general) synsets via hypernymy relations:

```
6 senses of circuit

Sense 1: {circuit, electrical circuit, electric circuit} => {electrical device}

Sense 2: {tour, circuit} => {journey, journeying}

Sense 3: {circuit} => {path, route, itinerary}
```

[1] WordNet 1.7 was the most recent version at the time of doing our experiments

```
Sense 4: {circuit (judicial division)} => {group, grouping}

Sense 5: {racing circuit, circuit} => {racetrack, racecourse, raceway, track}

Sense 6: {lap, circle, circuit} => {locomotion, travel}
```

Our algorithm associates *circuit 1* (electric circuit) with ODP directories such as:

```
business/industries/electronics and electrical/contract manufacturers
```

whereas the *circuit 5* (racing circuit) is tagged with directories such as:

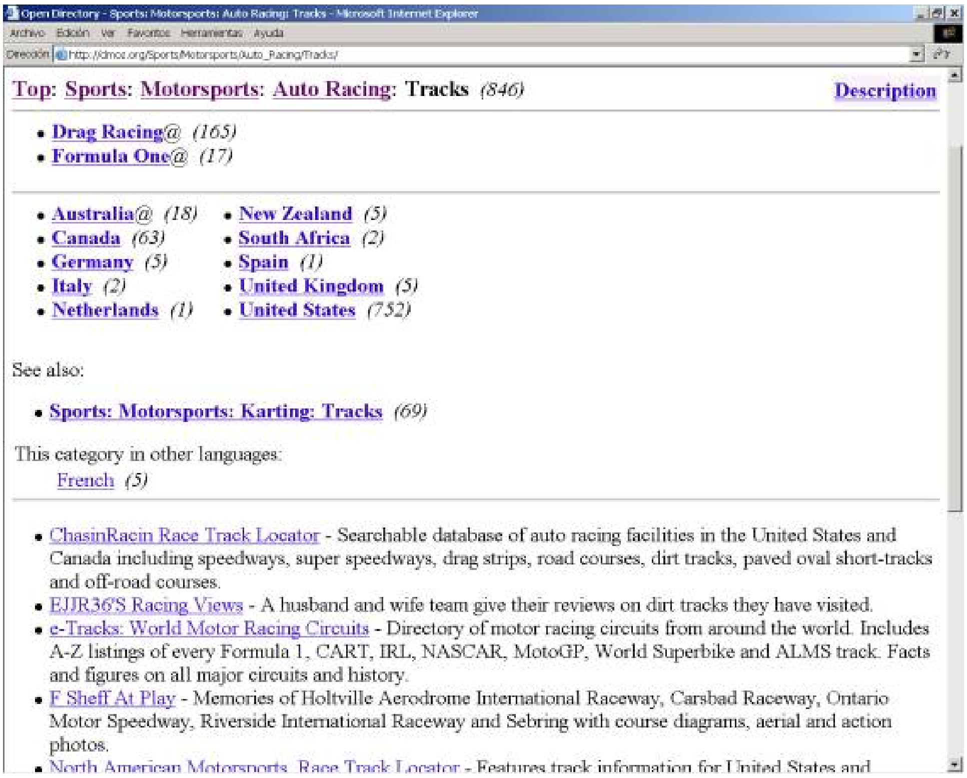
```
sports/motorsports/auto racing/tracks
sports/equestrian/racing/tracks
sports/motorsports/auto racing/formula one
```

As we have previously explained (see Section 2.1.1), every ODP directory has an associated URL, which contains a description of the directory and a number of Web sites that have been manually listed as pertaining to the directory topic, accompanied by brief descriptions of each site. This information is completed with a list of subdirectories, each containing more Web sites and subdirectories. Finally, some directories have also pointers to the same category in other languages. For instance, the Web page for the directory `sports/motorsports/auto racing/tracks` can be seen in Figure 3.1. This directory contains links and descriptions for 846 Web sites organized in 12 subdirectories, a link to a related directory (`sports/motosports/karting/tracks`) and a link to the same category in French.

The association of word senses with Web directories is related to the assignment of domain labels to WordNet synsets as described in [Magnini and Cavaglia, 2000], in which WordNet is (manually) enriched with domain categories from the Dewey Decimal Classification (DDC). Some clear differences between the two are that directories from the ODP are assigned automatically, are richer and deeper, and, more importantly, come with a large amount of associated information directly retrievable from the Web. DDC categories, on the other hand, are a stable domain characterization compared to Web directories.

As WordNet and ODP are both hierarchical structures, connecting them is also related to research in mapping thesauri for digital libraries, ontologies and data structures in compatible databases. A salient feature of our task is, however, that we do not intend to map both structures, as they are of a quite different nature (lexicalized English concepts vs. topics on the Web). Our goal is rather to associate

Figure 3.1: Contents of an ODP Web directory associated to *circuit 5* (racing circuit).


Open Directory - Sports: Motorsports: Auto Racing: Tracks - Microsoft Internet Explorer
Archivo Edición Ver Favoritos Herramientas Ayuda
Dirección http://dmoz.org/Sports/Motorsports/Auto_Racing/Tracks/

Top: Sports: Motorsports: Auto Racing: Tracks *(846)* Description

- Drag Racing@ *(165)*
- Formula One@ *(17)*

- Australia@ *(18)*
- Canada *(63)*
- Germany *(5)*
- Italy *(2)*
- Netherlands *(1)*
- New Zealand *(5)*
- South Africa *(2)*
- Spain *(1)*
- United Kingdom *(5)*
- United States *(752)*

See also:

- Sports: Motorsports: Karting: Tracks *(69)*

This category in other languages:
French *(5)*

- ChasinRacin Race Track Locator - Searchable database of auto racing facilities in the United States and Canada including speedways, super speedways, drag strips, road courses, dirt tracks, paved oval short-tracks and off-road courses.
- EJJR36'S Racing Views - A husband and wife team give their reviews on dirt tracks they have visited.
- e-Tracks: World Motor Racing Circuits - Directory of motor racing circuits from around the world. Includes A-Z listings of every Formula 1, CART, IRL, NASCAR, MotoGP, World Superbike and ALMS track. Facts and figures on all major circuits and history.
- F Sheff At Play - Memories of Holtville Aerodrome International Raceway, Carsbad Raceway, Ontario Motor Speedway, Riverside International Raceway and Sebring with course diagrams, aerial and action photos.
- North American Motorsports Race Track Locator - Features track information for United States and

individual items in a many-to-many fashion. A word sense may be characterized with several Web directories, and a Web directory may be suitable for many word senses.

The most direct applications of word sense/Web directory associations are:

- Clustering of senses with identical or very similar categories.
- Refinement of senses into specialized variants (e.g. *equestrian circuit*, *formula one circuit* as specializations of *racing circuit* in the example above).
- Extraction of sense-tagged corpora from the Web sites listed under the appropriate directories.

In Section 3.2, we describe the proposed algorithm. In Section 3.3, we evaluate the precision and recall of the algorithm for the set of nouns used in the Senseval

2 WSD competition. In Section 3.4, we make a preliminary experiment using the material from the ODP directories as training corpora for a supervised WSD system. In Section 3.5, we present the results of applying the algorithm to most WordNet 1.7 nouns. Finally, in Section 3.6, we draw some conclusions.

3.2 Algorithm

Overall, the system takes a WordNet 1.7 noun as input, generates and submits a set of queries into the ODP directories, filters the information obtained from the search engine and retrieves a set of directories classified as: (i) pseudo domain labels for some word sense, (ii) noise, and (iii) salient noise (i.e. directories which are not suitable for any sense in WordNet, but could reveal and characterize a new relevant sense of the noun). In case (i), the WordNet sense ↔ ODP directory association also receives a probability score.

A detailed description of the algorithm steps is presented as follows.

3.2.1 Querying ODP Structure

For every sense w_i of the noun w, a query q_i is generated, including w as compulsory term, the synonyms and direct hypernyms of w_i as optional terms, and the synonyms of other senses of w as negated (forbidden) terms. These queries are submitted to ODP, and a set of directories is retrieved. For instance, for *circuit* the following queries are generated and sent to the ODP search engine: [2]

```
 q1= [+circuit "electrical circuit" "electric circuit" "electrical
device" -tour
-"racing circuit" -lap -circle]
 q2= [+circuit tour journey journeying -"electrical circuit" -"electric
circuit"
-"electrical device" -"racing circuit" -lap -circle]
 q3= [+circuit path route itinerary -"electrical circuit" -"electric
circuit"
-"electrical device" -tour -"racing circuit" -lap -circle ]
 q4= [+circuit group grouping -"electrical circuit" -"electric circuit"
-"electrical device" -tour -"racing circuit" -lap -circle]
 q5= [+circuit "racing circuit" racetrack racecourse raceway track
-"electrical circuit"
-"electric circuit" -"electrical device" -tour -lap -circle]
```

[2] In ODP queries, compulsory terms are denoted by + and forbidden terms by -

```
 q6= [+circuit lap circle locomotion travel -"electrical circuit" -"electric
circuit"
-"electrical device" -tour -"racing circuit" -lap -circle]
```

3.2.2 Representing Retrieved Directory Descriptions

For each directory d, a list of words $l(d)$ is obtained removing stop words and preserving all content words in thedirectory path. For instance, one of the directories produced by the *circuit* queries is:

$d =$`business/industries/electronics and electrical/contract manufacturers`

which is characterized by the following word list:

$l(d) =$ `[business, industries, electronics, electrical, contract, manufacturers]`

3.2.3 Representing WordNet Senses

For every sense w_j, a list $l(w_j)$ of words is made with

- all nouns in the hypernymy chain of maximal length 6
- all hyponyms
- all meronyms, holonyms and coordinate terms

of w_j in WordNet. $l(w_j)$ is used as a description of the sense w_j. For instance, *circuit 1* receives the following description:

$l(\textit{circuit}_1) =$ `[electrical circuit, electric circuit, electrical device, bridge,`
```
bridge circuit, Wheatstone bridge, bridged-T, closed circuit, loop,
parallel circuit,
shunt circuit, computer circuit, gate, logic gate, AND circuit, AND
gate, NAND circuit,
NAND gate, OR circuit, OR gate, X-OR circuit, XOR circuit, XOR gate,
integrated circuit,
(..)
instrumentality, instrumentation, artifact, artefact, object, physical
object, entity]
```

3.2.4 Sense/Directory Comparisons

For every sense description $l(w_j)$, a comparison is made with the terms in the directory description $l(d)$. This comparison is based on the hypothesis that the terms in an appropriate directory for a word sense will have some correlation with the sense description via WordNet semantic relations. In other words, our assumption is that the path to the directory in the ODP topical structure will have some degree of overlapping with the hyponymy path to the word sense in the WordNet hierarchical structure.

For this comparison, we simply count the number of co-occurrences between words in $l(w_j)$ and words in $l(d)$. Repeated terms are not discarded, as repetition is correlated with stronger associations. Other, better-grounded comparisons, such as the cosine between $l(w_j)$ and $l(d)$, were empirically discarded because of the small size and small amount of overlapping of the average vectors.

3.2.5 Candidate Sense/Directory Associations

The association vector $v(d, w)$ has as many components as senses for w in WordNet 1.7; the i^{th} component, $v(d, w)_i$ represents the number of matches between the directory $l(d)$ and the sense descriptor $l(w_j)$. For instance, the association vector of

`business/industries/electronics and electrical/contract manufacturers`

with *circuit* is

$$v(d, \text{circuit}) = (6, 0, 0, 0, 0, 0)$$

that is, six coincidences for sense 1 (the *"electric circuit"* sense), which has the associated vector shown in the previous Section (which includes five occurrences of *electrical* and one occurrence of *electronic*). The rest of the sense descriptions have no coincidences with the directory description.

$v(d, w)$ is the basis for making candidate assignments of suitable senses for directory d: If one of the components $v(d, w)_j$ is not null, we assign the sense w_j to the directory d. If all components are null, the directory is provisionally classified as noise or new sense. If more than one component is not null, the senses i with maximal $v(d, w)_i$ are all considered as candidates.

These candidate assignments are confirmed or discarded after passing a number of filters and receiving a confidence score $C(d, w_j)$,both of which are described below.

3.2.6 Filters

Filters are simple heuristics that contribute to a more accurate classification of the relations predicted by the co-occurrence vector $v(d, w)$. We used two filters: one differentiates nouns and noun modifiers to prevent wrong associations and another detects sense specializations.

Modifiers

Frequently, the ODP search engine retrieves directories in which the noun to be searched, w, has as a noun modifier role. Such cases usually produce erroneous associations. For instance, the directory:

`library/sciences/animals & wildlife/mammals/tamarins/golden lion tamarin`

is erroneously associated to the mammal sense of "lion", which is here a modifier for "tamarin".

Modifiers are detected with a set of simple patterns, as the syntactic properties of descriptions in directories are quite simple. In particular, we discard most cases using the structure of the ODP hierarchy, as in this case. The filter analyzes the structure of the directory, detects that the parent category of *golden lion tamarin* is *tamarin*, and therefore assumes that *golden lion tamarin* is a specialization of *tamarin*, and assigns the directory to a suitable sense of *tamarin* (*tamarin 1* in WordNet).

An additional filter (weaker than the previous one) is discarding compounds according to the position (the searched noun precedes another noun), as in:

`personal/kids/arts & entertainment/movies/animals/lion king`

This directory could be associated with *lion 1* because in contains the word *animal*, but the assignment is rejected because of the modifier filter. In general, on such occasions the searched noun plays a modifier role (as adjective or noun); discarding all such cases favors precision over recall. In this case, the label is classified as noise.

Sense Specializations (Hyponyms)

A retrieved directory might be appropriate as a characterization of a sense specialization for some of the word senses being considered; our algorithm tries to detect such cases, creating a hyponym of the sense and characterizing the directory with the hyponym.

The filter identifies a directory as a candidate hyponym if it contains explicitly a `modifier w` pattern (where w is the noun being searched). This filter detects explicit specializations such as *office chair* as a hyponym of *chair 1*, or *family fox channel* as a hyponym of *channel 7*; but fails to identify, for instance, *memorial day* as a hyponym of *holiday*.

If the candidate hyponym, as a compound, is not present in WordNet, then it is incorporated and described with the directory. If it is already present in

WordNet, an additional checking of the hyponymy relation is made. For instance, the directory

```
   business/industries/electronics and electrical/components/integrated
circuits
```

is assigned to the WordNet entry *integrated circuit*, because *integrated circuit* is already a hyponym of *circuit* in WordNet.

3.2.7 Confidence Score

Finally, a confidence score $C(d, w_j)$ for every potential association (d, w_j) is calculated using four empirical criteria:

1. Checking whether d was directly retrieved for the query associated to w_j.
2. Checking whether the system associates d with one or more senses of the word w.
3. Checking the number of coincidences between $l(d)$ and $l(w_j)$.
4. Comparing the previous number with the number of coincidences between $l(d)$ and the other sense descriptions $\{l(w)_i, i \neq j\}$.

The confidence score is a linear combination of these factors, weighted according to an empirical estimation of their relevance:

$$C(d, w_j) = \sum_{i=1}^{4} \alpha_i C_i(d, w_j)$$

where

$$C_1(d, w_j) = \begin{cases} 1 & \text{if query}(w_j) \text{ retrieves } d \\ 0 & \text{otherwise} \end{cases}$$

$$C_2(d, w_j) = 1 - \frac{k}{n}$$

$$C_3(d, w_j) = \begin{cases} 1, & \text{if } v_j \geq 5 \\ (v_j + 5)/10, & \text{if } 1 < v_j \leq 4 \\ 0.5, & \text{if } v_j = 1 \end{cases}$$

$$C_4(d, w_j) = \frac{v_j - \max_{i \neq j}(v_i),}{\sum_{i=1}^{n} v_i}$$

where v is the association vector $v(d, w)$, n the number of senses, k the number of senses for which v_j is non-null, and α_i coefficients empirically adjusted to

$(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = (0.1, 0.15, 0.4, 0.35)$. The value of $C(d, w_j)$ ranges between 0 and 1 (all C_i range between 0 and 1, and the sum of the linear coefficients α_i is 1). Note that C_2 cannot reach 1 (but can get asymptotically close to 1), and note also that C_4 cannot take negative values, because - as (d, w_j) is a candidate association, v_j is maximal in $v(d, w)$ and therefore $v_j - \max_{i \neq j}(v_i)$ ranges between 0 and v_j.

Let us see an example of how this confidence measure works, calculating $C(d, w_j)$ for the directory `d=business/industries/electronics and electrical/contract manufacturers` with *circuit 1* (electric circuit):

- C_1. This directory has been retrieved from the query

  ```
  q1= [+circuit "electrical circuit" "electric circuit" "electrical
  device" -tour -"racing circuit" -lap -circle]
  ```

  corresponding to *circuit 1*, which agrees with the association made by the system. Hence $C_1 = 1$.

- C_2. The association vector $v(d, w) = (6, 0, 0, 0, 0, 0)$ presents only one non null coordinate; Therefore $C_2 = 1 - \frac{1}{6} = 0.83$. Note that, in general, this factor prevents C from reaching the upper bound 1.

- C_3. As $v_1 = 6$, $C_3 = 1$. This factor increases along with the number of coincidences between the sense and directory characterizations.

- C_4. As all other components of v are null, the highest value of the components different from sense 1 is also null ($\max_{i \neq j}(v_i) = 0$); therefore, $C_4 = 1$. This factor measures the strength of the association (d, w_1) compared with the other possibilities. It decreases when $v(d, w)$ includes more than one non null coordinate, and their values are similar.

- C. Finally, applying the α_i coefficients, we obtain $C(\mathtt{d}, \textit{circuit 1}) = 0.975$

The confidence score can be used to set a threshold for accepting/discarding associations; a higher threshold should produce a lower number of highly precise associations; a lower threshold would produce more associations with less accuracy. For the evaluation below, we have retained all directories, regardless of their confidence score, in order to assess how well this empirical measure correlates with correct and useful assignments.

An example of the results produced by the algorithm can be seen in Figure 3.2. The system assigns directories to senses 1, 2 and 5 of *circuit* (six, two and three directories, respectively). Some of them are shown in the Table, together with a sense specialization, *integrated circuit* for sense 1 (electrical circuit). Senses 3, 4 and 6, which did not receive any directory association, do not appear to have domain specificity, but are instead general terms.

Figure 3.2: Results of the association algorithm for *circuit*

circuit 1 (electrical circuit)

ODP directories	C
`business/industries/electronics and electrical/contract manufacturers`	0.98
`manufacturers/printed circuit boards/fabrication`	0.88
`computers/cad/electronic design automation`	0.78
...	
sense specializations (hyponyms)	
`business/industries/electronics and electrical/components/integrated circuits`	0.98

circuit 2 (tour, journey around a particular area)

ODP directories:	
`sports/cycling/travel/travelogues/europe/france`	0.58
`regional/asia/nepal/travel and tourism/travel guides`	0.66

circuit 5 (racing circuit)

ODP directories:	
`sports/motorsports/auto racing/stock cars/drivers and teams`	0.78
`sports/motorsports/auto racing/tracks`	0.82
`sports/motorsports/auto racing/driving schools`	0.78

3.3 Evaluation

We analyzed the results of the algorithm for the set of nouns in the Senseval 2[3] WSD English lexical sample test bed [Kilgarriff, 2001b]. The Senseval campaigns ([Kilgarriff and Palmer, 2000, Kilgarriff, 2001a, Kilgarriff, 2002, Mihalcea et al., 2004]), which have had a continuation in SemEval, were devoted to the comparative evaluation of Word Sense Disambiguation systems in many languages. In the Senseval 2 lexical sample task, a large number of instances (occurrences in context extracted from corpora) for a fixed set of words had to be tagged with the appropriate sense by the participating WSD systems. For English, the sense inventory was a pre-release of WordNet 1.7, and two sets of manually tagged instances were made available: a first set was intended for training supervised systems, and a second set for evaluation of all systems attempting the task. Altogether, the Senseval lexical samples test beds are one of the most widely used resources to study and compare Word Sense Disambiguation approaches.

For our evaluation, we considered the fraction of the Senseval 2 test bed that deals with English nouns: 29 polysemyc nouns with a total of 147 word senses. We applied the algorithm to this set of nouns, and examined the results in terms of coverage and quality of the sense/directory associations. Coverage measures how many senses can be characterized with directories, assuming that every domain-specific sense should receive at least one directory. Quality is measured in terms of precision (are the assignments correct?), relevance (are the assignments useful?), and confidence (does the confidence score correlate well with precision and relevance of the associations?).

3.3.1 Coverage

Table 3.1 shows the 148 directories retrieved by our algorithm, which makes an average of 1.0 directories per sense. The directories, however, are not evenly distributed among senses, covering only 43 different senses with unique directories, and 28 specialized (hyponym) senses. In addition, 9 senses are identified as part of potential clusters (i.e. having non unique directories).

In order to measure the real coverage of the system, we have to estimate how many word senses in the Senseval 2 sample are susceptible to receiving a domain label. For instance, *"sense"* in *"common sense"* is not associated to any particular topic or domain, whereas *"sense"* as *"word sense"* can be associated with linguistics or language-related topics.

The decision as to whether or not a word sense might receive a domain label or not is not always a simple, binary one. Hence we manually tagged all word senses

[3] Senseval 2 was the most recent Senseval/Semeval campaign at the time of doing this research.

Table 3.1: Coverage of nouns in the Senseval 2 test bed.

Senseval 2 nouns	**# senses**	**# directories**	**# labeled senses**	**# hyponyms**
art	4	6	1	1
authority	7	4	2	1
bar	13	3	3	0
bum	4	0	0	0
chair	4	4	1	0
channel	7	5	1	1
child	4	12	2	0
church	3	24	2	4
circuit	6	11	3	1
day	10	15	1	14
detention	2	1	1	0
dyke	2	1	1	0
facility	5	10	3	0
fatigue	4	0	0	0
feeling	6	2	1	0
grip	7	3	2	0
hearth	3	5	2	0
holiday	2	2	2	0
lady	3	0	0	0
material	5	9	2	3
mouth	8	0	0	0
nation	4	4	1	1
nature	5	0	0	0
post	8	14	5	0
restraint	6	4	3	0
sense	5	0	0	0
spade	3	3	1	1
stress	5	5	2	1
yew	2	1	1	0
Total	**147**	**148**	**43**	**28**

with two criteria (with each tagging performed by a different human annotator): a strict one (only word senses which can clearly receive a domain label are marked as positive), and a loose one (only word senses that are completely generic are marked as negative). The strict judgment gave 59 domain-specific senses in the sample; the loose judgment gave 71.

With these manual judgments, the coverage of the algorithm is between 73% (loose judgment) and 88% (strict judgment). This coverage can be increased by:

- Propagating a directory/word sense association to all members of the WordNet synset to which the word sense belongs.
- Propagating directories via hyponymy chains as in [Magnini and Cavaglia, 2000].

Table 3.2: Precision over Senseval 2 nouns.

Directories associated to WordNet senses	**# directories**	**# correct**	**# errors**
Unique sense	148	127	21
potential clustering	13	12	1
Total	161	139 (86%)	22 (14%)

3.3.2 Quality

We used three criteria to evaluate the directory/sense associations produced:

Precision Is the directory *correct* (suitable) for the word sense?

Relevance Is the directory *useful* to characterize the word sense?

Confidence How well is the confidence value $C(d, w_j)$ correlated with the precision and relevance of the associations?

Precision

An assignment (d, w_j) is considered correct (d is suitable for w_j) unless:

1. d adjusts better to some other sense w_i. For instance, the association of:

   `regional/north america/united states/government/agencies/`

   `independent/federal labor relations authority`

   as a hyponym of

   authority 4: assurance, self-assurance, confidence, self-confidence, authority, sureness

   is considered an error, as the directory would be better suited for a hyponym of sense 5 (authority as *administrative unit*).

2. The terms in $l(d)$ are contradictory with the definition of the word sense, or are better suited for a sense that is not listed in the dictionary. This is the case of:

   `arts/music/bands and artists/offspring`

   which is erroneously assigned to *child 2: human offspring of any age*.

The results of this manual evaluation can be seen in Table 3.2. The overall precision is 86%.

Regarding potential topical clusters (directories associated to more than one sense of the same word), these are considered correct if 1) the associated directory is correct for all the senses in the cluster, and 2) the occurrences of the word in the Web page associated with the directory can be loosely assigned to any of the cluster senses. 12 out of the 13 clusters extracted are correct according to this criterion.

Confidence Measures

Table 3.3 shows the distribution of directories according to the confidence measure. 84% of the directories have a confidence C over 0.7, and 41% over 0.8. This skewed distribution is consistent with the algorithm filters, designed to favor precision rather than recall.

Table 3.4 shows the distribution of errors in levels of confidence. The percentage of errors in directories with a confidence level below .6 is 25%. This error percentage decreases with increasing levels of confidence, down to 5% for associations with C over .8. This table indicates that the confidence value, which is assigned heuristically, is indeed correlated with precision.

Table 3.3: Confidence distribution.

Confidence	$C \leq 0.7$	$0.7 < C \leq 0.8$	$0.8 < C$
# directories	**24**	**63**	**61**

Table 3.4: Correlation between confidence and correctness

Confidence	**# directories**	**% errors**
$C \leq 0.7$	24	25%
$0.7 < C \leq 0.8$	63	19%
$C > 0.8$	61	5%
Total	148	14%

Relevance

Besides correctness of the associations, we want to measure the usefulness of the directories: how well can they be used to characterize the associated word senses? How much information do they provide about the word senses?

We performed a manual, qualitative classification of the directories extracted as irrelevant, mildly relevant or very relevant. An **irrelevant** directory is compatible with the word sense, but does not provide any useful characterization; a **mildly relevant** directory illustrates the word sense, but not centrally or in some particular aspect or domain. A **very relevant** directory provides a rich characterization per se, and can be considered as a domain label for the word sense.

An example of a very relevant directory is:

`business/industries/electronics and electrical/components/integrated circuit`

associated as hyponym of *circuit 1 (electrical circuit)* with a confidence of 98%.

An example of mildly relevant association is

`regional/north america/united states/texas/../society and culture/religion`

associated with *church 1 (Christian church)* with a 73% confidence. Obviously, Texas is not correlated with church, but the directory contains a lot of material (for instance the Web page of the "Northcrest Community Church" and many others) that might be used, for instance, to acquire topical signatures for the concept.

Hence the *mildly relevant* judgment.

Finally, an example of an irrelevant association is:

`regional/north america/united states/new york/localities/utica` associated with *art 1 (fine art)* with a confidence of 66% (the directory contains a section of Arts at Utica, which would be considered mildly relevant if pointed to explicitly by the label).

For the purposes of measuring relevance, all the directories that were judged as *incorrect* are counted as *irrelevant*.

The overall relevance figures, and the correlation of relevance with the confidence value, can be seen in Table 3.5. 67% of the directories are highly relevant to characterize word senses, which is an encouraging result. Also, the set of irrelevant directories (15%) is almost identical to the set of erroneous directories (with just one addition), indicating that (almost) all directories that are correct can be used to characterize word senses to some extent.

Table 3.5: Relevance of the directories in the test set.

Relevance	**Irrelevant**	**Mildly relevant**	**Highly relevant**
$C \leq 0.7$	7	4	13
$0.7 < C \leq 0.8$	13	12	38
$0.8 < C$	3	9	49
Total	23 (15%)	25 (17%)	100 (67%)

3.4 Example Application: Automatic Acquisition of Sense-Tagged Corpora

Each ODP directory contains links to related subdirectories, and to a large number of Web sites that have been manually classified there. Every link to a Web site includes the name of the site and a short description. For instance, under

`business/industries/electronics and electrical/components/integrated circuit`

we find over 30 descriptions such as `''Multilink Technology corporation: Manufacture of integrated circuits, modules, and boards for use in both data and telecommunications''`.

In order to perform a first experiment on extraction of sense-tagged corpora, we have used only such descriptions (without exploring the associated Web sites) to build a sense-tagged corpus for Senseval 2 nouns.

Notice that we are not using the contents of the Web sites that belong to a directory, but only the manually added description of Websites in the directory. Using the Web sites themselves is also an attractive possibility, that would produce a much larger corpus at the expense of lower precision.

The extraction is straightforward: When a word sense w_i has an associated directory d, we scan the site descriptions in the ODP page that corresponds to the directory d and extract all contexts in which w occurs, assuming that in all of them w is used in the sense i.

Some examples of the training material for *circuit* can be seen in Figure 3.3. On average, these examples are shorter than Senseval 2 training instances.

Figure 3.3: Examples of training material for *circuit*

circuit 1 (electrical circuit)

```
Electromechanical products for brand name firms;
offers printed circuit boards (..)
Offers surface mount, thru-hole, and flex circuit assembly,
in circuit and functional (..)
```

circuit 2 (tour, journey around a particular area)

```
The Tour du Mont-Blanc is a circuit of 322km
based in the northern French Alps.
A virtual tour of the circuit by Raimon Bach.
```

circuit 5 (racing circuit)

```
The Circuit is a smooth 536 yards of racing for Hot Rod
and Stock Car's at the East of (..)
(..)  History of the circuit and its banked track
and news of Formula 1 (..)
```

The goal is to compare the performance of a supervised word sense disambiguation system using Senseval 2 training data (hand made for the competition) to that using the sense-tagged corpus from ODP (automatically extracted). We have chosen the *Duluth* system [Pedersen, 2001] to perform the comparison. The Duluth system is a freely available supervised WSD system that participated in the Senseval 2 competition. As we are not concerned with absolute performance, we simply adopted the first of the many available versions of the system (*Duluth 1*).

An obstacle to performing such comparative evaluation is that, as expected, our algorithm assigns ODP directories only to a fraction of all word senses, partly because not every sense is domain-specific, partly because of lack of coverage. In order to circumvent this problem, we considered only the subset of 10 Senseval nouns for which our system tags at least two senses: *bar, child, circuit, facility, grip, holiday, material, post, restraint, stress*. We then projected the Senseval 2 training corpus, and the test material, onto the annotations for the word senses already in our ODP-based material. Hence we will evaluate the quality of the training material obtained from Web directories, not the coverage of the approach.

Table 3.6 shows the training material obtained for that subset of Senseval 2 nouns. A total of 66 directories are used as a source of training instances, of which 17% of them are incorrect and will presumably incorporate noise into the training.

Table 3.6: Training material obtained for the WSD experiment

word senses	**# directories per sense**	**# incorrect directories**	**# training instances**
bar 1,10	1,1	0,0	1,1
child 1,2	3,9	0,0	3,80
circuit 1,2,5	6,2,3	0,0,0	229,2,5
facility 1,4	4,5	0,0	4,18
grip 2,7	2,1	0,1	17,6
holiday 1,2	1,1	0,1	5,17
material 1,4	6,3	2,1	63,10
post 2,3,4,7,8	1,5,1,4,3	1,1,1,0,3	2,7,1,9,3
restraint 1,4,6	2,1,1	0,0,0	2,2,2
stress 1,2	1,4	0,0	8,50
Total	**66**	**11**	**547**

Table 3.7 compares the training material for the word senses in this sample, and the results of the supervised WSD algorithm with the Senseval and the ODP training instances.

We measured the performance of the system in terms of Senseval *recall*: the number of correctly disambiguated instances over the total number of test instances. Overall, using the Senseval training set gives .73 recall, and training with the automatically extracted ODP instances gives .58 (21% worse). A decrease of 21% is significant but nevertheless encouraging, because the Senseval training set is the

Table 3.7: Results of supervised WSD

word senses	**# instances** Senseval training	**# instances** ODP training	**# test instances**	**Recall** senseval training	**Recall** ODP training
bar 1,10	127,11	1,1	62,6	.91	.50
child 1,2	39,78	3,80	35,27	.57	.44
circuit 1,2,5	67,6,7	229,2,5	23,2,8	.70	.70
facility 1,4	26,61	4,18	15,28	.79	.67
grip 2,7	6,1	17,6	4,0	1	1
holiday 1,2	4,57	5,17	26,2	.96	.96
material 1,4	65,7	63,10	30,9	.79	.79
post 2,3,4,7,8	1,64,20,11,7	2,7,1,9,3	2,25,13,12,4	.45	.25
restraint 1,4,6	17,32,11	2,2,2	8,14,4	.65	.50
stress 1,2	3,45	8,50	1,19	.95	.95
Total	**773**	**547**	**379**	**.73**	**.58**

gold standard for the Senseval test set: it is larger than the ODP set (773 versus 547 instances in this subset), well balanced, built with redundant manual annotations, and part of the same corpus as the test set.

The most similar experiment in the literature is [Agirre and Martínez, 2000], where the sense-tagged instances obtained using a high-performance Web mining algorithm [Mihalcea and Moldovan, 1999] performed hardly better than a random baseline as WSD training instances. A difference between the two experiments is that Agirre et al. do not limit their experiments to the fraction of the test set for which they have automatically extracted training samples, hence a direct comparison of the results is not possible.

A detailed examination of the results indicates that the difference in performance is related to the smaller number of training instances rather than to the quality of individual instances:

- In all four cases where ODP provides a comparable -or larger- number of training instances (*circuit, grip, material, stress*), ODP training equals hand-tagged training. In one more case (*holiday*), the number of ODP instances is lower but still the recall is the same. For the other five words, the number of ODP instances is substantially lower and the recall is worse.

- Remarkably, incorrect directories harm recall substantially only for *post*, which accumulates six erroneous associations (out of 11 errors). The other five errors (in *material 1,4, holiday 2, grip 7*) do not affect the final recall for these words. There are two possible reasons for this behavior:

  - Erroneous directories tend to be less productive in terms of training instances. Indeed, this fact could be incorporated as an additional filter for candidate directories. This is the case, for instance, of *material 1*, for which correct directories provide much more training material than the incorrect one.

  - Erroneous directories are more frequent with rare (less frequent) word senses. This is correlated with a lower number of test instances (hence the influence on average recall is lower) and also of training instances (and then the reference, hand-tagged material does not provide good training data either). This is the case of *grip 7* or *holiday 2*, which have 0 and 2 test instances respectively.

Overall, our results suggest that directory-based instances, in spite of being shorter and automatically extracted, are not substantially worse for supervised WSD than the hand-tagged material provided by the Senseval organization. The limitation of the approach is currently the low coverage of word senses and the amount of training samples. Two strategies may help overcoming such limitations: first, propagating directories via synonymy (attaching directories to synsets rather than word senses) and semantic relationships (propagating directories via hyponymy relations); second, retrieving instances not only from the ODP page describing the directory contents, but from the Web pages listed in the directory.

The only fundamental limitation of our approach for the automatic extraction of annotated examples is the fact that directories are closely related to topics and domains and, therefore, word senses that do not pertain to any domain cannot receive directories and training instances from them. Still the approach can be very useful for language engineering applications in which only domain disambiguation (versus sense disambiguation) is required, such as information retrieval [Gonzalo et al., 1998], content-based user modelling [Magnini and Strapparava, 2000], etc.

3.5 Massive Processing of WordNet Nouns

We have applied the association algorithm to all non-compound nouns in WordNet without non-alphabetic characters (e.g. “sea lion” and “10” are not included in the bulk processing). The results can be seen in Table 3.8. Overall, the system associates at least one directory to 13,375 nouns (28% of the candidate set).

The most direct way of propagating directories in the WordNet structure is extending sense/directory associations to synset/directory relations (i.e., if a word sense receives a directory, then all word senses in the same synset receive the same directory). For instance, *cable 2* (transmission line) receives the following directories:

```
business/industries/electronics and electrical
business/industries/electronics and electrical/hardware/connectors and terminals
business/industries/electronics and electrical/contract manufacturers
```

As *cable 2* is part of the synset {*cable 2, line 9, transmission line 1*}, *line 9* and *transmission line 1* inherit the three directories.

With this (quite conservative) strategy, the number of characterized nouns and word senses almost doubles: 24,558 nouns and 27,383 senses, covering 34% of the candidate nouns plus 7,027 multi-word terms that were not in the candidate set.

The results of this massive processing, together with the results for the Senseval 2 test (including training material) are available for public inspection at `http://nlp.uned.es/ODP`.

Table 3.8: Massive association of ODP directories to WordNet 1.7 nouns

		with propagation
Candidate nouns	51,168	
candidate senses	73,612	
Associated directories	29,291	
Characterized nouns	13,375	24,558
Characterized senses	14,483	27,383
Hyponyms	1,800	

3.6 Conclusions

Our algorithm is able to associate ODP directories to WordNet senses with 86% accuracy over the Senseval 2 test, and with coverage between 73% and 88% of the domain specific senses. Such associations can be used as rich characterizations for word senses: as a source of information to cluster senses according to their topical relatedness, to extract topic signatures, to acquire sense-tagged corpora, etc. The only intrinsic limitation of the approach is that Web directories are not appropriate to characterize general word senses (versus domain-specific ones). If such characterization is necessary for a particular Natural Language application, the method should be complemented by other means of acquiring lexical information.

In the supervised WSD experiment we have carried out, the results suggest that the characterization of word senses with Web directories provide cleaner data, without further sophisticated filtering, than a direct use of the full Web. Indeed the WSD results using training material from ODP directories gives better results than could be expected from previous cross-validations of training and test WSD materials.

Perhaps the main conclusion of our work is that Web directories are a much more structured and reliable corpus than the whole Web. In spite of being manually supervised, Web directories offer immense amounts of structured corpora that deserve our attention as sources of linguistic information. In particular, listing word sense/ODP directory associations has the additional advantage, compared to other Web-mining approaches, of providing a wealth of lexical information in a very condensed manner.

In this Chapter, we have shown how we can extract useful lexical information that is implicit in a collaboratively authored Web resource (ODP). In the next chapter we will change the perspective of our research, and test whether a collaboratively authored Web encyclopedia (Wikipedia) can replace a conventional lexical database (WordNet) in an Information Access task: organization of Web search results.

Chapter 4

Use of Wikipedia to Organize Web Search Results

4.1 Introduction

The application of Word Sense Disambiguation (WSD) to Information Retrieval (IR) has been subject of a significant research effort in the recent past. The essential idea is that, by indexing and matching word senses (or even meanings) , the retrieval process could better handle polysemy and synonymy problems [Sanderson, 2000]. In practice, however, there are two main difficulties: (i) for long queries, IR models implicitly perform disambiguation, and thus there is little room for improvement. This is the case with most standard IR benchmarks, such as TREC[1] or CLEF[2] ad-hoc collections; (ii) for very short queries, disambiguation may not be possible or even desirable. This is often the case with one word and even two word queries in Web search engines.

In Web search, there are at least three ways of coping with ambiguity:

- Promoting diversity in the search results [Clarke et al., 2008]: given the query ”*oasis*”, the search engine may try to include representatives for different senses of the word (such as the Oasis band, the Organization for the Advancement of Structured Information Standards, the online fashion store, etc.) among the top results. Search engines are supposed to handle diversity as one of the multiple factors that influence the ranking.

- Presenting the results as a set of (labeled) clusters rather than as a ranked list [Carpineto et al., 2009].

[1] trec.nist.gov

[2] www.clef-campaign.org

- Complementing search results with search suggestions (e.g. ”*oasis band*”, ”*oasis fashion store*”) that serve to refine the query in the intended way [Anick, 2003].

All of them rely on the ability of the search engine to cluster search results, detecting topic similarities. In all of them, disambiguation is implicit, a side effect of the process but not its explicit target. Clustering may detect that documents about the Oasis band and the Oasis fashion store deal with unrelated topics, but it may as well detect a group of documents discussing why one of the Oasis band members is leaving the band, as opposed to another group of documents containing lyrics by the band; both are different aspects of the broad topic Oasis band. A perfect hierarchical clustering should distinguish among the different Oasis senses at a first level, and then discover different topics within each of the senses.

Is it possible to use sense inventories in order to improve search results for one word queries? To answer this question, we will focus on two broad-coverage lexical resources of a different nature: WordNet, as a de-facto standard used in Word Sense Disambiguation experiments and many other Natural Language Processing research fields; and Wikipedia, as a large coverage and updated encyclopedic resource which may have a better coverage of relevant senses in Web pages.

Our hypothesis is that, under appropriate conditions, any of the above mechanisms (clustering, search suggestions, diversity) might benefit from an explicit disambiguation (classification of pages in the top search results) using a wide-coverage sense inventory. In order to test this hypothesis, our research is focused on four relevant aspects of the problem:

1. Coverage: we study whether Wikipedia/WordNet senses are representative of the meanings of a word in search results for one word queries.

2. Web Search diversity: We estimate search results diversity using our sense inventories.

3. Sense frequencies: we estimate Web sense frequencies from currently available information.

4. Classification: we classify Web pages according to our sense inventories, using the results to promote diversity by re-ranking search results.

In order to address these issues, we built a corpus consisting of 40 nouns and 100 Google search results per noun, manually annotated with the most appropriate WordNet and Wikipedia senses. Section 4.2 describes how this corpus has been created, and in Section 4.3 we discuss WordNet and Wikipedia coverage of search results according to our test bed. As this initial results clearly discard WordNet as a sense inventory for the task, the remainder of the Chapter mainly focuses on Wikipedia. In Section 4.4 we estimate search results diversity from our test bed, finding that the use of Wikipedia could substantially improve diversity in the top results. In Section 4.5 we use the Wikipedia internal link structure and the number of visits per page to estimate relative frequencies for Wikipedia senses, obtaining an estimation which is highly correlated with actual data in our test bed. Finally, in Section 4.6 we discuss a few strategies to classify Web pages into word senses, and apply the best classifier to enhance diversity in search results.

4.2 Test Bed

Our goal is to compare WordNet and Wikipedia as resources to distinguish alternative meanings of one-word queries in Web search results. Thus we need to create a corpus comprising:

1. A set of polysemic nouns, large enough to be representative.
2. An inventory of Wikipedia senses for the nouns.
3. An alternative inventory of senses (WordNet senses)
4. Lexical information associated with the senses, extracted from Wikipedia
5. A set of documents associated with the nouns as search results provided by a Web search engine
6. A gold standard, manual classification of the documents according to the senses in WordNet and Wikipedia.

4.2.1 Set of Nouns

One of the most critical steps for building our test set was choosing the set of words to be considered. We were looking for words susceptible to form a one-word query for a Web search engine, and therefore we wanted to focus on nouns used to denote one or more named entities. At the same time, we wanted to have some degree of comparability with previous research on Word Sense Disambiguation, which pointed us to noun sets used in Senseval/SemEval evaluation campaigns[3]. Our budget for corpus annotation was enough for two people-month, which limited us to handle 40 nouns (usually enough to establish statistically significant differences between WSD algorithms, although obviously limited to reach solid figures about the general behavior of words in the Web).

[3] http://senseval.org

With these arguments in mind, we decided to choose: (i) the 15 nouns from the Senseval-3 lexical sample dataset previously employed by [Mihalcea, 2007] in a related experiment (see Section 2.3.2); (ii) 25 additional nouns satisfying two conditions: all being highly ambiguous and all being names for music bands in one of their senses (not necessarily the most salient). Figure 4.1 shows the two sets chosen.

Figure 4.1: Nouns selected for our test bed

Senseval-3 Subset

`argument arm atmosphere bank degree difference disc image paper party performance plan shelter sort source`

Band Names

`amazon apple camel cell columbia cream foreigner fox genesis jaguar oasis pioneer police puma rainbow shell skin sun tesla thunder total traffic trapeze triumph yes`

4.2.2 Sense Inventories

For each noun, we looked up all its possible senses in WordNet 3.0 and in Wikipedia (using Wikipedia disambiguation pages). Note that for a conventional dictionary, a high ambiguity sometimes indicate an excess of granularity; for an encyclopedic resource such as Wikipedia, however, it is usually an indication of larger coverage.

Wikipedia senses

Generating a Wikipedia sense inventory involved seeking for senses through the world's largest encyclopedia. One possible strategy was to directly acquire the Wikipedia entries containing explicitly the noun, but we decided to use the Wikipedia disambiguation pages for each noun, as they are themselves a collaborative selection of senses. [Mihalcea, 2007] suggests the possibility of extracting lexical information from the Wikipedia disambiguation pages, but discards this option, whereas [Sanderson, 2008] opts for disambiguation pages as we do.

It should be noted that the Wikipedia disambiguation pages are not sense inventories; potential senses are presented in unordered lists, which are organized by arbitrary criteria, and do not follow strict formatting or conceptual rules (although there are some structures that are commonly used ([Sanderson, 2008])). Processing the disambiguation page for a given noun involves accepting some elements of the lists as actual senses and discarding others (e.g *Argumentation theory*, for *argument*). We filtered senses fitting one of these formats:

- Plural occurrences of the noun N_i.
- For any three words w_i, w_j, w_k, instances of "w_i N_i", "w_i w_j N_i" and "w_i w_j w_k N_i". (definitions and possible specializations) (e.g. for the noun *argument*, we stored *argument*, *deductive argument, heuristic argument*)
- Instances of "N_i, w_i". (e.g. for *bank, "Banks, Alabama"*)
- Instances of N_i (w_i). (e.g. for *circuit, "circuit (band)"*).
- Instances of a unique w_i, different of N_i. (e.g. for *atmosphere, "mood"*); w_i is interpreted as a synonym of N_i.
- Redirected items, if N_i appears explicitly and at least one of the previous conditions happens to be true.

WordNet Senses

The process is simpler for WordNet senses. We obtained the list of senses from WordNet 3.0 and we enriched these senses with their glosses (definitions of the senses in WordNet), to generate descriptors for them (for the manual annotation process).

Table 4.1 displays the number of senses acquired per noun. The average is 22 for Wikipedia and 4.25 for WordNet senses; Wikipedia is five times larger than WordNet.

As an example, Figure 4.2 shows Wikipedia senses for the noun *camel*, together with their definitions. We did not get lexical information for the sense *camel (paint)*, as there was neither a Wikipedia entry nor a definition in the disambiguation page for it. Figure 4.3 displays the only WordNet sense for camel.

4.2.3 Information about Wikipedia Senses

For each sense, we considered several types of information (mostly internal to Wikipedia):

- Wikipedia entry for the sense: the basic element in Wikipedia is the article, a page which explains the concept that gives title to it. Most senses listed in disambiguation pages become titles of Wikipedia articles providing a detailed description of the sense, although a small part of them do not have a dedicated Wikipedia entry or have an incomplete description.

  In Figure 4.2 Wikipedia senses for *camel* are shown. For the first seven senses, the first sentence of the corresponding article in which the noun *camel*

	#Senses	
	Wikipedia	**WordNet**
amazon	19	4
apple	11	2
argument	23	7
arm	8	6
atmosphere	13	6
bank	19	10
camel	10	1
cell	17	7
columbia	57	5
cream	6	3
degree	22	7
difference	5	5
disc	12	4
foreigner	9	2
fox	139	7
genesis	33	2
image	29	9
jaguar	21	1
oasis	26	2
paper	14	7
party	14	5
performance	8	5
pioneer	35	2
plan	19	3
police	15	1
puma	18	1
rainbow	43	2
shell	19	10
shelter	19	5
skin	25	6
sort	5	4
source	32	9
sun	54	5
tesla	8	2
thunder	16	3
total	6	2
traffic	13	4
trapeze	5	1
triumph	23	2
yes	12	1
Total	**882**	**170**

Table 4.1: Number of senses for the test nouns: Wikipedia vs WordNet

Figure 4.2: Wikipedia senses and definitions for *camel*

Apache Camel

Apache Camel is a rule-based routing and mediation engine which provides a POJO based implementation of the Enterprise Integration Patterns using an API (or declarative Java Domain Specific Language) to configure routing and mediation rules.

Camel (band)

Camel are an English progressive rock band formed in 1971.

Camel (cigarette)

Camel is a brand of cigarettes that was introduced by American company Reynolds Tobacco in the summer of 1913.

Camel

Camels are even-toed ungulates within the genus Camelus.

River Camel

The River Camel is a river in Cornwall, UK.

Sopwith Camel

The Sopwith Camel was a British World War I unique-seat fighter biplane, famous for its manoeuvrability.

Camel (album)

Camel is the first studio album by English progressive rock band Camel. It was released in 1973.

JISC infoNet Camel

Collaborative Approaches to the Management of e-Learning, a UK-based e-learning project and model in higher and further education for an intentional community of practice

CAML

CAML (Collaborative Application Markup Language) is an XML based markup language used with the family of Microsoft SharePoint technologies.

Camel (paint)

Figure 4.3: WordNet senses for *camel*

Camel

Cud-chewing mammal used as a draft or saddle animal in desert regions

explicitly appears is considered as definition. For the eighth and ninth senses, the noun *camel* was not explicitly mentioned in the corresponding Wikipedia entry, and then we consider the definition provided by the disambiguation page. Finally, for the tenth sense, there was neither a Wikipedia entry, nor a definition in the disambiguation page, as we have discussed before.

- Incoming links: Wikipedia has a complex structure of hyperlinking connecting articles by their titles. Such internal structure can be exploited by collecting, for each sense, the articles that link to the sense entry, which can be seen as manually disambiguated instances of the word.

- Outgoing links: we found two kinds of outgoing links in Wikipedia entries

  - Internal: Links to other Wikipedia articles.
  - External: Links to other Web pages.

The size of Wikipedia entries and the number of outgoing links are really variable, ranging from one to thousand sentences and from zero (very unusual) to over a thousand links. We have empirically found that some of these features are related to the relevance in Web Search Results of the concept described; this point will be addressed in Section 4.5.

Processing each of these sources, we obtained four sets of information associated with the nouns in our test bed. Obviously, the usability of the different types of information provided for the senses was expected to be quite different, depending on the source. As a matter of fact, we decided to discard the training corpora generated by external links, after a manual inspection of the information generated.

4.2.4 Set of Documents

We retrieved the 150 first ranked documents for each noun, by submitting the nouns as queries to the Google Web search engine. Then, for each document, we stored both the snippet (text excerpt extracted by Google as a description of the document in relation to the query) and the whole HTML document.

This collection of documents contains an implicit new inventory of senses, based on Web searches, as top-ranked documents retrieved by a noun query should be strongly associated with some sense of the noun. For this reason we assumed a ”one sense per document” scenario. This assumption turned out to be correct except for a few exceptional cases, such as Wikipedia disambiguation pages.

4.2.5 Manual Annotation

The main goal of our test bed is to establish relations between WordNet/Wikipedia inventories and Web Search results. Therefore, we need to provide a manual annotation of each retrieved document in terms of our sense inventories.

We implemented an annotation interface which stored all documents and a short description for every WordNet and Wikipedia sense. The annotators had to decide, for every document, whether there was one or more appropriate senses in each of the dictionaries. They were instructed to provide usable annotations for 100 documents per name; if an URL in the list was corrupt or not available, it had to be discarded. We provided 150 documents per name to ensure that the figure of 100 usable documents per name could be reached without problems.

Each judge provided annotations for the 4,000 documents in the final data set. In a second round, they met and discussed their independent annotations together, reaching a consensus judgement for every document.

This is the information provided to the annotators:

1. The inventory of Wikipedia senses, enriched with short descriptions, extracted from the Wikipedia entry for the sense, together with the Wikipedia categories associated with the sense.
2. The inventory of WordNet senses, enriched with the gloss associated with the synset containing the word sense, together with the synset hypernyms.
3. The set of 150 retrieved documents for each noun, together with their snippets. The interface showed snippets by default; the annotators could make a decision based on the snippet alone, or click to inspect the full document.

Figure 4.4 displays the flowchart of the annotation process. Columns 2 and 4 of Table 4.2 show the number of clustered documents into Wikipedia and WordNet senses, respectively. The first six senses of camel in Figure 4.2, for instance, were all represented in search results.

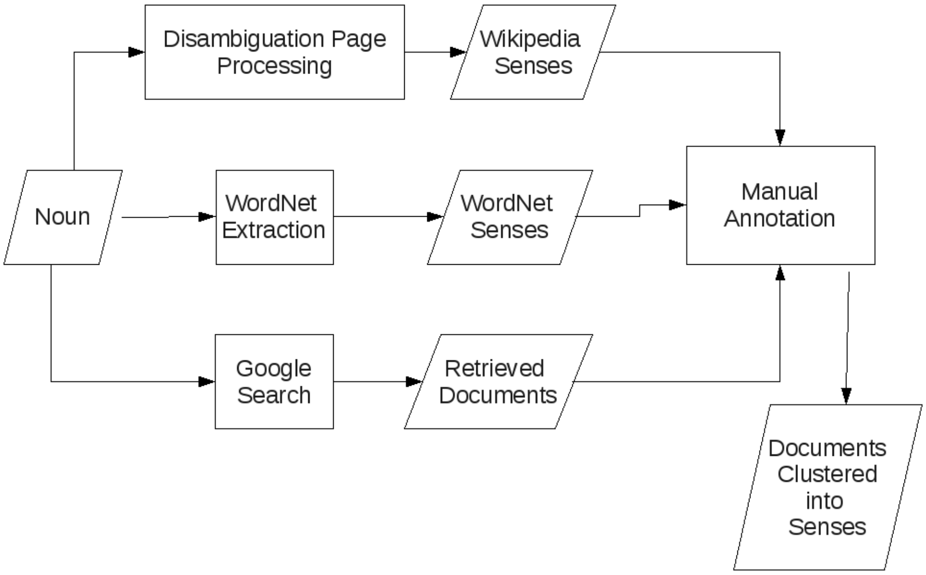


Figure 4.4: Annotation process: classification of search results into the appropriates senses

	Wikipedia		WordNet	
	#sens/#with docs	#clustered docs	#sens/#with docs	#clustered docs
amazon	19/7	91	4/2	5
apple	11/2	73	2/2	12
argument	23/14	64	7/6	25
arm	8/4	15	6/3	14
atmosphere	13/8	56	6/5	33
bank	19/3	86	10/2	77
camel	10/6	51	1/1	29
cell	17/6	82	7/3	70
columbia	57/12	30	5/5	28
cream	6/5	26	3/3	8
degree	22/10	77	7/4	49
difference	5/2	31	5/4	40
disc	12/7	53	4/3	34
foreigner	9/6	76	2/2	28
fox	139/12	61	7/1	3
genesis	33/6	33	2/1	11
image	29/6	50	9/1	40
jaguar	21/7	74	1/1	20
oasis	26/7	21	2/1	1
paper	14/6	71	7/3	60
party	14/5	75	5/2	70
performance	8/4	51	5/5	47
pioneer	35/8	25	2/1	7
plan	19/10	80	3/3	67
police	15/3	99	1/1	91
puma	18/10	63	1/1	14
rainbow	43/9	26	2/1	13
shell	19/6	72	10/2	10
shelter	19/10	70	5/5	38
skin	25/6	77	6/3	49
sort	5/3	63	4/3	77
source	32/8	35	9/3	25
sun	54/19	46	5/1	15
tesla	8/6	86	2/1	62
thunder	16/7	21	3/3	6
total	6/3	16	2/0	0
traffic	13/6	91	4/1	78
trapeze	5/3	51	1/1	37
triumph	23/8	53	2/0	0
yes	12/4	14	1/1	2
Total	**882/274**	**2235 (.56)**	**170/91**	**1295 (.32)**

Table 4.2: Detected senses and coverage of search results: Wikipedia vs WordNet

4.3 Coverage of Web Search Results: Wikipedia vs WordNet

Table 4.3 shows how well Wikipedia and WordNet cover the senses present in search results. We report each noun subset separately (*Senseval* and *Bands* subsets) as well as aggregated figures. Let us discuss these results.

Sense Relevance

274 Wikipedia and 91 WordNet senses received at least one document, which

Table 4.3: Coverage of search results: Wikipedia vs. WordNet

	Wikipedia		**WordNet**	
	#senses / with docs	**#clustered docs**	**#senses / with docs**	**# clustered docs**
Senseval set	**242/100**	**877 (59%)**	**92/52**	**696 (46%)**
Bands set	**640/174**	**1358 (54%)**	**78/39**	**599 (24%)**
Total	**882/274**	**2235 (56%)**	**170/91**	**1295 (32%)**

means that Wikipedia triples the number of relevant WordNet senses. The number of Wikipedia relevant senses is much larger in both subsets, but the difference is sharper for the Bands subset (where by definition at least one sense was really a named entity, which are rarely covered by WordNet). The ratios of relevant senses are, for Wikipedia: Senseval set, .41; Bands set, .27; and for WordNet: Senseval nouns .56, Band nouns .50. Note that the problem of WordNet is not that its senses are not relevant, but that many senses are missing.

The worst ratio of relevant senses is for Wikipedia senses on the Bands set, but at the same time this is the most productive set (for which more different senses receive annotations).

Search Results Coverage

As for the number of documents assigned to some word sense, for the Senseval nouns we reach 46% with WordNet and 59% with the Wikipedia inventory, whereas for the Bands set the percentages are 54% (Wikipedia) and 24% (WordNet). The coverage of Wikipedia is much better than WordNet in both subsets. The best coverage is for the Senseval set / Wikipedia senses case, indicating that the names in the Bands set seem more susceptible of being used unpredictably for all kinds of named entities.

In summary, the most relevant fact is that Wikipedia senses (as expected) cover much more search results (56%) than WordNet (32%). In the Bands subset (which should be more representative of plausible web queries), Wikipedia covers near 70% of the top ten results in average. This is an indication that it can indeed be useful for clustering purposes; and, at the same time, that it cannot be the only reference for clustering, because even the best scenario there would still be a 30% of the Web documents which cannot be associated with any of the Wikipedia senses.

4.4 Diversity in Google Search Results

Short queries are sometimes inherently ambiguous; the lack of context makes it impossible to apply explicit or implicit disambiguation. In these cases, search results should promote diversity, covering alternative senses in the top ranked results.

Once we know that Wikipedia senses are a representative subset of the actual

Web senses (covering more than a half of the first 100 documents retrieved by the search engine), we can test how well search results respect diversity in terms of this subset of senses.

In Tables A.1, A.2, A.3, A.4, A.5 (see Appendix A) we represent the results of the manual annotation for, respectively, the first 10, 25, 50, 75 and 100 documents retrieved by the Band nouns as queries to Google. The analogous results for Senseval nouns are detailed in Tables A.6, A.7, A.8, A.9, A.10.

Number of senses

Table 4.4 displays the aggregated number of senses (for all nouns) present in search results at different points in the ranking, using WordNet and Wikipedia as a reference. Table 4.5 shows the average number of senses per noun. These results are also graphically depicted in Figure 4.5. The data provides some hints about search results diversity:

	Wikipedia			WordNet		
	Band Nouns	**Senseval**	**Total**	**Band nouns**	**Senseval**	**Total**
First 10 Docs	72	48	120	19	24	43
First 25	111	72	183	25	39	68
First 50	139	82	221	34	43	77
First 75	164	95	259	37	49	86
First 100	174	100	274	39	52	91

Table 4.4: Search results: detected senses presented by subsets and aggregated

	Wikipedia			WordNet		
	Band Nouns	**Senseval**	**Total**	**Band nouns**	**Senseval**	**Total**
First 10 Docs	2.88	3.2	3	.76	1.6	1.08
First 25	4.44	4.8	4.58	1.16	2,6	1.7
First 50	5.56	5.47	5.53	1.36	2.87	1.93
First 75	6.56	6.33	6.48	1.48	3.27	2.15
First 100	6.96	6.67	6.85	1.56	3.47	2.28

Table 4.5: Search results: averages of detected senses per noun

- The average number of different Wikipedia senses in search results goes approximately from 3 (in the top ten documents) to almost 7 (in the top 100). An immediate conclusion is that there is room for improving diversity in the top ten results: only 3 out of 7 detected senses, in average, are represented in the first page of search results.

- Wikipedia is more representative of actual diversity in search results, and also has a more robust coverage: results for the Bands and the Senseval set are very similar. WordNet is less representative and also more sensitive to the type of nouns, with much lower figures for the Bands set than for the Senseval set.

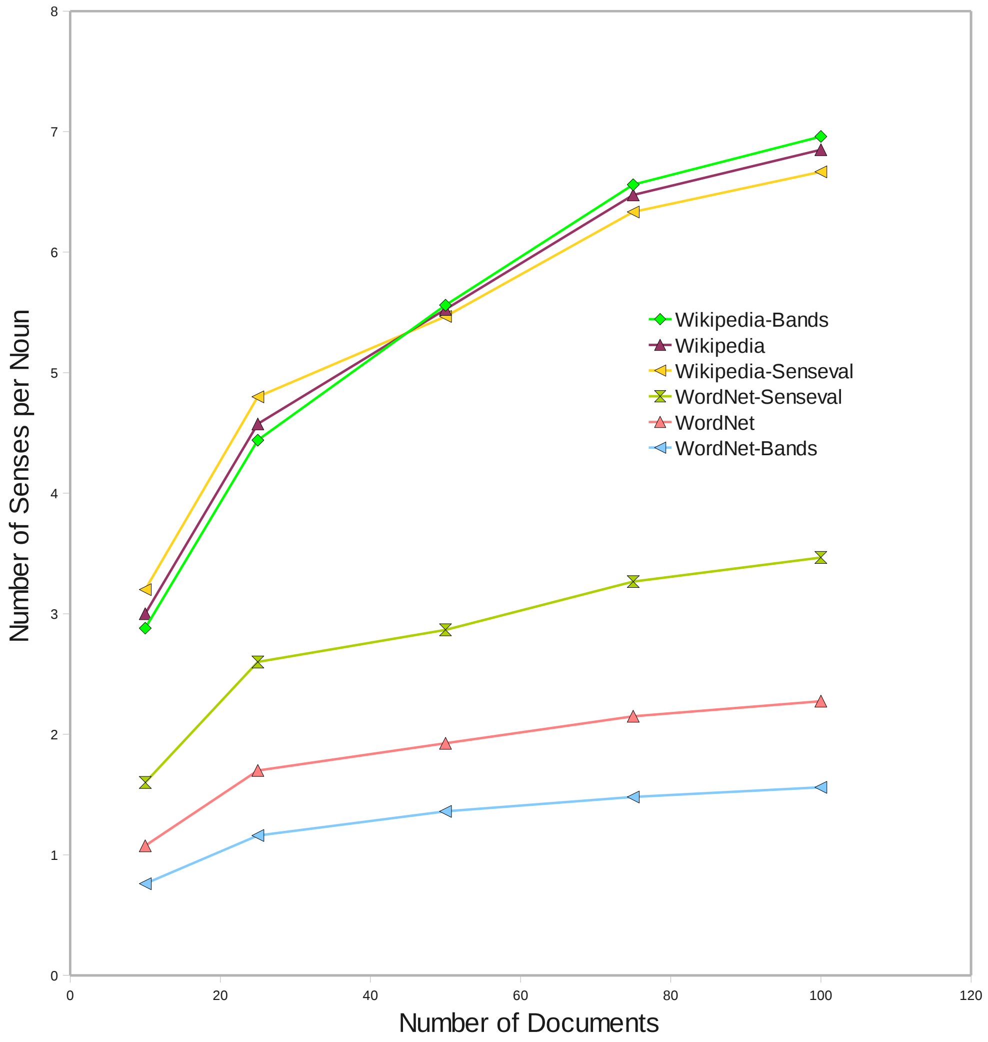


Figure 4.5: Average number of senses per noun at different points in the search engine rank.

Ratio of senses covered by search results

It is also interesting to look at the percentage of senses (listed in Wikipedia or WordNet) which are covered by search results at different points in the ranking. This is shown in Table 4.6 (details are given in Appendix A, Tables A.1 to A.10) and displayed as a graph in Figure 4.6. The situation now is the opposite: as WordNet compiles much less senses, its coverage in search results is better, starting with 34% in the top ten results and going up to 63% in the top 100 results. For Wikipedia, top ten search results only cover 21% of the senses, and the top 100 results include 41% of the listed senses. Note that, given the larger coverage of Wikipedia of actual senses in search results, it represents a much better estimation of search results diversity than WordNet.

	Wikipedia			**WordNet**		
	Band Nouns	**Senseval**	**Total**	**Band nouns**	**Senseval**	**Total**
First 10 Docs	**.21**	**.21**	**.21**	**.36**	**.30**	**.34**
First 25	**.28**	**.33**	**.30**	**.50**	**.50**	**.50**
First 50	**.33**	**.36**	**.34**	**.58**	**.54**	**.56**
First 75	**.37**	**.43**	**.39**	**.60**	**.60**	**.60**
First 100	**.38**	**.45**	**.41**	**.63**	**.64**	**.63**

Table 4.6: Search results: averages of ratios of detected senses

Coverage

We can also revisit how well Wikipedia and WordNet cover search results, studying how coverage changes at different points of the ranking. Table 4.7 displays the numbers and Figure 4.7 gives a graphical view of them. Note how the behavior of Wikipedia and WordNet is very different:

- Coverage of Wikipedia is maximal at the top ten results, and then decreases steadily up to the top 80 results. In other words, Wikipedia seems to have a better coverage of the senses that are more relevant for the search engine (and therefore are promoted to the first results).

- Coverage of WordNet follows a different pattern: it is much lower than Wikipedia at the top ten results (26% versus 62%) and then grows slowly up to 32% (which is, in any case, much lower than the 56% reached by Wikipedia). Therefore, coverage of WordNet is not only much lower in average, but also particularly weak for the most relevant senses according to the search engine. The most likely reason is that the search engine gives more importance to named entities (which are more likely interpretations of a noun as a Web Search query) than to standard, lexicographical senses of nouns.

Most frequent sense

Another relevant figure is the frequency of the most frequent sense for each word: in average, 63% of the pages in search results belong to the most frequent sense of the query word. This is roughly comparable with most frequent sense

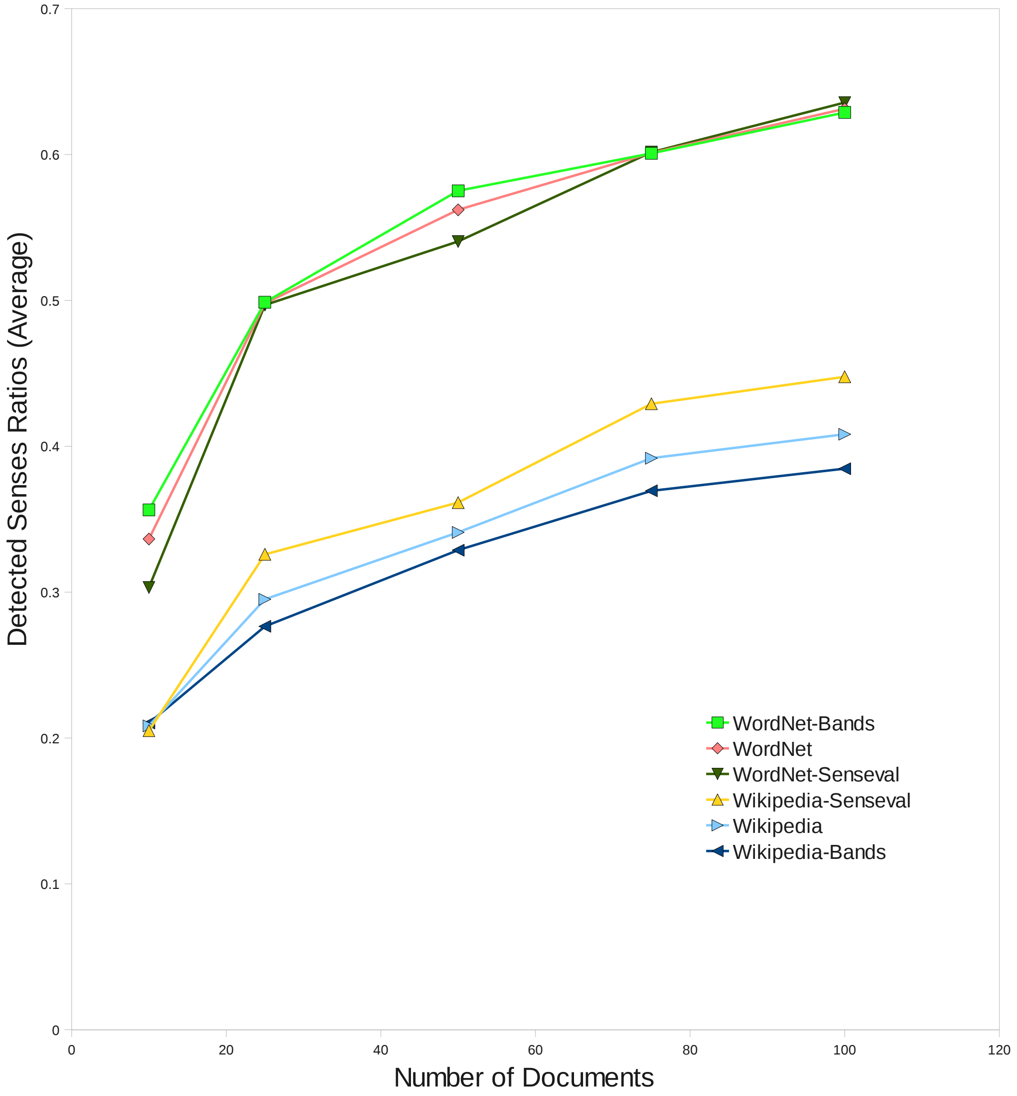


Figure 4.6: Ratio of Wikipedia/WordNet senses covered by search results

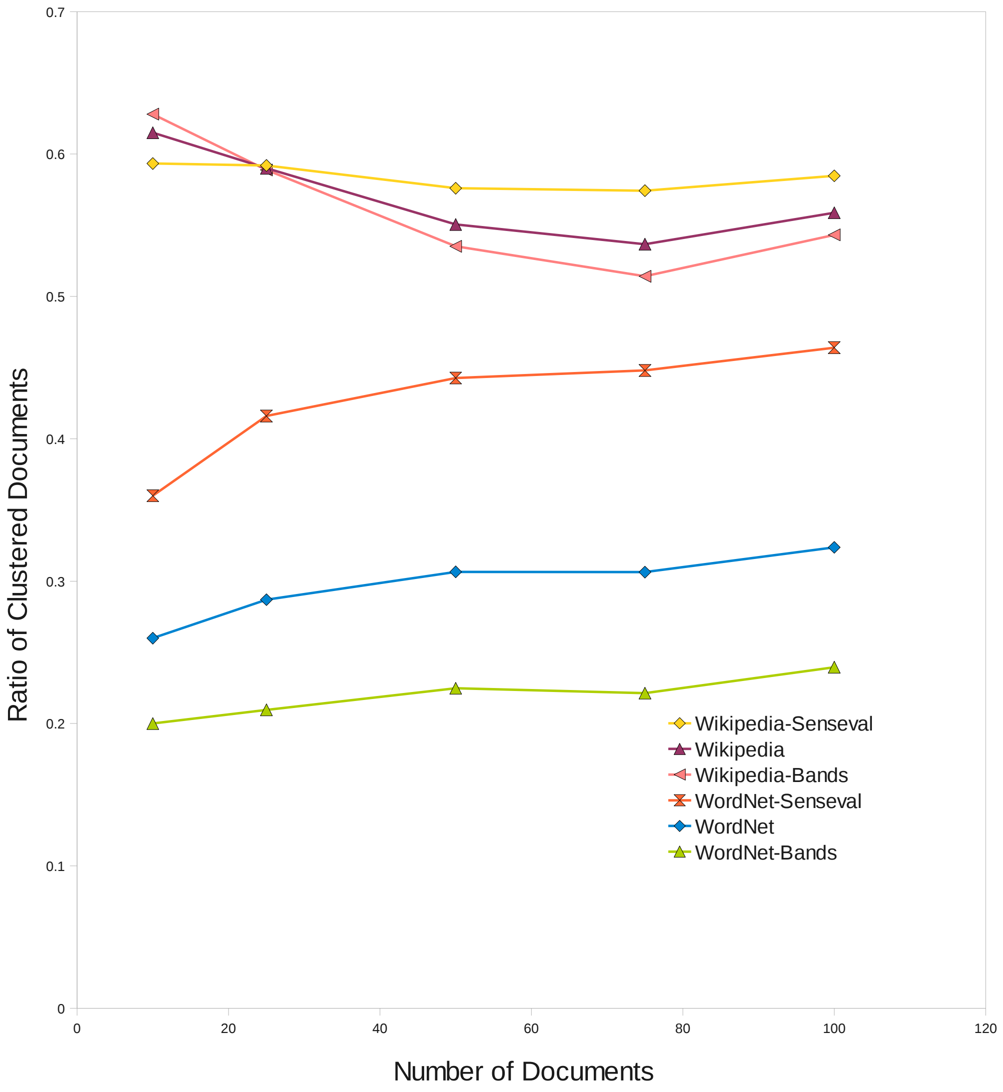


Figure 4.7: Coverage of search results by Wikipedia/WordNet senses

	Wikipedia			WordNet		
	Band Nouns	**Senseval**	**Total**	**Band Nouns**	**Senseval**	**Total**
First 10 Docs	**.63**	**.59**	**.62**	**.20**	**.36**	**.26**
First 25	**.59**	**.59**	**.59**	**.21**	**.42**	**.29**
First 50	**.54**	**.58**	**.55**	**.22**	**.44**	**.31**
First 75	**.51**	**.57**	**.54**	**.22**	**.45**	**.31**
First 100	**.54**	**.58**	**.56**	**.24**	**.46**	**.32**

Table 4.7: Search results: Coverage

figures in standard annotated corpora such as SemCor [Miller et al., 1993] and the Senseval/Semeval data sets, which again suggests that diversity does not play a major role in the current Google ranking algorithm.

Of course this result must be taken with care, because variability between words is high and unpredictable, and we are using only 40 nouns for our experiment. But what we have is a positive indication that Wikipedia could be used to cluster search results: potentially the first top ten results could cover at least 6.5 different senses in average, which would be a substantial growth.

4.5 Sense Frequency Estimators for Wikipedia

Wikipedia contains no explicit information about the relative importance of word senses. Such information, however, is crucial in a lexicon, because sense distributions tend to be skewed, and knowing them helps disambiguation algorithms. Fortunately, Wikipedia provides implicit information that can be exploited to estimate the relative importance of word senses. We use two sources of evidence: one internal (relative amount of incoming links for each word sense) and one external (relative number of visits for each word sense). Both estimations can be assessed using our test bed, which provides relative sense frequencies in web search results.

4.5.1 Estimators

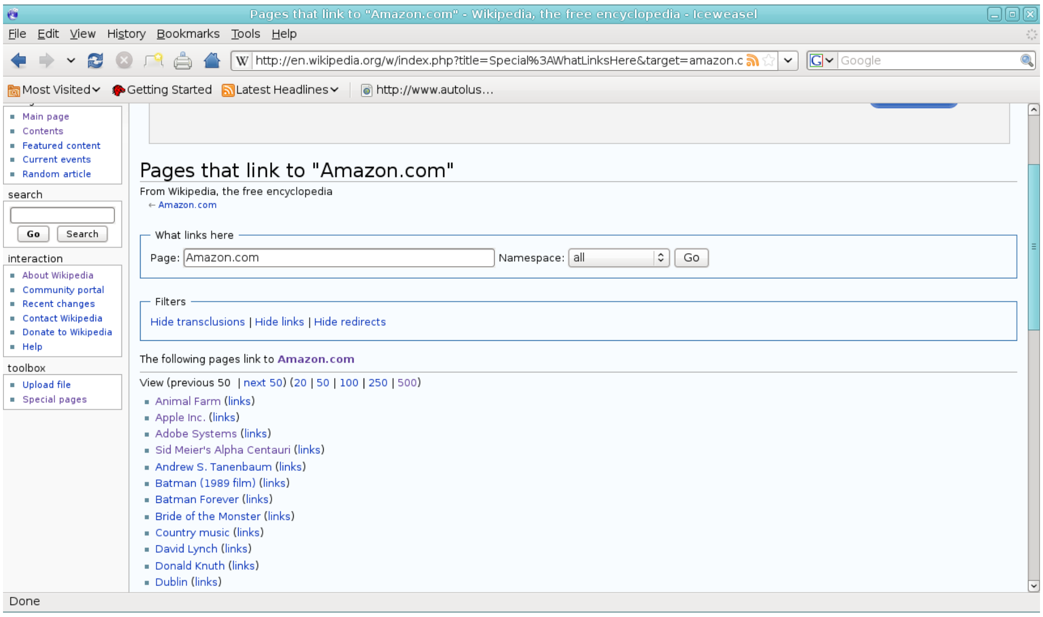
To assign relevance values to Wikipedia senses, we have used two sources of quantitative information about word senses: the first one, focused on the internal relevance of senses (how central they are to the internal Wikipedia structure), and the second one focused on external relevance, measuring how important a sense is for Wikipedia users.

Internal relevance

Wikipedia articles include hyperlinks to other articles (internal links), and a reasonable assumption is that the number of Wikipedia pages that link to a given sense[4], or number of incoming links, can be a good estimation of sense relevance

[4] Source http://en.wikipedia.org/wiki/Special:WhatLinksHere

(Figure 4.8 shows an example for *amazon.com*.)



Pages that link to "Amazon.com"

From Wikipedia, the free encyclopedia
← Amazon.com

What links here
Page: Amazon.com Namespace: all Go

Filters
Hide transclusions | Hide links | Hide redirects

The following pages link to **Amazon.com**

View (previous 50 | next 50) (20 | 50 | 100 | 250 | 500)

- Animal Farm (links)
- Apple Inc. (links)
- Adobe Systems (links)
- Sid Meier's Alpha Centauri (links)
- Andrew S. Tanenbaum (links)
- Batman (1989 film) (links)
- Batman Forever (links)
- Bride of the Monster (links)
- Country music (links)
- David Lynch (links)
- Donald Knuth (links)
- Dublin (links)

Figure 4.8: Incoming links example: all these Wikipedia pages contain at least one link to the article for *amazon.com*

External relevance

A way of measuring the relative impact of Wikipedia senses is using the number of visits received (available in http://stats.grok.se). Figure 4.9 shows an example. This information is an indicator of the popularity of a sense. Our source provides the number of visits to Wikipedia entries per month.

As a first preliminary analysis of the correlation of our estimators with actual sense frequencies in search results, we performed a preliminary analysis shown in Figures 4.10 and 4.11.

Incoming links

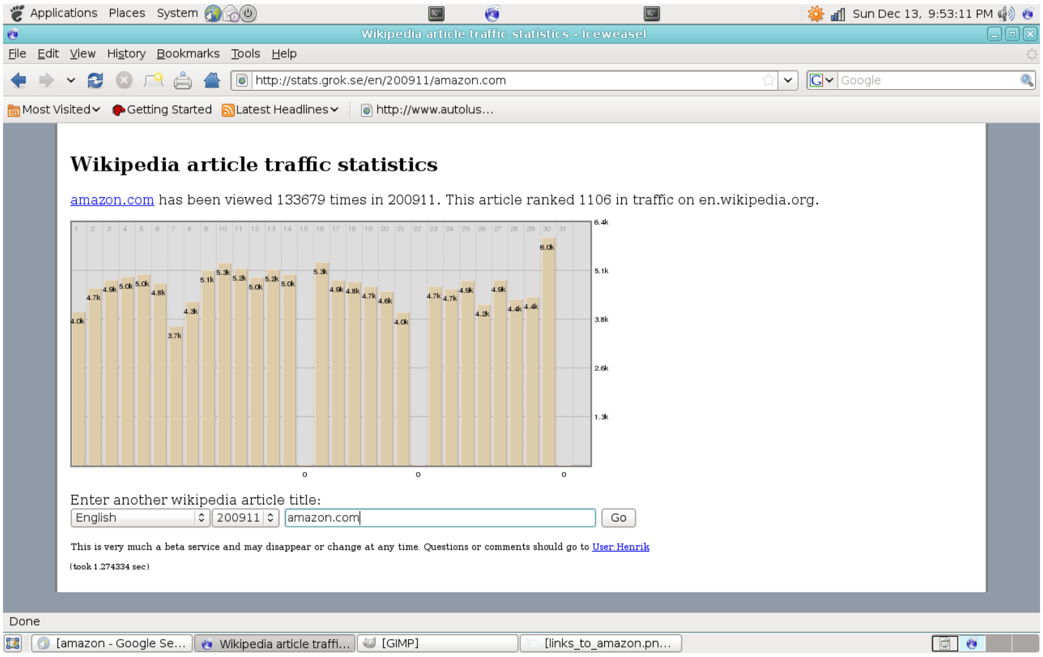
In Figure 4.10, upper bars represent the average number of incoming links for senses which are represented in the search results, and lower bars for the others.

Visits

In Figure 4.11, upper bars represent the average number of visits for senses present in the search results, and lower bars for the others.

The results show that senses present in search results receive much more incoming links and visits, supporting the existence of a strong connection between relevant senses in Wikipedia and senses relevant in (Google) search results.

In Table 4.8 we find another confirmation of the connection between relevance in Wikipedia and relevance in search results: the probability of finding a sense in the search results increases with the relative number of visits and with the relative number of incoming links. These relationships, as the table shows, are always satisfied for the first five positions of the visits and incoming links ranks.


Applications Places System
Sun Dec 13, 9:53:11 PM
Wikipedia article traffic statistics - Iceweasel
File Edit View History Bookmarks Tools Help
http://stats.grok.se/en/200911/amazon.com
Google
Most Visited
Getting Started
Latest Headlines
http://www.autolus...
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
4.0k 4.7k 4.9k 5.0k 5.0k 4.8k 3.7k 4.3k 5.1k 5.3k 5.2k 5.0k 5.2k 5.0k 5.3k 4.9k 4.8k 4.7k 4.6k 4.0k 4.7k 4.7k 4.9k 4.2k 4.9k 4.4k 4.4k 6.0k
6.4k
5.1k
3.8k
2.6k
1.3k
0
0
0
Done
[amazon - Google Se...
Wikipedia article traffi...
[GIMP]
[links_to_amazon.pn...

Wikipedia article traffic statistics

amazon.com has been viewed 133679 times in 200911. This article ranked 1106 in traffic on en.wikipedia.org.

Enter another wikipedia article title:

English 200911 amazon.com Go

This is very much a beta service and may disappear or change at any time. Questions or comments should go to User:Henrik

(took 1.274334 sec)

Figure 4.9: Visits to the Wikipedia article for *amazon.com*

	% senses represented in search results
Most visited sense	90
2nd most visited sense	87.5
3rd most visited sense	77.5
4th most visited sense	75
Most linked-to sense	86
2nd linked-to senses	85
3rd most linked-to senses	78.3
4th most linked-to senses	72.5

Table 4.8: Percentage of Wikipedia relevant senses which are represented in search results

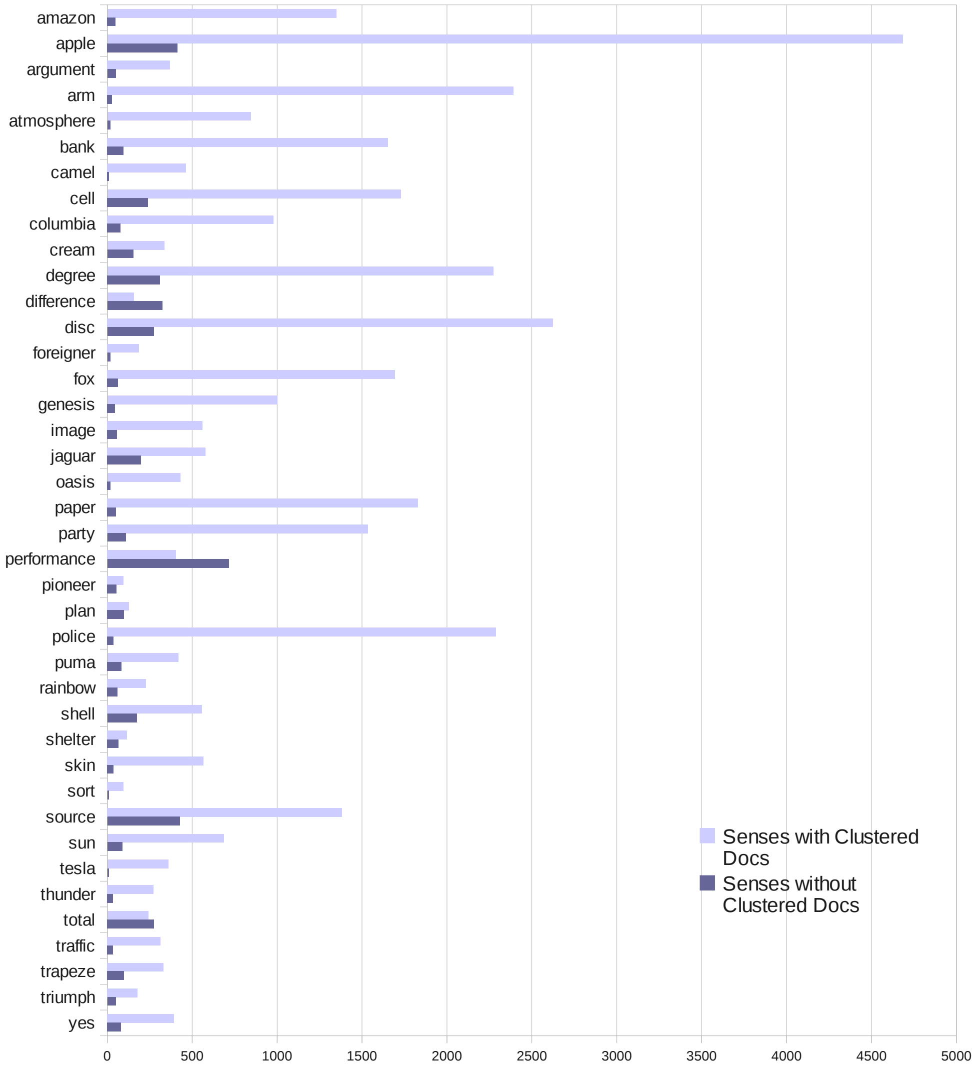


Figure 4.10: Incoming Links: upper bars represent the average number of incoming links for senses which are represented in the search results, and lower bars for the others.

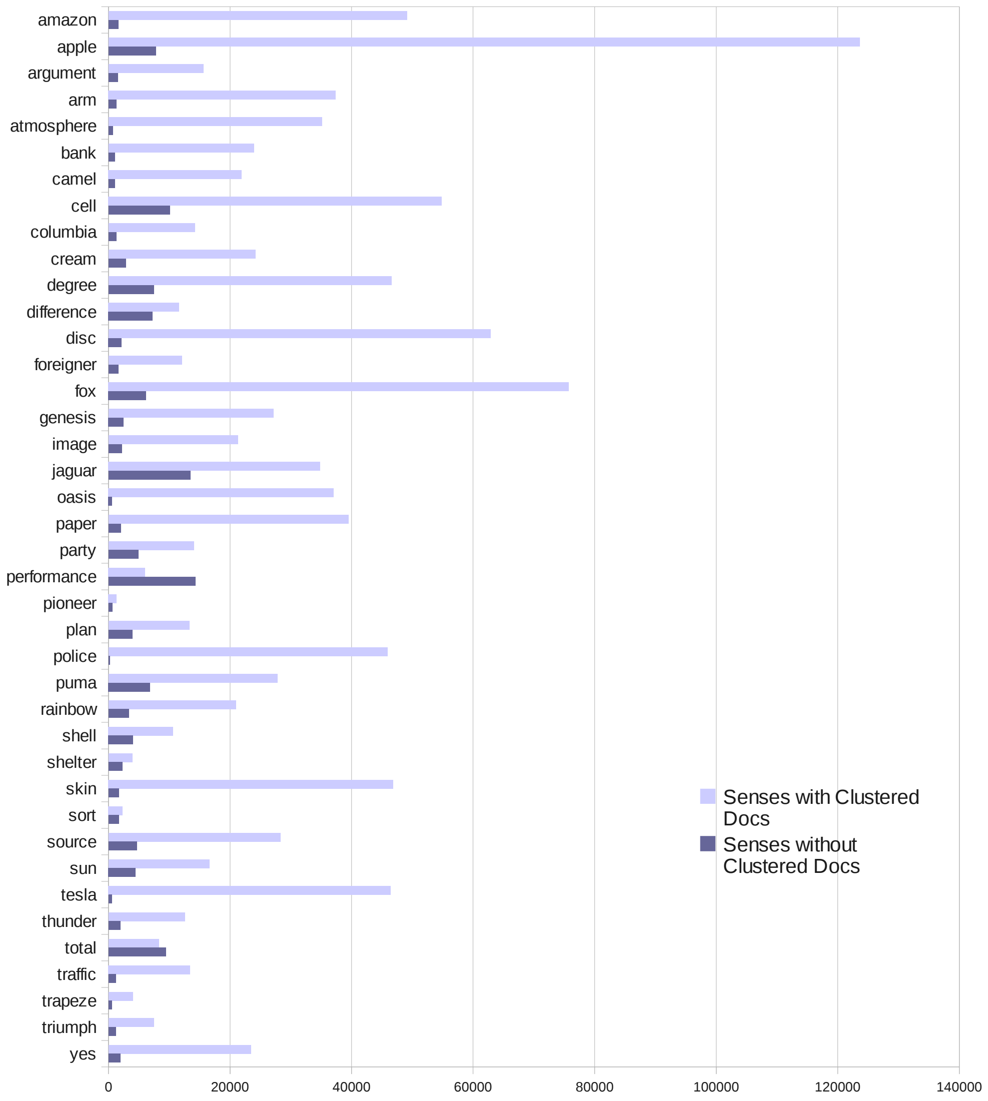


Figure 4.11: Visits in May 2009: upper bars represent the average for senses present in the search results, lower bars for the others.

Stability of Results

Wikipedia, as other collaboratively authored Web resources, is continuously enlarged and updated. But it is also a stable resource, where inclusions and modifications are censored by a large community of contributors. Hence we expect incoming links to be a relatively stable source of information, or at least not subject to random behavior with time.

Visits to Wikipedia articles, however, are subject to the shifting interests of web users, which can show drastic and unpredictable changes in time. Therefore, we need further analysis before assuming that the number of visits is a stable estimation of relative sense importance.

We performed a comparison of the number of visits received by our nouns during the months of May (Figure 4.11), June (Figure 4.12) and July (Figure 4.13) 2009. As shown in the figures, results are rather stable, except for one notorious exception: the number of visits to *Tesla*, which raised dramatically in July.

The explanation happened to be the date of birth of Nikola Tesla (10 July 1856). On July 10th, a special Google logo was released as a tribute to the scientist (Figure 4.14), which directed users to the corresponding Wikipedia entry (see the logo in Figure 4.15). This exception confirms that the number of visits is more vulnerable than the number of links but, at the same time, the absence of other alterations in our data is an indicator that the number of visits can indeed be a reasonable source of information. We decided to use an average of the number of visits in a period of at least three months.

4.5.2 Pairwise Frequency Correlations

For a given noun W_i, if sense $W_i s_j$ has higher frequency for Wikipedia relevance values (visits or links-to) than sense $W_i s_k$, can we assume that $W_i s_j$ will be more represented in search results than $W_i s_k$?

To answer this question, we define two equivalence relations:

For each noun W_i, and for each pair of senses $W_i s_j$ and $W_i s_k$

1. $f_l(W_i s_j) > f_l(W_i s_k) \iff f(W_i s_j) > f(W_i s_k)$
2. $f_v(W_i s_j) > f_v(W_i s_k) \iff f(W_i s_j) > f(W_i s_k)$

Checking whether such relationships hold for all sense pairs, we obtained the global results detailed in Table 4.9. The probability of both equivalences is $.66$. As $.50$ would be a random baseline, we can conclude that there is a indeed a correlation, but not a particularly strong one. Note that for this task we are considering all the Wikipedia senses detected for each noun, no matter whether they are represented in the search results (according to the manual annotation) or not.

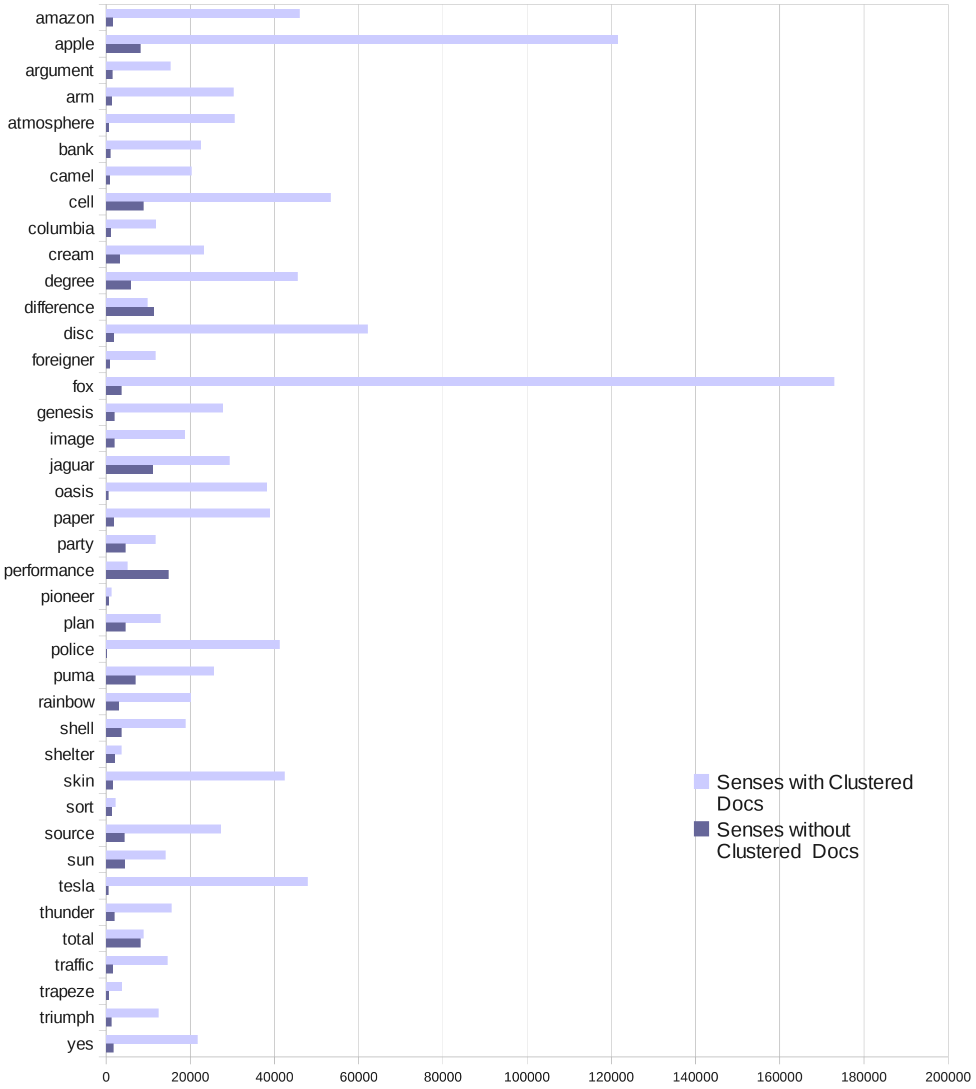


Figure 4.12: Average of visits to senses in June 2009

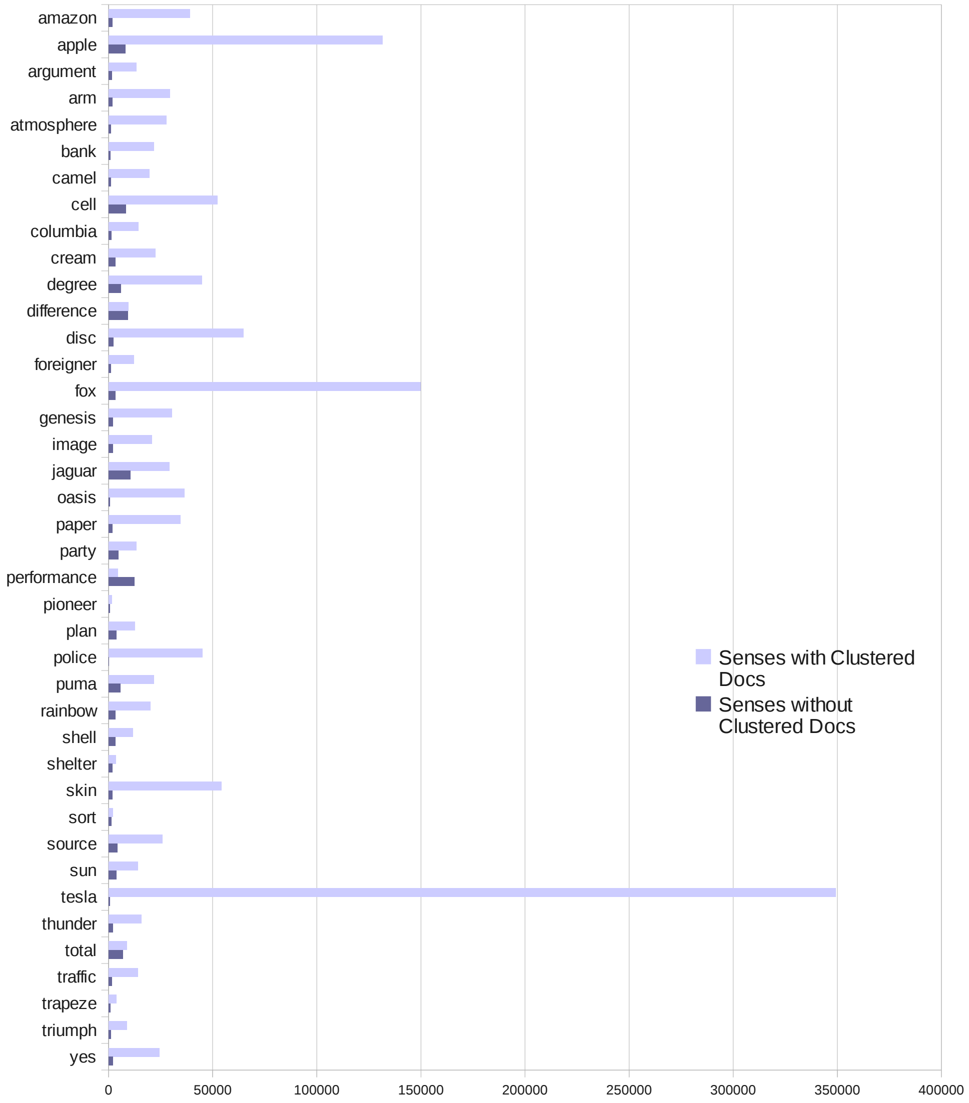


Figure 4.13: Averages of visits in July 2009: unexpected increase of visits for the noun tesla

Figure 4.14: Google logo in tribute to Nikola Tesla, linking to the Wikipedia entry for *Nikola Tesla*.

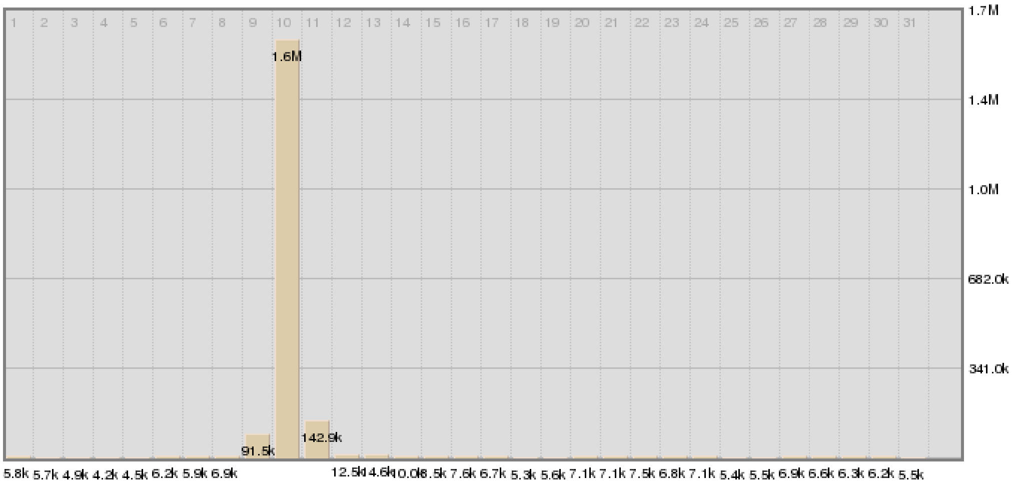


Figure 4.15: Accesses to tesla in July 2009

Equivalence between ordering criteria	True
Links and Documents	.66
Visits and Documents	.66

Table 4.9: Pairwise ordering equivalence probabilities (more frequency of links/visits ⟺ more frequency of documents)

4.5.3 Correlations between Frequency Estimators

The next step is a direct measurement of the correlation between the relative frequencies of our two estimators (incoming links and visits) and the relative frequencies observed in our gold standard. Let us consider, for each noun W_i and for each sense $W_i s_j$, the following three values:

1. Number of documents manually assigned to the sense $W_i s_j$. These documents have been retrieved submitting W_i as a Google search.

2. Number of incoming links to the sense $W_i s_j$. These links are available in *http://en.wikipedia.org/wiki/Special:WhatLinksHere*.

3. Number of visits received by the senses $W_i s_j$. (as listed in *http://stats.grok.se/*).

Then for each noun W_i, for each sense $W_i s_j$, we consider

Definition of frequencies

1. Frequency of documents associated to $W_i s_j$ in Google searches,

$$f_d(W_i s_j) = \frac{\#\text{docs assigned to } W_i s_j}{\#\text{docs retrieved for } W_i} \tag{4.1}$$

2. Frequency of links to $W_i s_j$,

$$f_l(W_i s_j) = \frac{\#\text{links to } W_i s_j}{\#\text{links to } W_i} \tag{4.2}$$

3. frequency of visits to $W_i s_j$,

$$f_v(W_i s_j) = \frac{\#\text{visits to } W_i s_j}{\#\text{visits to } W_i} \tag{4.3}$$

In order to assess the correlations between frequencies, for each noun W_i and for each sense $W_i s_j$, we use the standard linear regression correlation coefficient (equation 4.4),

Correlation coefficients

$$\rho = \frac{n(\Sigma x_i y_i) - (\Sigma x_i)(\Sigma y_i)}{\sqrt{n(\Sigma x_i^2) - (\Sigma x_i)^2}\sqrt{n(\Sigma y_i^2) - (\Sigma y_i)^2}} \tag{4.4}$$

which leads to the following correlation coefficients

1. Correlation between relative frequencies of word senses in the documents and relative number of incoming links (Equation 4.5):

$$\rho_{ld} = \frac{n(\Sigma f_d(W_i s_j) f_l(W_i s_j)) - (\Sigma f_d(W_i s_j))(\Sigma f_l(W_i s_j))}{\sqrt{n(\Sigma f_d(W_i s_j)^2) - (\Sigma f_d(W_i s_j))^2}\sqrt{n(\Sigma f_l(W_i s_j)^2) - (\Sigma f_l(W_i s_j))^2}} \tag{4.5}$$

2. Correlation coefficient between frequencies of documents and frequencies of visits to senses (equation 4.6)

$$\rho_{vd} = \frac{n(\Sigma f_d(W_i s_j) f_v(W_i s_j)) - (\Sigma f_d(W_i s_j))(\Sigma f_v(W_i s_j))}{\sqrt{n(\Sigma f_d(W_i s_j)^2) - (\Sigma f_d(W_i s_j))^2}\sqrt{n(\Sigma f_v(W_i s_j)^2) - (\Sigma f_v(W_i s_j))^2}} \tag{4.6}$$

Again, we considered all the Wikipedia senses detected for each noun, whether they are represented in search results or not.

Results

Results by noun are shown in Figure 4.16. If we look only at the sign of the correlation, the agreement between both indicators (visits/incoming links) is high; only in one case (*degree*) one of the correlations in positive and the other negative. If we look at the actual correlation values, in general both estimators have a similar tendency, with some exceptions: *arm, difference, fox, shelter* and *skin.*

To compute the global correlation, we considered the set of 40 nouns and weighted the separate correlation according to the number of senses per noun (for a total amount of 863). The global weighted average correlation between frequency of documents and frequency of visits is $.54$, whereas for incoming links is .71.

4.5.4 Rank Correlations

It is also interesting to measure the correlation between the sense orderings (ranks) induced by our estimators and the actual ranks in the test bed. The ranks are representative of the sense ordering that one expects to find in a sense inventory or dictionary.

We consider three ranks for Wikipedia senses: (i) ordering the senses by number of visits received, (ii) ordering them by number of incoming links, and (iii) ranking the senses by frequency of manually clustered documents (connected to Google searches). The formal definition of these ranks is as follows: for each noun W_i, for each sense $W_i s_j$,

1. Rank of visits for $W_i s_j$, $R_v(W_i s_j)$ is the position of $f_v(W_i s_j)$ in the set of frequencies $f_v(W_i s_k)$ (k = 1,2,...m senses for W_i) ordered by decreasing values.

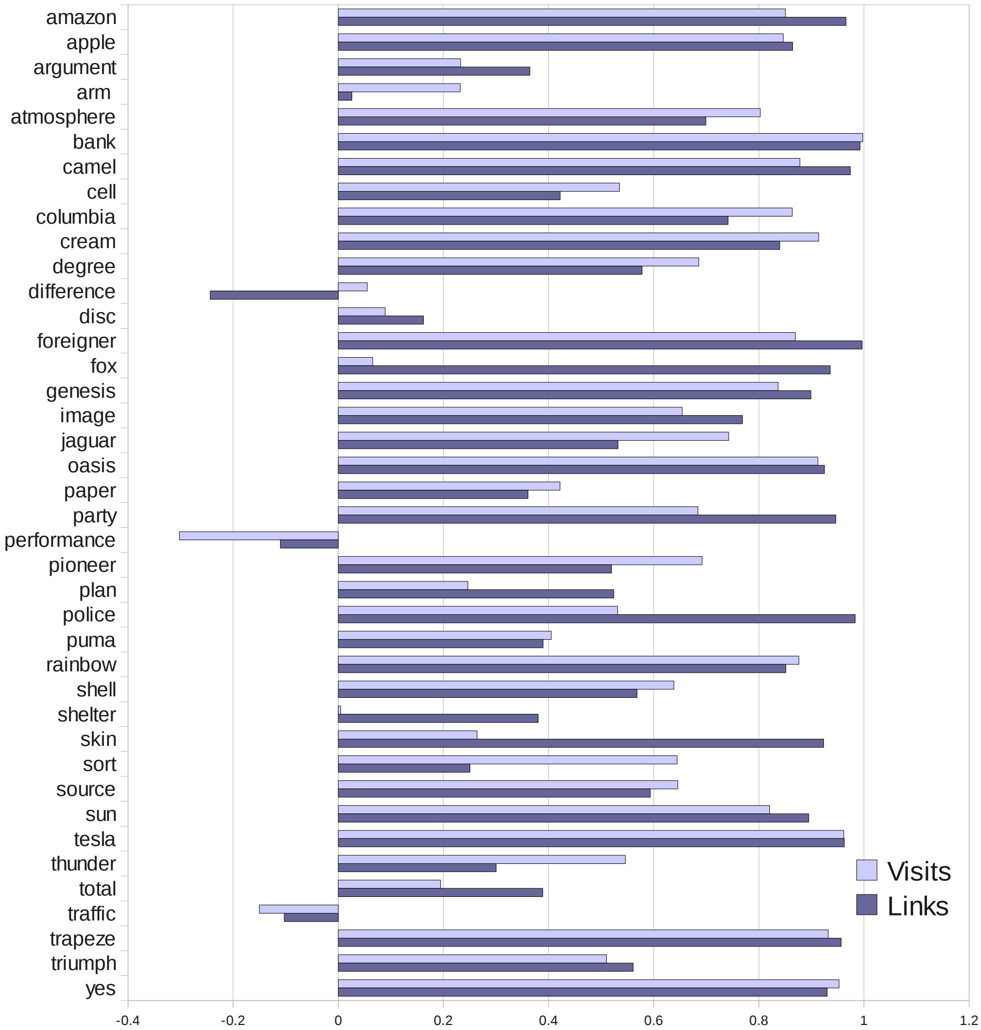


Figure 4.16: Averages correlation between relative frequency estimators (visits and links) and actual relative frequencies in our test bed.

2. Rank of links for W_is_j, $R_l(W_is_j)$ is the position of $f_l(W_is_j)$ in the set of frequencies $f_l(W_is_k)$ (k = 1,2,...m senses for W_i) ordered by decreasing values.

3. Rank of documents for W_is_j, $R_d(W_is_j)$ is the position of $f_d(W_is_j)$ in the set of frequencies $f_d(W_is_k)$ (k = 1,2,...m senses for W_i) ordered by decreasing values.

We have analyzed the correlations between both Wikipedia ranks and documents rank by implementing a commonly used non-parametric measure of statistical dependence between two variables, the Spearman's rank correlation coefficient (equation 4.7):

$$\rho = 1 - \frac{6\Sigma(x_i - y_i)^2}{n(n^2-1)} \tag{4.7}$$

where

- the n raw scores X_i, Y_i are converted to ranks x_i, y_i.
- n is the number of values in the ordered sets.
- tied ranks have been replaced by their average rank.

Hence, for each noun W_i, we computed the following coefficients:

1. For rank of documents, rank of visits correlations,

$$\rho_{vd} = 1 - \frac{6\Sigma(R_v(W_is_j) - R_d(W_is_j))^2}{n(n^2-1)} \tag{4.8}$$

2. For rank of documents, rank of links correlations,

$$\rho_{ld} = 1 - \frac{6\Sigma(R_l(W_is_j) - R_d(W_is_j))^2}{n(n^2-1)} \tag{4.9}$$

The detailed results by noun, for both correlations, are depicted in Figure 4.17. As in Section 4.5.3, we calculated the global correlations, by considering the 40 nouns and weighted the separate correlation according to the number of senses per noun, to finally obtain a global weighted average of .56 for rank of visits-rank of documents correlation, and a global weighted average of .56 for rank of incoming links-rank of documents correlation.

Comparing Figures 4.16 and 4.17, we can see that global results for frequencies correlations, and especially for documents-incoming links correlation (.71 of global

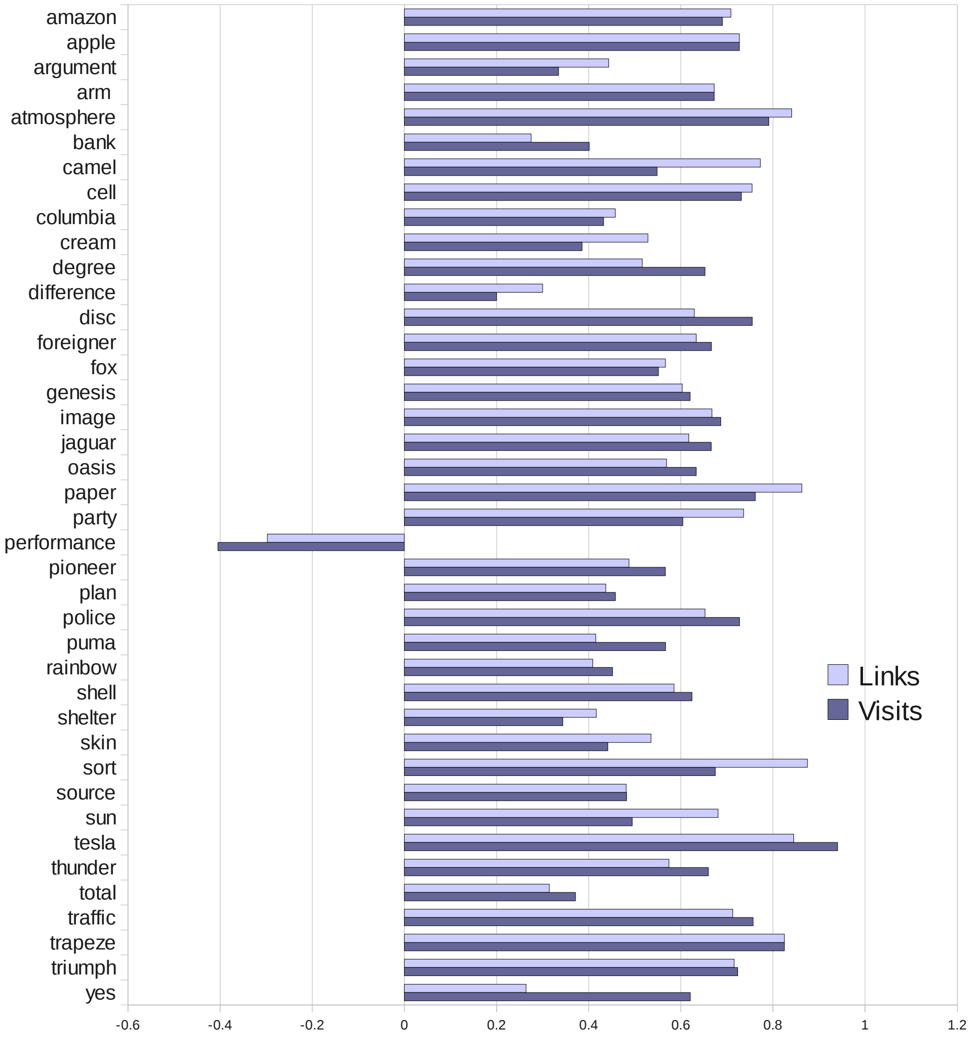


Figure 4.17: Average correlation between frequency ranks from our estimators (links and visits) and actual ranks in our test bed

accuracy) exceed the ones for rank correlations; Indeed, focusing on frequency correlations, for 19 nouns we reach a correlation higher than 80%, whereas for ranks, only 5 words present such a strong correlation. These evidence lead us to prefer relative frequencies rather than ranks as information to enrich Wikipedia senses.

4.5.5 Optimized Frequency Estimator for Wikipedia Senses

Although incoming links provide the best correlation with actual sense frequencies in our test bed, the number of visits has also proved to be a valuable source of information, and both can be combined into a more complete and robust estimator. Therefore, we have adopted a linear combination of both criteria, setting the weights empirically: we have tested weight pairs of the form $[n/10, 1 - n/10]$ with $n \in 0 \ldots 10$, finding the best fit at 0.9 for incoming links and 0.1 for visits. Therefore, our frequency estimator for every sense $W_i s_j$ belonging to word W_i is

$$V_{l-v}(W_i s_j) = .9 f_l(W_i s_j) + .1 f_v(W_i s_j) \tag{4.10}$$

Figure 4.18 shows the correlation results for this optimized frequency estimator. The global weighted average for all senses is 0.73. This weighted estimator improves the use of incoming links only, but not substantially (.73 vs .71 for incoming links only). Overall, we have an estimator which has a strong (although not perfect) correlation with the distribution of senses in our test bed. In Section 4.6, we will test its utility for disambiguation tasks.

4.6 Association of Wikipedia Senses to Web Pages

Wikipedia senses extracted from disambiguation pages are associated to Wikipedia articles, which contain valuable explicit and implicit information about a word sense (see Section 4.2). We now want to test whether this information can be used to associate each page in the search results with the appropriate word sense.

Word Sense Disambiguation algorithms typically decide the correct sense of an ambiguous noun occurrence in a given context. What we want to do is to decide the correct sense of the word for the whole document. Therefore, we can see our task as a Word Sense Disambiguation task under a ”one sense per discourse” hypothesis.

Using our test bed, we could experiment with supervised learning approaches, using a part of our manual assignments for training and a part for testing. But the resulting algorithms could not be used in practice, because it is unfeasible to build and maintain an annotated corpus for all nouns in a language. Therefore, we have restricted ourselves to (i) unsupervised classification strategies and (ii) supervised

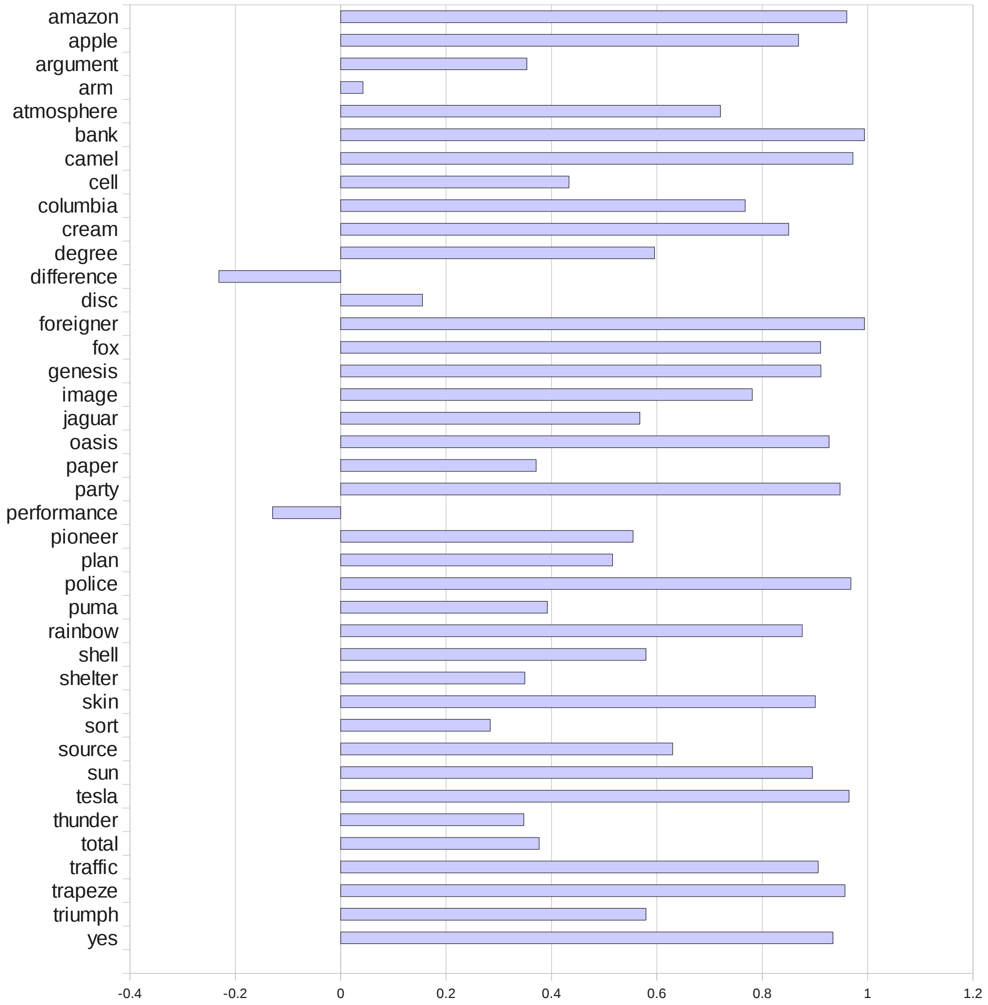


Figure 4.18: Optimized frequency estimator: correlation with actual sense frequencies

learning strategies which only use information that can be automatically extracted from Wikipedia, and does not involve manual creation of training samples.

Given a Web page p returned by the search engine for the query w, and the set of senses $w_1 \dots w_n$ listed in Wikipedia, the task is to assign the best candidate sense to p. We have considered two different techniques:

- An Information Retrieval approach, where the documents and the Wikipedia pages are represented using a Vector Space Model (VSM) and compared with a standard cosine measure. This is a basic approach which, if successful, can be used efficiently to classify search results.

- An approach based on a state-of-the-art supervised WSD system, extracting training examples automatically from Wikipedia content.

We also computed two baselines:

- A random assignment of senses (precision is computed as the inverse of the number of senses, for every test case). For each noun, we calculated the probability of success, by randomly selecting a sense, and then we took the average for the 40 nouns (the first column of Table 4.2 *senses with docs* shows the distribution of the 274 considered senses). This random strategy has a precision of .19 in our corpus.

- A most frequent sense heuristic which uses our estimation of sense frequencies and assigns the same sense (the most frequent) to all documents. For each noun, we assigned to all documents the sense with maximum optimized frequency estimator value (see Section 4.5.5), and then we compared the results to the manual assignations. In this case we reached a precision of $.46$. Table 4.10 shows the most frequent sense, with the number of documents manually assigned to this sense.

Both are naive baselines, but it must be noted that the most frequent sense heuristic is usually hard to beat for unsupervised WSD algorithms in most standard data sets.

We discarded the documents with no Wikipedia sense assigned (clustered to *OTHERS*) to measure classification performance in our experiments, for a better comparison with WSD tasks (such as Senseval/Semeval competitions) which do not consider out-of-dictionary senses. That gives a restricted list of senses displayed in the first column of Table 4.2 (senses with docs).

We now describe each of these two main approaches in detail.

	#correct	#annotated	precision
amazon	69	91	.76
apple	62	73	.85
argument	2	61	.03
arm	1	13	.08
atmosphere	21	56	.38
bank	84	86	.98
camel	29	51	.57
cell	4	82	.05
columbia	13	30	.43
cream	17	26	.65
degree	21	77	.27
difference	27	31	.87
disc	13	53	.25
foreigner	45	76	.59
fox	21	61	.34
genesis	11	33	.33
image	22	32	.69
jaguar	22	73	.30
oasis	12	20	.60
paper	7	68	.10
party	44	74	.59
performance	16	51	.31
pioneer	0	25	.00
plan	60	77	.78
police	91	99	.92
puma	14	61	.23
rainbow	13	26	.5
shell	21	72	.29
shelter	24	67	.36
skin	49	77	.64
sort	2	63	.03
source	0	35	.00
sun	4	45	.31
tesla	58	86	.67
thunder	9	17	.53
total	0	16	.00
traffic	78	90	.87
trapeze	37	51	.73
triumph	28	39	.72
yes	9	14	.64
Total	1079	2178	.46 average

Table 4.10: Baseline precision by using the most frequent sense according to the combined-frequencies value

4.6.1 VSM Approach

For each word sense, we represented its Wikipedia page in a (unigram) vector space model, assigning standard tf*idf weights to the words in the document. idf weights are computed in three different ways:

1. Experiment **VSM** computes inverse document frequencies in the collection of retrieved documents (for the word being considered).
2. Experiment **VSM-GT** uses the statistics provided by the Google Terabyte collection [Brants and Franz, 2006], i.e. it replaces the collection of documents with statistics from a representative snapshot of the Web.
3. Experiment **VSM-mixed** combines statistics from the collection and from the Google Terabyte collection, following [Chen et al., 2009].

The document p is represented in the same vector space as the Wikipedia senses, and it is compared with each of the candidate senses w_i via the cosine similarity metric. The sense with the highest similarity to p is assigned to the document. In case of ties (which are rare), we picked the first sense in the Wikipedia disambiguation page (which in practice is like a random decision, because senses in disambiguation pages do not seem to be ordered according to any clear criteria).

We also tested a variant of this approach which uses the estimation of sense frequencies presented above: once the similarities were computed, we considered those cases in which two or more senses had a similar score (in particular, all senses with a score greater or equal than the sense with the highest score). In that cases, instead of using the small similarity differences to select a sense, we picked up the one which had the largest frequency according to our estimator. We applied this strategy to the best performing system, VSM-GT, resulting in experiment **VSM-GT+freq**.

4.6.2 WSD Approach

We used TiMBL ([Daelemans et al., 2001]), a state-of-the-art supervised WSD system which uses Memory-Based Learning. The key, in this case, is how to extract learning examples from Wikipedia automatically. As we have seen in Section 4.2.3, for each word sense, we basically have three sources of examples: (i) occurrences of the word in the Wikipedia page for the word sense; (ii) occurrences of the word in Wikipedia pages pointing to the page for the word sense; (iii) occurrences of the word in external pages linked in the Wikipedia page for the word sense.

After an initial manual inspection, we decided to discard external pages as being too noisy, and we focused on the first two options. We tried three alternatives:

- **TiMBL-core** uses only the examples found in the page for the sense being trained.
- **TiMBL-inlinks** uses the examples found in Wikipedia pages pointing to the sense being trained.
- **TiMBL-all** uses both sources of examples.

We also experimented with a variant of the approach that uses our estimation of sense frequencies, similarly to what we did with the VSM approach.

4.6.3 WSD-Based Algorithm

We followed these main steps:

1. Design models to transform Wikipedia pages into a set of training examples.
2. Design a model to transform documents in search results into a list of test sentences for the query word.
3. For each noun, train the TiMBL supervised WSD system with the examples acquired.
4. For each document, disambiguate each of the test sentences using TiMBL.
5. Design a way of using the output of TiMBL to assign a unique sense to each document in the search results.

Figure 4.19 shows the flowchart of the method. We now detail each of the steps.

Acquisition of training samples from Wikipedia pages

In Figure 4.19, sense representation is the process that turns the Wikipedia lexical information for the senses into the training examples. The Training corpus consists of textual information stored as raw data for each sense. As our aim was to exploit these data for disambiguation purposes, we had to convert them into a structured set of features which characterize the corresponding senses. Given a noun W_i and given a sense belonging to W_i, W_is_j, our representation process involves the following steps:

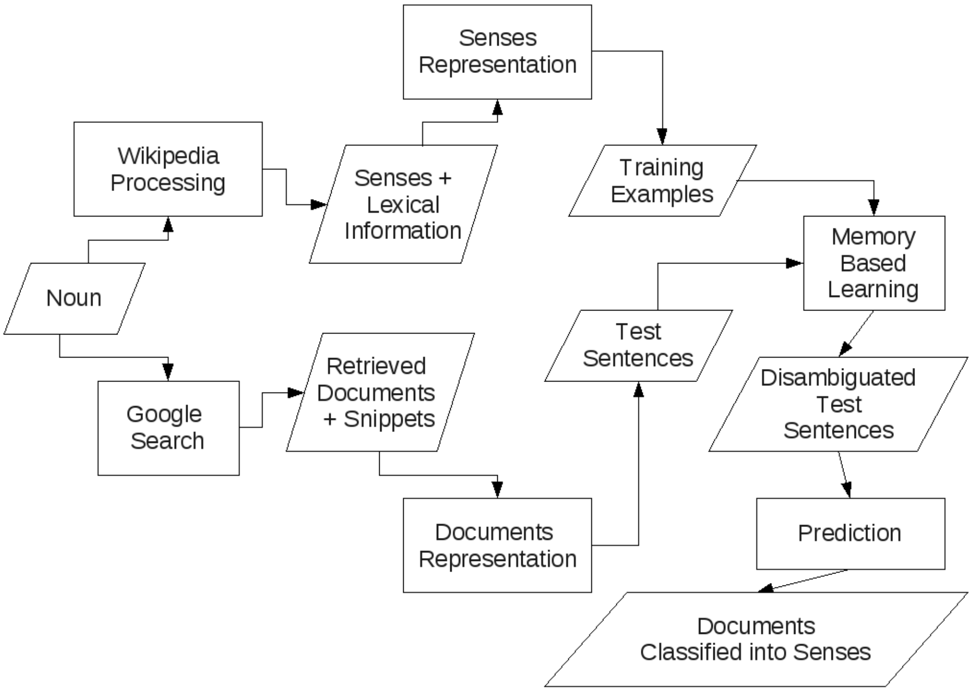


Figure 4.19: WSD-based assignment of senses to documents in search results

Wikipedia dump pre-processing We started from Wikiprep[5], a software developed by E. Gabrilovich and S. Markovitch that pre-processes the entire Wikipedia dump[6], and then we adapted this preprocessor to be applied to individual articles, extracting the text associated with Wikipedia pages for $W_i s_j$ as well as the text in the pages linking to the entry for $W_i s_j$.

Sentence selection We selected and stored the sentences which explicitly included W_i.

POS tagging The next step was to add part of speech tags to the stored sentences, to identify the nouns, verbs or adjectives near W_i, assuming that these words would be meaningful and closely related to $W_i s_j$

Training examples Assuming that every article containing W_i likely associates to a unique sense $W_i s_j$, we take Wikipedia articles as contexts for nouns. Hence, we decided to build training features by taking meaningful words located near W_i; more precisely, we ignored the stop words, collecting all the occurrences of

- (N or V or ADJ)+W_i+(N or V or ADJ) (Pattern 1)
- (N or V or ADJ)+(N or V or ADJ)+W_i (Pattern 2)
- W_i+(N or V or ADJ)+(N or V or ADJ) (Pattern 3)

Proceeding as described above, we generated a set of training features for each sense $W_i s_j$, (representation of the sense $W_i s_j$). This model is inspired in the SenseLearner semantic models [Csomai, 2005].

Figure 4.20 shows a part of the results for the sense *Amazon Basin* (amazon#1 in the Wikipedia sense inventory). We applied the process to our different training corpora (see Section 4.2), generating three training models (flowchart for them is shown in Figure 4.21):

Wikipedia page Generated using Wikipedia sense entries as Training corpus.

Incoming links Based on the training corpus built from the Wikipedia pages that link to each sense page.

Mixed The union of the two sets above.

[5] http://www.cs.technion.ac.il/ gabr/resources/code/wikiprep

[6] We used a dump of English Wikipedia, downloaded on December 3 2008

Figure 4.20: Example of sense representation: above, text from the Wikipedia entry for *Amazon Basin*; below, features extracted from the text

```
Amazon River basin
The Amazon Basin is the part of South America drained
by the Amazon River and its tributaries.
The South American rain forest of the Amazon is the largest
in the world, covering about 8,235,430 km 2 with dense tropical forest.
Not all of the big plant and animal life in the Amazon Basin
are known because of its huge unexplored areas.
One tropical fruit tree that is native to the Amazon is the abiu.
The Amazon Basin includes a diversity of traditional inhabitants
as well as biodiversity in both flora and fauna.
The Amazon basin has been continuously inhabited for more
than 12,000 years, since the first proven arrivals
of people in South America.
```

`amazon,river,basin,amazon#1` (Pattern 3)
`drain,amazon,river,amazon#1` (Pattern 1)
`america,drain,amazon,amazon#1` (Pattern 2)
`amazon,basin,part,amazon#1` (Pattern 3)
`amazon,river,tributary,amazon#1` (Pattern 3)
`forest,amazon,large,amazon#1` (Pattern 1)
`rain,forest,amazon,amazon#1` (Pattern 2)
`amazon,large,world,amazon#1` (Pattern 3)
`life,amazon,basin,amazon#1` (Pattern 1)
`animal,life,amazon,amazon#1` (Pattern 2)
`amazon,basin,know,amazon#1` (Pattern 3)
`native,amazon,abiu,amazon#1` (Pattern 1)
`tree,native,amazon,amazon#1` (Pattern 2)
`amazon,basin,include,amazon#1` (Pattern 3)
`amazon,basin,inhabit,amazon#1` (Pattern 3)

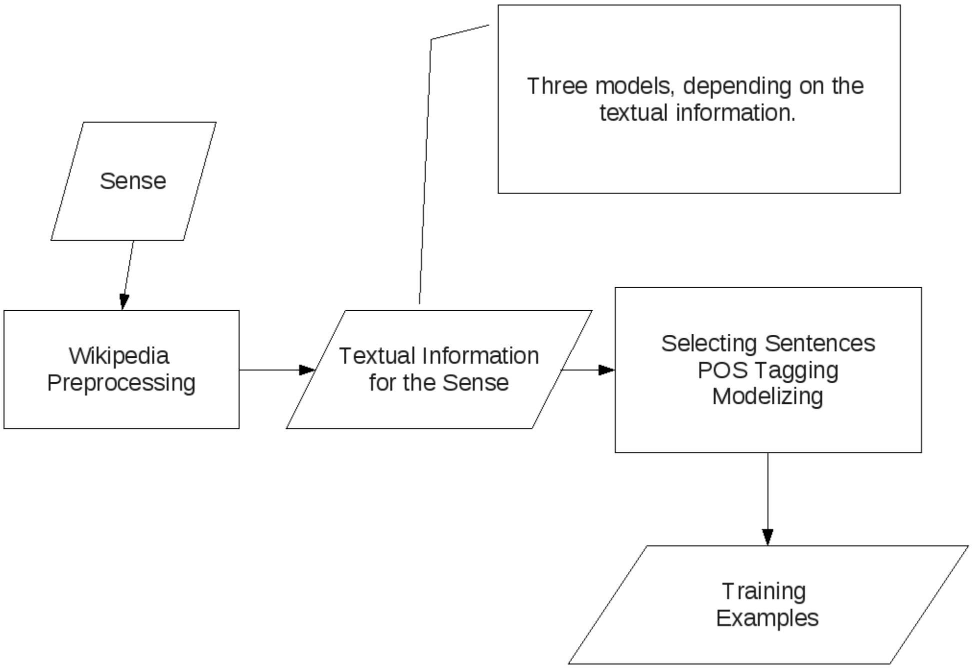


Figure 4.21: Training model: Acquisition of examples

Document Representation

We followed a similar procedure to represent the documents to be classified. We considered the documents retrieved for each noun as contexts for a specific sense, implicit in the document, as we did with Wikipedia articles containing the noun. In Figure4.19, document representation is the process that turns the information about the document into test sentences.

During the creation of the test bed (Section 4.2), we found two sources of lexical information available for each document, (i) the snippet description and (ii) the textual information in the whole HTML document. The snippet is supposed to highlight the contents most related to the query (i.e. the noun in our case), therefore we decided to add it to the document representation. Figure 4.22 shows the flowchart, from documents to test sentences, and Figures 4.23 and 4.24 a partial sight of a document example (document 014 for the query *amazon*) can be seen. Figure 4.25 shows the features extracted for this document. In Figures A.1 and A.2 the complete textual information extracted from these partial sights is presented. Note that only the lines with the noun *amazon* explicitly mentioned are considered for the generation of features.

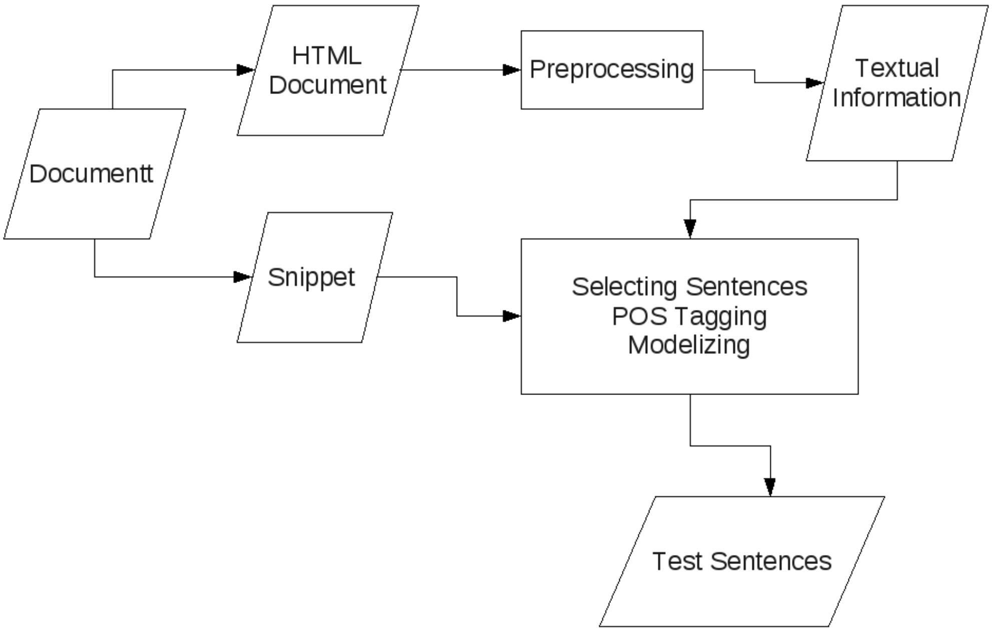
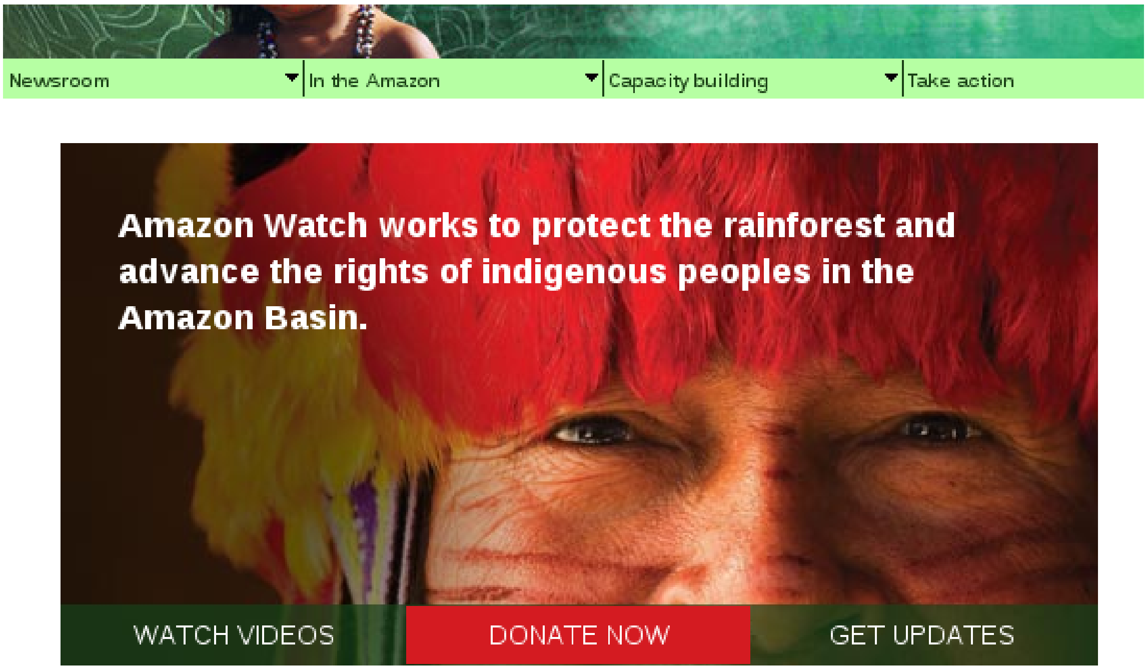
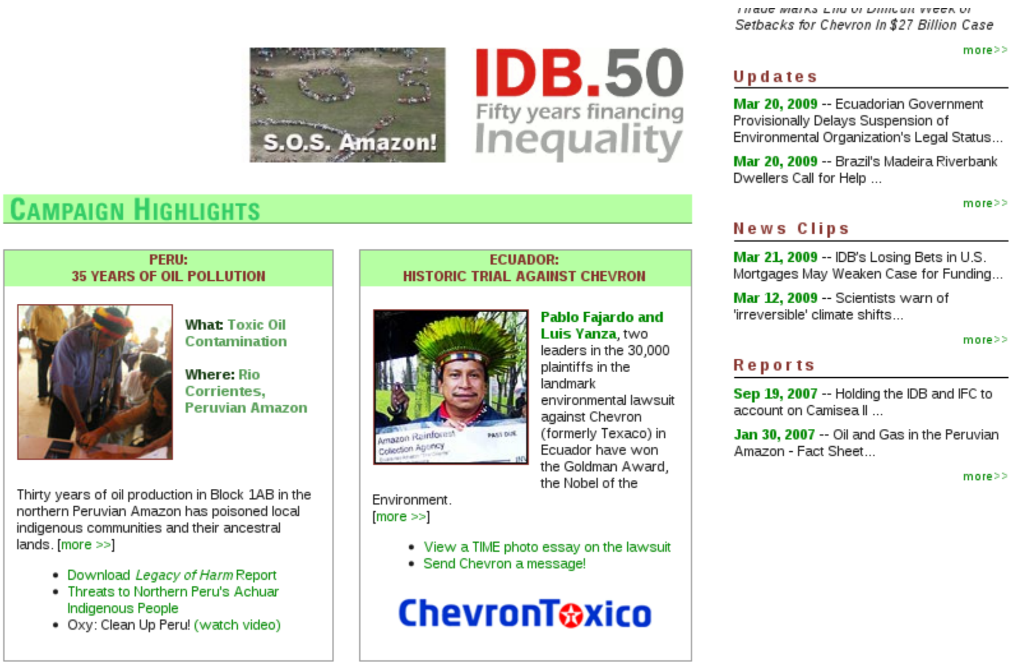


Figure 4.22: Test Model: representation of the document



Amazon Watch works to protect the rainforest and advance the rights of indigenous peoples in the Amazon Basin.

Figure 4.23: Document 014 for *amazon* (partial sight)



Trade Marks End of Difficult Week of Setbacks for Chevron In $27 Billion Case

more>>

CAMPAIGN HIGHLIGHTS

PERU:
35 YEARS OF OIL POLLUTION

What: Toxic Oil Contamination

Where: Rio Corrientes, Peruvian Amazon

Thirty years of oil production in Block 1AB in the northern Peruvian Amazon has poisoned local indigenous communities and their ancestral lands. [more >>]

- Download *Legacy of Harm* Report
- Threats to Northern Peru's Achuar Indigenous People
- Oxy: Clean Up Peru! (watch video)

ECUADOR:
HISTORIC TRIAL AGAINST CHEVRON

Pablo Fajardo and Luis Yanza, two leaders in the 30,000 plaintiffs in the landmark environmental lawsuit against Chevron (formerly Texaco) in Ecuador have won the Goldman Award, the Nobel of the Environment. [more >>]

- View a TIME photo essay on the lawsuit
- Send Chevron a message!

Updates

Mar 20, 2009 -- Ecuadorian Government Provisionally Delays Suspension of Environmental Organization's Legal Status...

Mar 20, 2009 -- Brazil's Madeira Riverbank Dwellers Call for Help ...

more>>

News Clips

Mar 21, 2009 -- IDB's Losing Bets in U.S. Mortgages May Weaken Case for Funding...

Mar 12, 2009 -- Scientists warn of 'irreversible' climate shifts...

more>>

Reports

Sep 19, 2007 -- Holding the IDB and IFC to account on Camisea II ...

Jan 30, 2007 -- Oil and Gas in the Peruvian Amazon - Fact Sheet...

more>>

Figure 4.24: Document 014 for *amazon* (partial sight)

Figure 4.25: Example of document representation: document 014 for *amazon*

Document text (partial sight)

```
Newsroom | In the Amazon | Capacity building | Take action | About us
Amazon Watch works to protect the
peoples in the Amazon Basin.
Peruvian Amazon environmental lawsuit images]
Amazon has poisoned Goldman Award, the Nobel Environmental
in the Peruvian Amazon -
```

Extracted features

```
newsroom,amazon,capacity,amazon#0 (Pattern 1)
amazon,capacity,building,amazon#0 (Pattern 3)
amazon,watch,work,amazon#0(Pattern 3)
people,amazon,basin,amazon#0(Pattern 1)
peruvian,amazon,environmental,amazon#0(Pattern 1)
amazon,environmental,lawsuit,amazon#0 (Pattern 3)
amazon,poison,goldman,amazon#0 (Pattern 3)
```

Document snippet

```
Amazon Watch.
Amazon Watch works with indigenous and environmental organizations
in the Amazon Basin to defend the environment and
advance indigenous peoples rights in .
```

Extracted features

```
organization,amazon,basin,amazon#0 (Pattern 1)
environmental,organization,amazon,amazon#0(Pattern 2)
amazon,watch,work,amazon#0 (Pattern 3)
amazon,basin,defend,amazon#0 (Pattern 3)
```

Web Page Classification into Word Senses

In this Section, we describe the final processes shown in Figure 4.19, namely learning and prediction.

Word Sense Disambiguation

The first step is applying a WSD system, in order to provide separate sense predictions for all sentences in the documents to be classified. We have used the TiMBL memory based learning algorithm [Daelemans et al., 2001], a state-of-the-art supervised WSD system which has been widely employed for the task (see for instance [Hoste et al., 2002], [Mihalcea, 2002b]). TiMBL is fed with the examples extracted from Wikipedia pages for the learning process, and then applied on each occurrence of the noun in the document to be classified.

Prediction Criteria

To accomplish the goal of providing all documents with a unique sense prediction, we established some rules to deal with the annotations provided by the WSD system: as TiMBL predicts a sense for each sentence in the document, we established the following classification criteria:

1. If possible, we select the sense assigned by TiMBL to a maximum number of test sentences in the representation of the document.

2. In case of two or more senses with the same number of assigned noun occurrences, the one appearing first in the list of senses (from the Wikipedia disambiguation page) is chosen. This choice is similar to a random decision, because senses in disambiguation pages are not ranked according to any particular rule.

We label this classification strategy as **TiMBL-core** (when using only training examples from the page for the sense being trained), **TiMBL-inlinks** (when using the examples found in Wikipedia pages pointing to the sense being trained) or **TiMBL-all** (uses both) depending on the training set.

As we do not require a confidence threshold to make the assignment, we reach a maximum coverage at the expense of precision. We will show the effects of establishing a threshold in Section 4.6.5.

Adding the Sense Frequency Estimator

In this case, we modified the prediction criteria above as follows: (i) when there is a tie between two or more senses (which is much more likely than in the VSM approach), we pick up the sense with the highest frequency according to our estimator (see Section 4.5.5) and (ii) when no sense reaches 30% of the cases in the page to be disambiguated, we also resort to the most frequent sense heuristic (among the candidates for the page). We applied these criteria in the experiment

TiMBL-core+freq (we discarded ”inlinks” and ”all” versions because they were clearly worse than ”core”).

Figure 4.26 shows the modified algorithm; it is the same as the one presented in Figure 4.19, apart from the inclusion of combined-frequencies values in the decision stage.

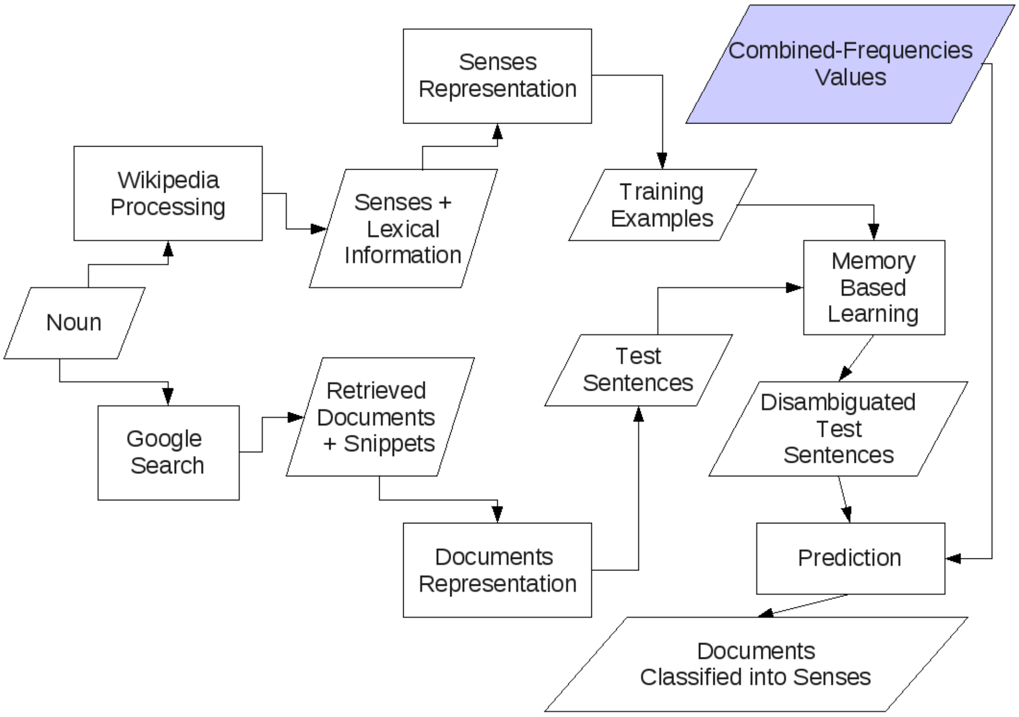

Senses Representation
Combined-Frequencies Values
Wikipedia Processing
Senses + Lexical Information
Training Examples
Memory Based Learning
Noun
Test Sentences
Retrieved Documents + Snippets
Disambiguated Test Sentences
Google Search
Documents Representation
Prediction
Documents Classified into Senses

Figure 4.26: WSD-based algorithm using frequencies information

4.6.4 Classification Results

Table 4.11 shows the classification results for both VSM and WSD approaches. The accuracy of systems is reported as precision, i.e. the number of pages correctly classified divided by the total number of predictions. This is approximately the same as recall (correctly classified pages divided by total number of pages) for our systems, because the algorithms provide an answer for every page containing text. Indeed, the actual coverage was 94%, because of two main reasons (i) there were some senses without any textual information, and that prevented them from generating training material (some senses included in the sense inventory did have no associated Wikipedia entry at all) and (ii), some documents were manually assessed on the basis of visual information (text included as images, photographs, logotypes) but did not provide any text that could be used by the classification algorithms.

Table 4.11: Classification Results

Experiment	Precision
random	.19
most frequent sense (estimation)	.46
TiMBL-core	.60
TiMBL-inlinks	.50
TiMBL-all	.58
TiMBL-core+freq	**.67**
VSM	.67
VSM-GT	.68
VSM-mixed	.67
VSM-GT+freq	**.69**

TiMBL results

All systems are significantly better than the random and most frequent sense baselines (at $p < 0.05$ using a standard t-test). Overall, both approaches (using TiMBL WSD machinery and using VSM) lead to similar results ($.67$ vs. $.69$), which would make VSM preferable because it is a simpler and more efficient approach. Taking a closer look at the results with TiMBL, there are a couple of interesting facts:

- As we have previously discussed, there is a substantial difference between using only examples taken from the Wikipedia Web page for the sense being trained (TiMBL-core, .60) and using examples from the Wikipedia pages pointing to that page (TiMBL-inlinks, .50). Examples taken from related pages (even if the relationship is close as in this case) seem to be too noisy for the task. This result is compatible with our findings in [Santamaría et al., 2003] (see previous chapter) using the Open Directory Project to extract examples automatically.

- Our estimation of sense frequencies turns out to be very helpful for cases where our TiMBL-based algorithm cannot provide an answer: precision rises from $.60$ (TiMBL-core) to $.67$ (TiMBL-core+freq). The difference is statistically significant at $p < 0.05$.

VSM results

As for the experiments with VSM, the variations tested do not provide substantial improvements to the baseline (which is $.67$). Using idf frequencies obtained from the Google Terabyte corpus (instead of frequencies obtained from the set of retrieved documents) provides only a small improvement (VSM-GT, $.68$), and adding the estimation of sense frequencies gives another small improvement ($.69$). Comparing the baseline VSM with the optimal setting (VSM-GT+freq), the difference is small ($.67$ vs $.69$) but relatively robust ($p = 0.066$ according to the t-test).

Remarkably, the use of frequency estimations is very helpful for the WSD approach but not for the SVM one, and they both end up with similar performance figures; this might indicate that using frequency estimations is only helpful up to certain precision ceiling.

Detailed results by noun

The detailed results per noun for experiments based on TiMBL can be seen in the following tables of Appendix A:

- **TiMBL-core** results in Table A.11
- **TiMBL-inlinks** results in Table A.12
- **TiMBL-all** results in Table A.13.
- **TiMBL-core+freq** results in Table A.14.

Figure 4.27 compares the results per noun of all four experiments. Figure 4.28 compares the best run (TiMBL-core+freq)with the random and most-frequent-sense baselines.

4.6.5 Precision/Coverage Trade-off

All the above experiments are done at maximal coverage, i.e., all systems assign a sense for every document in the test collection (at least for every document with textual content). But it is possible to enhance search results diversity without annotating every document (in fact, not every document can be assigned to a Wikipedia sense, as we have previously discussed). Thus, it would be useful to investigate which is the precision/coverage trade-off in our dataset. We experimented with the best performing system (VSM-GT+freq), introducing a similarity threshold: assignment of a document to a sense was only done if the similarity of the document to the Wikipedia page for the sense exceeded the similarity threshold.

We computed precision and coverage for every threshold in the range $[0.00 - 0.90]$ (beyond 0.90 coverage was null) and represented the results in Figure 4.29. The graph shows that we can classify around 20% of the documents with a precision above $.90$, and around 60% of the documents with a precision of $.80$.

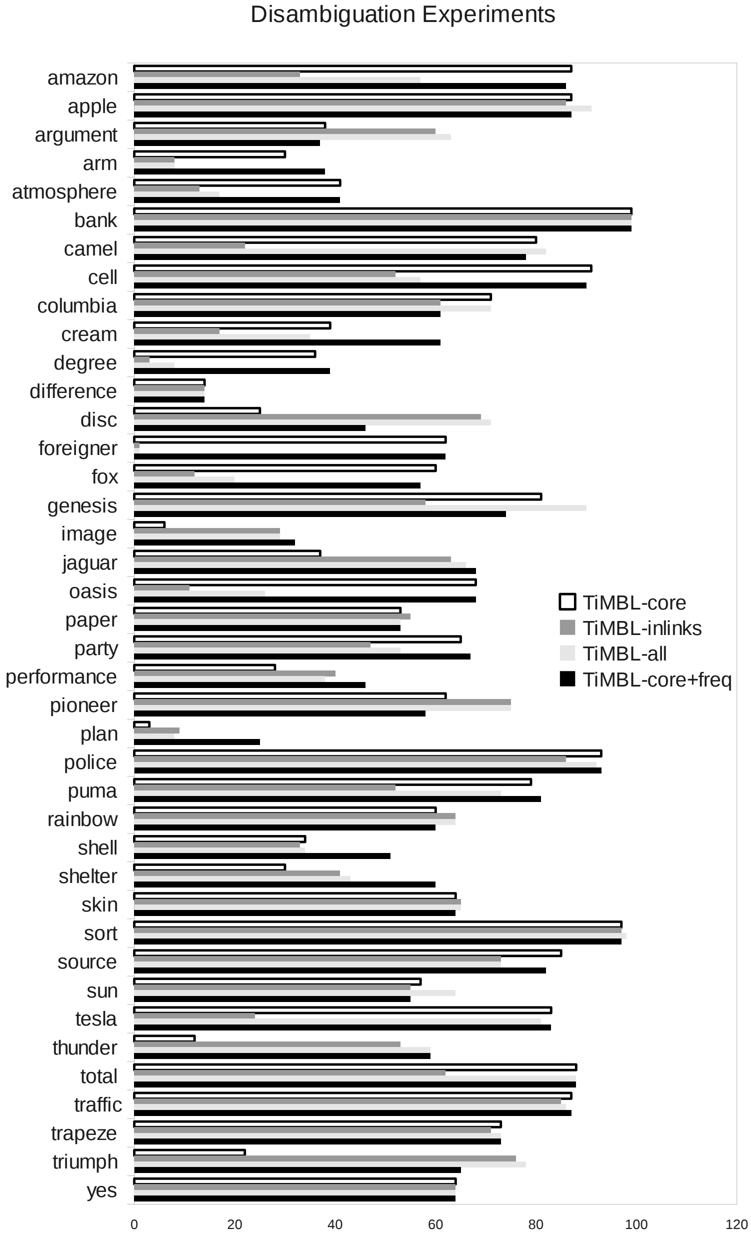


Figure 4.27: Results of classification experiments using variations of TiMBL

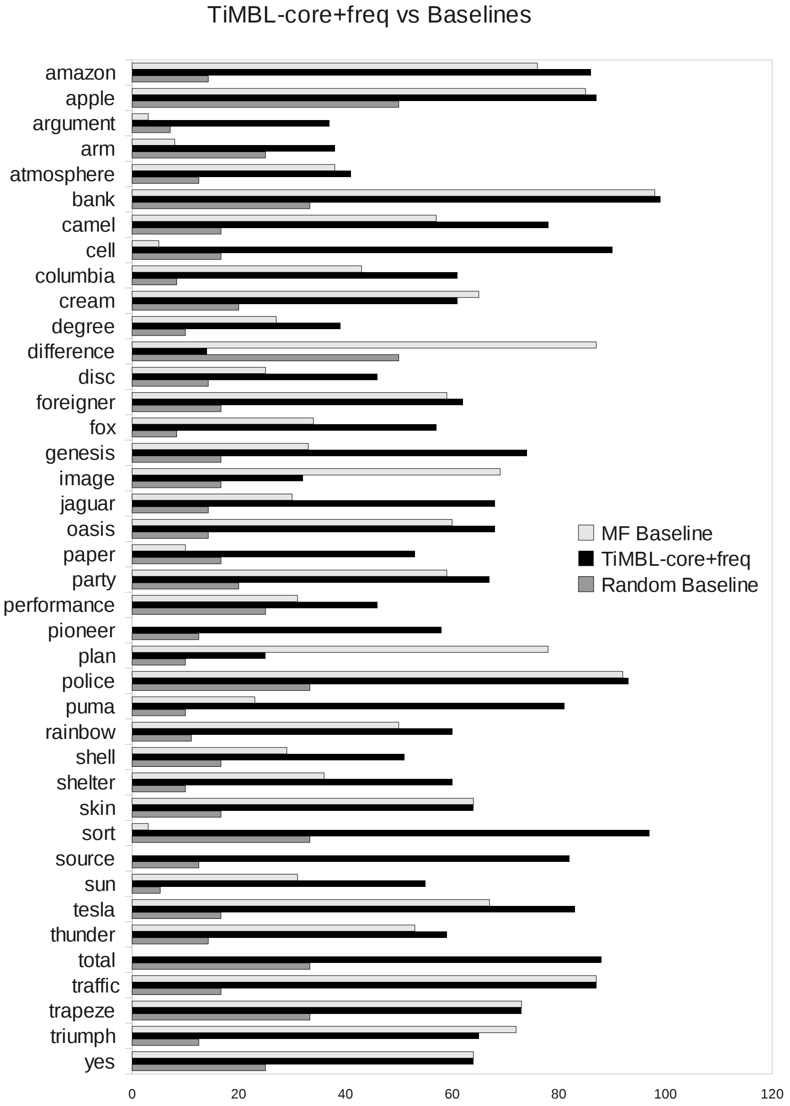


Figure 4.28: Results for the TiMBL-core+freq run compared to the baselines

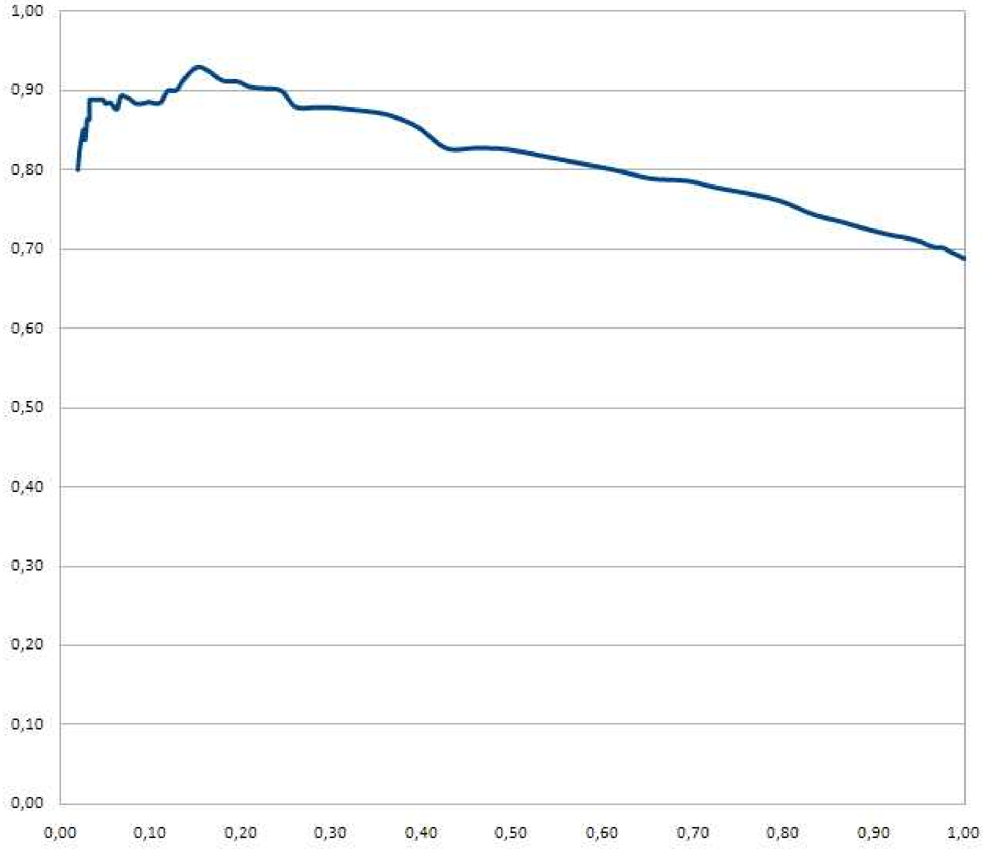


Figure 4.29: Precision/Coverage curve for VSM-GT+freq classification algorithm

4.6.6 Using Classification to Promote Diversity

Our final goal was to estimate how the reported classification accuracy might perform in practice to organize search results. In particular, we wanted to enhance diversity in search results. In order to provide an initial answer to this question, we have re-ranked the documents for the 40 nouns in our test bed, using our best classifier (VSM-GT+freq) and making a list of the top-ten documents with the primary criterion of maximizing the number of senses represented in the set, and the secondary criterion of maximizing the similarity scores of the documents to their assigned senses.

Results are presented in Table 4.12. Diversity in the top ten documents increases from an average of 3.00 Wikipedia senses represented in the original search engine rank, to 5.18 senses in the modified rank (the ceiling would be 6.5 in our test bed), with the coverage of senses going from 21% to 84%. Therefore, using a simple VSM algorithm, the coverage of (Wikipedia) senses in the top ten results becomes four times larger than in the original ranking.

Of course this does not imply that the modified rank is better than the original one: there are many other factors that influence the final ranking provided by a search engine. What our results indicate is that, with simple and efficient algorithms, Wikipedia can be used as a reference to improve search results diversity for one-word queries.

Table 4.12: Enhancement of Search Results Diversity

rank@10	**# senses**	**sense coverage**
original rank	3.00	21%
modified rank	5.18	84%

4.7 Related Work

Besides the work reviewed in Chapter 2, there are a couple of research issues related to our work with Wikipedia that we discuss here: (i) applications of Wikipedia (and relevance of Word Senses) for the problem of Web Search Diversity, and (ii) establishing relative frequencies for word senses in a corpus.

4.7.1 Diversity

As previously discussed, an alternative way of dealing with ambiguity in IR tasks consists of promoting diversity in the search results. This approach has not yet been exploited, but we believe that it is very relevant for disambiguating Web search results, although disambiguation is not the main goal of diversity. In fact, diversity is used both to represent sub-themes in a broad topic, or to consider alternative interpretations for ambiguous queries ([Agrawal et al., 2009]), which is one of our interests in this research.

To our knowledge, Wikipedia has not explicitly been used before to promote diversity in search results. [Li et al., 2007] however, use an approach similar to ours. Their goal, however, is to reduce diversity by giving more importance to certain retrieved results via query expansion (see Section 2.2.4).

[Clough et al., 2009] is the only article known to us that explicitly links diversity in search results with Wikipedia as a sense inventory. They analyze query diversity from a user's perspective, using a search log (from Microsoft Live) where they compute click entropy and query reformulation as indicators of queries which would require a diversity treatment. Click entropy measures how many search results have been clicked on by users - something which may indicate that they are looking for different kinds of things with the same query- , whereas query reformulations may indicate that search results do not satisfy the requirements of the user, which again can be directly related to diversity ([Radlinski and Dumais, 2006]). According to the click entropy measures performed, the queries with high diversity represented 18% of queries.

To analyze ambiguity, a list of all ambiguous terms (words and phrases) in WordNet and Wikipedia (using the disambiguation pages) is generated in [Clough et al., 2009], together with a list of the number of senses for each term. They did not find any significant correlation between the number of senses of a word in Wikipedia and the indicators used to discover diverse queries, although there is some positive correlation between the length of a Wikipedia article and the click entropy. This result does not discard, however, the usefulness of Wikipedia for queries that can benefit from an explicit treatment of diversity. And, in addition, it remains to be tested how well their diversity predictors are correlated with real user needs.

A relevant problem in this part of our research was that standard IR test collections do not usually consider ambiguous queries, and are thus inappropriate to test systems that promote diversity. In [Sanderson, 2008], the impact of ambiguity on the performance of search engines is studied, considering, in particular, terms which are not usually reflected in conventional dictionaries, by analyzing the disambiguation pages of Wikipedia. A comparison with WordNet suggests that it would be necessary to consider ambiguity more seriously for IR tasks. By examining the logs of several search engines, it is established the importance of ambiguity in actual queries, finding that these queries are relatively common, even more in the most frequent queries submitted to search engines. To explore the performance of IR systems with ambiguous queries, and given the lack of appropriate test collections, Sanderson decides to create such a collection, by adapting the technique of pseudo-queries, testing then a IR system. The results show that this system is not able to handle ambiguous queries with effectiveness. The final conclusion points to the need of creating new test collections for a better treatment of this kind of queries in IR research.

Indeed, it is only recently that appropriate test collections have being built, such as (i) [Paramita et al., 2009], a collection generated for the Image CLEF Photo Retrieval Task 2009 (a part of the CLEF evaluation campaign focused on image search and diversity), consisting of approximately half a million images with English annotations, and (ii) [Artiles et al., 2009] in which data sets are created for the WePS (Web People Search) Evaluation campaigns, focused on person name search.

The test bed that we have generated (see Section 4.2), is complementary to those ones, and we expect that it can contribute to foster research on search results diversity.

4.7.2 Establishing Relative Frequencies for Senses

One of the aims of our work has been to associate sense frequency estimators to an inventory of senses extracted from Wikipedia (see Section 4.5). This knowledge is crucial, because sense distributions tend to be skewed, and knowing them could be useful in disambiguation tasks.

As we have seen in Section 2.3.2, in [Medelyan et al., 2008] the relative frequency of senses being used as links is measured. This frequency is an internal measure of the relevance of a sense, in the same line that one of our proposed frequencies, although it is used with a different aim, specifically for WSD purposes.

As related work on the topic of assigning sense frequencies, [McCarthy et al., 2004] and [Mohammad and Hirst, 2006] present methods for automatically determining dominant senses of ambiguous words. Determining the dominant sense for a word is considered a very valuable information that could be used to take decisions, in an unsupervised system, when there is no better evidence. For WSD systems, the *most common sense* heuristic, which consists of selecting the predominant sense for a word, is often used as a hard to beat baseline which outperforms many of such systems.

[McCarthy et al., 2004] propose and evaluate a method for obtaining the predominant sense of a word, by using the neighbors from automatically extracted thesauri and semantic similarity measures to weight the impact of these neighbors in the senses of the word. For the evaluation process, the authors use the data in SemCor as a gold-standard. The accuracy of finding (i) the predominant sense, when it actually exists for the given word in SemCor, and also (ii) the WSD accuracy that would be obtained on SemCor, when using the proposed first sense in all contexts are calculated. Using the jcn and Lesk WordNet similarity measures [Pedersen et al., 2004], the results are, respectively, 54% and 48% for the Lesk and 54% and 46% for the jcn. The random baselines are 32% and 24%, and the upper-bound is 67% (first sense in SemCor). Comparing the automatically acquired predominant senses with the manually annotated resources SemCor for the Senseval-2 English all-words task, the obtained precisions are 64% and 69% respectively. In this work, it is shown that the rank of the senses may likely vary in domain-dependent contexts.

[Mohammad and Hirst, 2006] present a more efficient method using bootstrapping techniques to create a co-occurrence matrix, which does not require similarity measures, based on the hypothesis that the sense is indicated in the surrounding words of the given word and that the prevalence of a particular sense is proportional to the relative strength of the association between it and co-occurring words in the given text. Although they are not comparable with the obtained in [McCarthy et al., 2004] because the used thesauri are different, in [Mohammad and Hirst, 2006], the reached results are also near the upper-bound.

Note that our estimation of word sense frequencies in Web Search results is comparable to those of McCarthy and Mohammad, in spite of the fact that we are using only evidence obtained from Wikipedia (inlinks and visits), without using the corpus itself.

4.8 Conclusions

Our comparison between Wikipedia and WordNet reveals that, for Web documents, the coverage of word senses in Wikipedia is substantially better than in WordNet. This suggest that the long pursued goal of improving Web Search with semantic annotations should perhaps be based on a direct use of Wikipedia (and possibly other collaboratively authored Web Contents) instead of conventional lexical databases, which might have better formal properties but cannot be compared in coverage and updating rate with the Wikipedia.

Our results, in fact, support the hypothesis that Wikipedia can be used to organize search results for short, ambiguous queries: Wikipedia covers a substantial proportion of the actual meanings of nouns (used as one word queries) in search results, and it is feasible to classify pages in search results according to the most appropriate meaning of the noun in the Wikipedia.

We have also provided a way of estimating the relative frequencies of Wikipedia senses for a word in search results, which is independent from corpus evidence: it only uses internal Wikipedia structure (incoming links to a Wikipedia entry) and statistics on the number of visits received by each page. We find this (high) correlation between prominence in Wikipedia and prominence in Web Search results particularly useful, and also an interesting proof that world-wide collaborative authoring has interesting scale effects.

We expect that the test bed created for this research will complement the - currently short - set of benchmarking test sets to explore search results diversity and query ambiguity. And we believe that our results endorse further investigation on the use of Wikipedia to organize search results.

The negative side of our results is that, although our classification task resembled a Word Sense Disambiguation problem under a ”one sense per discourse” hypothesis, we have not been able to prove the usefulness of WSD strategies for the task: a simple (and more efficient) VSM approach performs better than a state-of-the-art disambiguation system (TiMBL).

Some limitations of our research must be noted: (i) the nature of our test bed (with every search result manually annotated in terms of two sense inventories) makes it too small to extract any solid conclusion on Web searches (ii) our work does not involve any study of diversity from the point of view of Web users (i.e. when a Web query addresses many different use needs in practice); (iii) we have tested our classifiers with a simple re-ordering of search results to test how much diversity can be improved, but a search results ranking depends on many other factors, some of them more crucial than diversity; it remains to be tested how can we use document/Wikipedia associations to improve search results clustering (for instance, providing seeds for the clustering process) and to provide search suggestions.

Chapter 5

Conclusions

In this thesis, we have investigated the role of collaboratively authored Web contents as resources to enrich or replace lexical resources for Word Sense Disambiguation and Discovery. We have focused on the two most prominent (non specialized) resources of this type: ODP and Wikipedia.

In our research, ODP has been used to enrich WordNet with rich domain information (associating word sense with ODP directories) providing ways of acquiring WSD examples and discovering new senses automatically. Wikipedia, on the other hand, has been compared with WordNet for the task of organizing search results in ambiguous queries.

We have focused on collaboratively authored Web contents because (i) they offer information about virtually any topic, and this information is organized in an accessible way, (ii) they are continually edited, updated and enlarged (iii) they provide periodical dumps which are easy to download and to handle, facilitating the processing and reproducibility of data and results. ODP hierarchically organizes Web sites by domains, thus containing implicit information about such topics, whereas Wikipedia is a large coverage, updated encyclopedic repository of explicit knowledge.

5.1 Main Contributions

5.1.1 Results

Enrichment of Linguistic Resources

We have described an algorithm that combines lexical information from WordNet with Web directories from the ODP to associate word senses with such directories, with 86% accuracy over the Senseval 2 English lexical sample test bed, and with coverage ranging between 73% and 88% of the domain specific senses in the test bed. We have seen that these associations can be used as rich characterizations for word senses: as a source of information to cluster senses according to their topical relatedness, to enrich WordNet with sense specializations, and to acquire sense-tagged corpora.

In an extrinsic evaluation of the richness of Web directories as sense characterizations in a supervised Word Sense Disambiguation task, we have shown that for correct directory/word sense associations, the samples automatically acquired from the Web directories are nearly as valid for training as the original Senseval 2 training instances, although they have been obtained without manual intervention. In the supervised WSD experiment we have carried out, the results suggest that the characterization of word senses with Web directories provide cleaner data, without further sophisticated filtering, than a direct use of the whole Web. In fact, the WSD results using training material from ODP directories give better results than what could be expected from previous cross-validations of (manually built) training and test WSD materials.

An intrinsic weakness of the method is that, as the categories are strongly related to topics and domains, word senses that do not belong to any domain cannot be associated to web directories (and therefore training instances cannot be acquired automatically). A less serious problem is that the measured coverage of our approach is low, although this problem could evolve positively with the growth of the ODP and its contents.

This part of our research has been published in the *Computational Linguistics* Journal (vol. 29, 2003) and has received more than 35 references at the time of finishing this Ph.D. dissertation.

Replacement of Linguistic Resources

The comparison performed between WordNet and Wikipedia by producing two sense inventories based on Wikipedia and WordNet respectively, leads us to the conclusion that the senses provided by Wikipedia are more relevant and offer a better coverage in search results; Indeed, Wikipedia senses cover much more search results (56%) than WordNet (32%). In the Bands subset (which should be more representative of plausible web queries, because there is at least one named entity named after each of the nouns in this set), Wikipedia covers near 70% of the top ten results in average. These results support the viability of using Wikipedia for clustering purposes, although not as a unique reference.

An interesting point in this research has been to study the relations between the intrinsic relevance of senses according to Wikipedia and the extrinsic relevance of such senses according to search results. We have shown that the distribution of senses in search results can be estimated by using the internal graph structure of the Wikipedia and the relative number of visits received by each sense in Wikipedia, We have performed some experiments with the main goal of automatically classifying search results into the most appropriate senses, comparing two approaches: a basic Information Retrieval technique, in which the documents and the Wikipedia pages are represented using a Vector Space Model (VSM) and compared with a standard cosine measure, and a WSD-based system trained with different sets of examples automatically acquired from Wikipedia pages. All systems are significantly better than the random and most frequent sense baselines (using $p < 0.05$ for a standard t-test). Both approaches (using TiMBL WSD machinery and using VSM) lead to similar results ($.67$ vs. $.69$).

Analyzing the behavior of search results regarding diversity in terms of Wikipedia senses, our results suggest that diversity is not a major priority for the current Google ranking algorithm, and an indication that Wikipedia could be used to cluster search results. Although the limited scope of our research and the unpredictable variability among words do not allow to extract definitive conclusions, we have signs that Wikipedia could be used to cluster search results: potentially the first top ten results could cover at least 6.5 different senses in average, which would be a substantial growth with respect to the current diversity in Google results (as measured in our test bed).

As a final point, we have seen that generic lexical resources can be used to promote diversity in Web search results for one-word, ambiguous queries, as a matter of fact, our results indicate that associating Web pages to Wikipedia senses with simple and efficient algorithms, we can produce modified rankings that cover four times more Wikipedia senses than the original search engine rankings.

There are some limitations that should be considered: (i) the nature of our test bed (with every search result manually annotated in terms of two sense inventories) makes it too small to extract any solid conclusion on Web searches (ii) our work does not involve any study of diversity from the point of view of Web users (i.e. when a Web query addresses many different use needs in practice); research in [Clough et al., 2009] suggests that word ambiguity in Wikipedia might not be related with diversity of search needs; (iii) we have tested our classifiers with a simple re-ordering of search results to test how much diversity can be improved, but a search results ranking depends on many other factors, some of them more crucial than diversity.

5.1.2 Resources

During the development of this thesis, we have created resources for Language Engineering which could be useful for the research community, and have been made available in our research group homepage (nlp.uned.es):

Massive Association of Web Directories to Word Senses We have applied the association algorithm described in Chapter 3 to all non-compound nouns in WordNet without non-alphabetic characters. In the resource built, at least one directory is associated to 13,375 nouns (28% of the candidate set). Propagating these sense/directory associations to synset/directory relations, the number of characterized nouns and word senses almost doubles: 24,558 nouns and 27,383 senses, covering 34% of the candidate nouns plus 7,027 multi-word terms that were not in the candidate set.

Search Results Diversity Test bed We have created a test bed for experiments in search results diversity, consisting of (i) 40 highly ambiguous nouns, (ii) two alternative inventories of senses (derived from Wikipedia and WordNet respectively) together with useful lexical information for the senses, and (iii) a collection of 4000 documents, manually associated with the most appropriate senses in both inventories.

We expect that the test bed created for this research will complement the - currently short - set of benchmarking test sets to explore search results diversity and query ambiguity.

As an overall conclusion, we have shown that collaboratively authored Web contents are a very valuable source of lexical information, either for enriching linguistic resources, as we have done with ODP, or to replace them in specific applications as in our study of diversity using Wikipedia.

5.2 Further Work

Our plans for future work involve three main lines of research: better exploitation of our results for Word Sense Disambiguation tasks, improvement of the resources built, and exploration of other search results organization techniques.

All-Words WSD tasks. In our experiments with ODP, we have restricted ourselves to the English lexical sample task to be able to compare manually annotated training material (as provided in the Senseval 2 test bed) with our automatically acquired examples. A natural step is now to use our massive annotation of WordNet sense with web directories to acquire examples for all possible WordNet nouns, and then apply them in an all-words task (where training material is scarce for everyone). Potentially the results should outperform current unsupervised approaches to the problem.

Enrichment of the Multilingual Content Repository.Another direction for future work is extending the algorithm to include sense/directory associations to the Multilingual Content Repository, which contains much more information than WordNet and should provide anchors for improved precision and recall of our algorithm. Alternatively, another interesting option would be to abandon WordNet derivatives and focus on establishing a direct connection between ODP and Wikipedia. If successful, ODP would provide a massive amount of links between Web pages and Wikipedia entries, which could potentially enrich sense distinctions in Wikipedia and enhance its value as a lexical resource for the whole Web.

Search Results Organization. We are also interested in considering the effects of Wikipedia inventories in other Information Access tasks. For instance, we have directly applied Wikipedia to promote diversity in search results, but it remains to be explored how can we use document/Wikipedia associations to explicitly cluster search results (including cluster labeling) and to provide search suggestions. An interesting option for clustering would be to use web page / Wikipedia sense annotations with a high confidence value as seeds for a clustering algorithm, which could in turn be used to improve web page classification.

Since the moment we began our research, collaboratively authored Web contents have grown exponentially, both in size and quality. At the same time, the Web has become a much noisier resource, because a substantial proportion of its contents are placed with the only purpose of manipulating search engine results. In this context, the Wikipedia has already become a fundamental resource for Natural Language Engineering; we have proved that it can even replace lexical databases for certain applications, and we have also shown that other resources such as the ODP can also play a key role in the development of large-scale linguistic resources equipped with massive amounts of world knowledge and usable for NLP at a web scale.

Bibliography

[Adafre and de Rijke, 2006] Adafre, S. and de Rijke, M. (2006). Finding similar sentences across multiple languages in Wikipedia. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 62–69.

[Agirre et al., 2000] Agirre, E., Ansa, O., Hovy, E., and Martínez, D. (2000). Enriching very large ontologies using the www. In *Proceedings of the Ontology Learning Workshop, European Conference on Artificial Intelligence (ECAI), Berlin, Germany.*

[Agirre et al., 1994] Agirre, E., Arregi, X., Artola, X., de Ilarraza, A., Evrard, F., Sarasola, K., and Soroa, A. (1994). Intelligent dictionary help system. *Applications and implications of current language for special purposes research*, pages 174–183.

[Agirre and de Lacalle, 2004] Agirre, E. and de Lacalle, O. (2004). Publicly available topic signatures for all WordNet nominal senses. In *Proceedings of the Language Resources and Evaluation Conference (LREC), Lisbon, Portugal.*

[Agirre and Martínez, 2000] Agirre, E. and Martínez, D. (2000). Exploring automatic word sense disambiguation with decision lists and the web. In *Proceedings of the COLING Workshop on Semantic Annotation and Intelligent Annotation, Luxembourg, 11–19.*

[Agirre and Martinez, 2001] Agirre, E. and Martinez, D. (2001). Learning class-to-class selectional preferences. In *Proceedings of the workshop on Computational Natural Language Learning-Volume 7*, pages 1–8.

[Agirre and Martinez, 2002] Agirre, E. and Martinez, D. (2002). Integrating selectional preferences in WordNet. In *Proceedings of the first International WordNet Conference in Mysore, India*, pages 21–25.

[Agirre and Martinez, 2004] Agirre, E. and Martinez, D. (2004). Unsupervised WSD based on automatically retrieved examples: the importance of bias. In

Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Barcelona, Spain, pages 25–32.

[Agirre and Rigau, 1996] Agirre, E. and Rigau, G. (1996). Word sense disambiguation using conceptual density. In *Proceedings of 16th International Conference on Computational Linguistics (COLING)*, volume 96, pages 16–22.

[Agirre and Rigau, 1997] Agirre, E. and Rigau, G. (1997). A proposal for word sense disambiguation using conceptual distance. *Amsterdam studies in the Theory and History of Linguistic Science Series 4*, pages 161–172.

[Agrawal et al., 2009] Agrawal, R., Gollapudi, S., Halverson, A., and Leong, S. (2009). Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining, Barcelona, Spain*, pages 5–14.

[Ahn et al., 2004] 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 the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland*.

[Álvez et al., 2008] Álvez, J., Atserias, J., Carrera, J., Climent, S., Oliver, A., and Rigau, G. (2008). Consistent annotation of EuroWordNet with the top concept ontology. In *Proceedings of the 4th Global WordNet Conference, Szeged, Hungary*.

[Anick, 2003] Anick, P. (2003). Using terminological feedback for Web search refinement : a log-based study. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto, Canada*, pages 88–95.

[Artiles et al., 2009] Artiles, J., Gonzalo, J., and Sekine, S. (2009). WePS 2 evaluation campaign: overview of the Web people search clustering task. In *2nd Web People Search Evaluation Workshop (WePS), 18th International World Wide Web Conference, Madrid, Spain*.

[Atserias et al., 2004] Atserias, J., Villarejo, L., Rigau, G., Agirre, E., Carroll, J., Magnini, B., and Vossen, P. (2004). The Meaning Multilingual Central Repository. In *Proceedings of the Second International WordNet Conference, Brno, Czech Republic*, pages 80–210.

[Brants and Franz, 2006] Brants, T. and Franz, A. (2006). Web 1T 5-gram version 1. *Linguistic Data Consortium, Philadelphia*.

[Bunescu and Pasca, 2006] Bunescu, R. and Pasca, M. (2006). Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy*, volume 6, pages 9–16.

[Buscaldi and Rosso, 2006] Buscaldi, D. and Rosso, P. (2006). Mining knowledge from Wikipedia for the question answering task. In *Proceedings of the International Conference on Language Resources and Evaluation, Genoa, Italy*.

[Carmel et al., 2009] Carmel, D., Roitman, H., and Zwerdling, N. (2009). Enhancing cluster labeling using Wikipedia. In *Proceedings of the 32nd international ACM SIGIR conference on Research and Development in Information Retrieval, Boston, Massachusetts*, pages 139–146.

[Carpineto et al., 2009] Carpineto, C., Osinski, S., Romano, G., and Weiss, D. (2009). A Survey of Web clustering engines. *ACM Computing Surveys*, 41(3).

[Chen et al., 2009] Chen, Y., Yat Mei Lee, S., and Huang, C. (2009). PolyUHK: a robust information extraction system for Web personal names. In *2nd Web People Search Evaluation Workshop (WePS), 18th International World Wide Web Conference, Madrid, Spain*.

[Clarke et al., 2008] Clarke, C., Kolla, M., Cormack, G., Vechtomova, O., Ashkan, A., Büttcher, S., and MacKinnon, I. (2008). Novelty and diversity in information retrieval evaluation. In *The 31th Annual International ACM SIGIR Conference, Singapore*, pages 659–666.

[Clough et al., 2005] Clough, P., Joho, H., and Sanderson, M. (2005). Automatically organising images using concept hierarchies. In *Proceedings of the SIGIR Workshop on Multimedia Information Retrieval,Salvador, Brazil*.

[Clough et al., 2009] Clough, P., Sanderson, M., Abouammoh, M., Navarro, S., and Paramita, M. (2009). Multiple approaches to analysing query diversity. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval,Boston, Massachusetts*, pages 734–735.

[Csomai and Mihalcea, 2007] Csomai, A. and Mihalcea, R. (2007). Linking educational materials to encyclopedic knowledge. *Frontiers in Artificial Intelligence and Applications*, 158:557–559.

[Csomai, 2005] Csomai, A.and Mihalcea, R. (2005). Senselearner: word sense disambiguation for all words in unrestricted text. In *in Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics, Michigan*.

[Cuadros et al., 2005] Cuadros, M., Padro, L., Rigau, G., Unibertsitatea, E., and Irdiazabal, P. (2005). Comparing methods for automatic acquisition of topic signatures. In *Proceedings of the International Conference in Recent Advances in Natural Language Processing, Borovets, Bulgaria*.

[Cuadros and Rigau, 2008] Cuadros, M. and Rigau, G. (2008). Bases de conocimiento multilíngües para el procesamiento semántico a gran escala. In *Acceso y visibilidad de la información multilíngüe en la red: el rol de la semántica*, pages 27–54.

[Cucerzan, 2007] Cucerzan, S. (2007). Large-scale named entity disambiguation based on Wikipedia data. In *Conference on Empirical Methods in Natural Language Processing. Conference on Computational Natural Language Learning, Prague, Czech Republic*, pages 708–716.

[Daelemans et al., 2001] Daelemans, W., Zavrel, J., van der Sloot, K., and van den Bosch, A. (2001). TiMBL: Tilburg Memory Based Learner, version 4.0, Reference Guide. Technical report, University of Antwerp.

[Fernández-Amorós et al., 2001] Fernández-Amorós, D., Gonzalo, J., and Verdejo, F. (2001). The role of conceptual relations in word sense disambiguation. In *Proceedings of the 6th International Conference on Application of Natural Language to Information Systems*, pages 87–98.

[Finkelstein et al., 2002] Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2002). Placing search in context: the concept revisited. *ACM Transactions on Information Systems (TOIS)*, 20(1):116–131.

[Gabay et al., 2008] Gabay, D., Eliahu, Z., and Elhadad, M. (2008). Using Wikipedia links to construct word segmentation corpora. In *Proceedings of the AAAI Workshop on Wikipedia and artificial intelligence, Chicago*.

[Gabrilovich and Markovitch, 2005] Gabrilovich, E. and Markovitch, S. (2005). Feature generation for text categorization using world knowledge. In *International Joint Conference on Artificial Intelligence, Edinburgh, Scotland*, volume 19, pages 1048–1053.

[Gabrilovich and Markovitch, 2007] Gabrilovich, E. and Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of The 20th International Joint Conference on Artificial Intelligence (IJCAI), Hyderabad, India*, pages 6–12.

[Gonzalo et al., 1998] Gonzalo, J., Verdejo, F., Chugur, I., and Cigarran, J. (1998). Indexing with WordNet synsets can improve text retrieval. *Proceedings of the COLING/ACL Workshop on Usage of WordNet for Natural Language Processing, Montreal, Canada*.

[Hoste et al., 2002] Hoste, V., Daelemans, W., Hendrickx, I., and Van Den Bosch, A. (2002). Evaluating the results of a memory-based word-expert approach to unrestricted word sense disambiguation. In *Proceedings of the ACL workshop on Word sense disambiguation: recent successes and future directions, Philadelphia*, volume 8, pages 95–101.

[Kaisser, 2008] Kaisser, M. (2008). The QuALiM question answering demo: supplementing answers with paragraphs drawn from Wikipedia. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Demo Session, Columbus, Ohio*, pages 32–35.

[Kassner et al., 2008] Kassner, L., Nastase, V., and Strube, M. (2008). Acquiring a taxonomy from the German Wikipedia. In *Proceedings of the 6th International Conference on Language Resources and Evaluation, Marrakech, Morocco*, pages 2143–2146.

[Kilgarriff, 2001a] Kilgarriff, A. (2001a). *Senseval 2: second international workshop on evaluating word sense disambiguation systems, Toulouse, France.*

[Kilgarriff, 2001b] Kilgarriff, A. (2001b). English lexical sample task description. In *Proceedings of Senseval 2, Second International Workshop on Evaluating Word Sense Disambiguation Systems, Toulouse, France*, pages 17–20.

[Kilgarriff, 2001c] Kilgarriff, A. (2001c). Web as corpus. In *Proceedings of Corpus Linguistics 2001, Lancaster, UK*.

[Kilgarriff, 2002] Kilgarriff, A. (2002). English lexical sample task description. In *Proceedings of Senseval 2: Second International Workshop on Evaluating Word Sense Disambiguation Systems, Toulouse, France*.

[Kilgarriff and Palmer, 2000] Kilgarriff, A. and Palmer, M. (2000). Introduction to the special issue on Senseval. *Computers and the Humanities*, 34(1):1–13.

[Kilgarriff and Rosenzweig, 2000] Kilgarriff, A. and Rosenzweig, J. (2000). Framework and results for English Senseval. *Computers and the Humanities*, 34(1):15–48.

[Leacock et al., 1998] Leacock, C., Chodorow, M., and Miller, G. (1998). Using corpus statistics and wordnet relations for sense identification. *Computational Linguistics.*

[Lee and Ng, 2002] Lee, Y. and Ng, H. (2002). An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In *Proceedings of the ACL conference on Empirical methods in natural language processing, Pennsylvania, Philadelphia*, volume 10.

[Li et al., 2007] Li, Y., Luk, W., Ho, K., and Chung, F. (2007). Improving weak ad-hoc queries using Wikipedia as external corpus. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, Amsterdam, The Netherlands*.

[Magnini and Cavaglia, 2000] Magnini, B. and Cavaglia, G. (2000). Integrating subject field codes into WordNet. In *Proceedings of the Second International Conference on Language Resources and Evaluation, Athens, Greece*, pages 1413–1418.

[Magnini and Strapparava, 2000] Magnini, B. and Strapparava, C. (2000). Experiments in word domain disambiguation for parallel texts. In *Proceedings of the ACL-2000 workshop on Word senses and multi-linguality*, pages 27–33.

[McCarthy, 2001] McCarthy, D. (2001). *Lexical acquisition at the syntax-semantics interface: diathesis alternations, subcategorization frames and selectional preferences*.

[McCarthy et al., 2004] McCarthy, D., Koeling, R., Weeds, J., and Carroll, J. (2004). Finding predominant word senses in untagged text. In *Proceedings of the 42th annual meeting of the Association for Computational Linguistics, Barcelona, Spain*.

[Medelyan et al., 2008] Medelyan, O., Witten, I., and Milne, D. (2008). Topic indexing with Wikipedia. In *Proceedings of the AAAI WikiAI workshop,Chicago*.

[Mihalcea, 2002a] Mihalcea, R. (2002a). Bootstrapping large sense tagged corpora. In *Proceedings of the Language Resources and Evaluation Conference (LREC), Las Palmas, Spain.*

[Mihalcea, 2002b] Mihalcea, R. (2002b). Instance based learning with automatic feature selection applied to word sense disambiguation. In *Proceedings of the 19th international conference on Computational linguistics, Taipei, Taiwan*, volume 1.

[Mihalcea, 2003] Mihalcea, R. (2003). Word sense disambiguation with pattern learning and automatic feature selection. *Natural Language Engineering*, 8(04):343–358.

[Mihalcea, 2007] Mihalcea, R. (2007). Using Wikipedia for automatic word sense disambiguation. In *Proceedings of the Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics, Rochester, New York.*

[Mihalcea et al., 2004] Mihalcea, R., Chklovski, T., and Kilgarriff, A. (2004). The Senseval 3 English lexical sample task. In *Senseval 3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Barcelona, Spain*, pages 25–28.

[Mihalcea and Csomai, 2007] Mihalcea, R. and Csomai, A. (2007). Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the 16th ACM conference on Conference on information and knowledge management, Lisbon, Portugal*, pages 233–242.

[Mihalcea and Moldovan, 1999] Mihalcea, R. and Moldovan, D. (1999). An automatic method for generating sense tagged corpora. In *Proceedings of the American Association for Artificial Intelligence (AAAI), Orlando*, page 461–466.

[Mihalcea and Moldovan, 2000] Mihalcea, R. and Moldovan, D. (2000). An iterative approach to word sense disambiguation. In *In Proceedings of Flairs 2000, Orlando.*

[Mihalcea and Moldovan, 2001] Mihalcea, R. and Moldovan, D. (2001). eXtended WordNet: progress report. In *Proceedings of NAACL Workshop on WordNet and Other Lexical Resources, Pittsburgh, PA.*, pages 95–100.

[Miller and Charles, 1991] Miller, G. and Charles, W. (1991). Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28.

[Miller et al., 1993] Miller, G., Leacock, C., Tengi, R., and T., B. R. (1993). A semantic concordance. In *Proceedings of the ARPA WorkShop on Human Language Technology, San Francisco, Morgan Kaufman.*

[Miller et al., 1990] Miller, G., R. Beckwith, C., Fellbaum, D., Gross, and Miller, K. (1990). Wordnet: an on-line lexical database. *International Journal of Lexicograph*, 3(4).

[Milne, 2007] Milne, D. (2007). Computing semantic relatedness using Wikipedia link structure. In *Proceedings of the New Zealand Computer Science Research Student Conference, Hamilton, New Zealand.*, volume 7.

[Milne et al., 2006] Milne, D., Medelyan, O., and Witten, I. (2006). Mining domain-specific thesauri from Wikipedia: a case study. In *Proceedings of the*

IEEE/WIC/ACM International Conference on Web Intelligence, Hong Kong, pages 442–448.

[Mohammad and Hirst, 2006] Mohammad, S. and Hirst, G. (2006). Determining word sense dominance using a thesaurus. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy*, pages 121–128.

[Nastase, 2008] Nastase, V. (2008). Topic-driven multi-document summarization with encyclopedic knowledge and spreading activation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, Waikiki, Honolulu, Hawaii*, pages 763–772.

[Nastase and Strube, 2008] Nastase, V. and Strube, M. (2008). Decoding wikipedia categories for knowledge acquisition. In *Proceedings of the AAAI Conference on Artificial Intelligence, Trento, Italy*, volume 8, pages 1219–1224.

[Nelken and Yamangil, 2008] Nelken, R. and Yamangil, E. (2008). Mining Wikipedia's article revision history for training computational linguistics algorithms. In *Proceedings of the WIKI-AI: Wikipedia and AI Workshop at the AAAI Conference, Chicago*.

[Niles and Pease, 2001] Niles, I. and Pease, A. (2001). Towards a standard upper ontology. In *Proceedings of the international conference on Formal Ontology in Information System, Ogunquit, Maine*, pages 2–9.

[Paramita et al., 2009] Paramita, M., Sanderson, M., and Clough, P. (2009). Diversity in photo retrieval: overview of the ImageCLEFPhoto task 2009. *CLEF working notes*, 2009.

[Pedersen, 2001] Pedersen, T. (2001). Machine learning with lexical features: The duluth approach to senseval 2. In *Proceedings of Senseval 2, Toulouse, France*.

[Pedersen et al., 2004] Pedersen, T., Patwardhan, S., and Michelizzi, J. (2004). Wordnet: similarity-measuring the relatedness of concepts. In *Proceedings of the National Conference on Artificial Intelligence, San Jose, California*, pages 1024–1025.

[Ponzetto and Strube, 2006] Ponzetto, S. and Strube, M. (2006). Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In *Proceedings of Human Language Technology Conference/North American chapter of the Association for Computational Linguistics Annual Meeting, New York City, New York*, volume 6, pages 192–199.

[Ponzetto and Strube, 2007] Ponzetto, S. and Strube, M. (2007). Deriving a large scale taxonomy from Wikipedia. In *Proceedings of the national conference on artificial intelligence, Vancouver, British Columbia, Canada*, volume 22, page 1440.

[Rada et al., 1989] Rada, R., Mili, H., Bicknell, E., and Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE transactions on systems, man and cybernetics*, 19(1):17–30.

[Radlinski and Dumais, 2006] Radlinski, F. and Dumais, S. (2006). Improving personalized web search using result diversification. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, Seattle, Washington*.

[Resnik, 1995a] Resnik, P. (1995a). Disambiguating noun groupings with respect to WordNet senses. In *Proceedings of the Third Workshop on Very Large Corpora, Cambridge, Massachusetts*, pages 54–68.

[Resnik, 1995b] Resnik, P. (1995b). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, Montreal, Quebec, Canada*, volume 14, pages 448–453.

[Rigau et al., 2002] Rigau, G., Magnini, B., Agirre, E., Vossen, P., and Carroll, J. (2002). Meaning: A roadmap to knowledge technologies. In *In Proceedings of COLING Workshop on A Roadmap for Computational Linguistics, Taipei, Taiwan*.

[Rubenstein and Goodenough, 1965] Rubenstein, H. and Goodenough, J. (1965). Contextual correlates of synonymy. *Communications of the ACM*, 8(10).

[Ruiz-Casado et al., 2005] Ruiz-Casado, M., Alfonseca, E., and Castells, P. (2005). Automatic assignment of Wikipedia encyclopedic entries to WordNet synsets. *Proceedings of Advances in Web Intelligence, Lozd, Poland*, 3528:380–386.

[Sanderson, 1999] Sanderson, M. (1999). Deriving concept hierarchies from text. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, Berkeley, California*, pages 206–213.

[Sanderson, 2000] Sanderson, M. (2000). Retrieving with good sense. *Information Retrieval*, 2(1):49–69.

[Sanderson, 2008] Sanderson, M. (2008). Ambiguous queries: test collections need more sense. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, Singapore*, pages 499–506.

[Santamaría et al., 2003] Santamaría, C., Gonzalo, J., and Verdejo, F. (2003). Automatic association of Web directories to word senses. *Computational Linguistics*, 29(3):485–502.

[Schmitz et al., 2006] Schmitz, C., Hotho, A., Jäschke, R., and Stumme, G. (2006). Mining association rules in folksonomies. In *Proceedings of the IFCS 2006 Conference Data Science and Classification, Ljubljana, Slovenia*, pages 261–270.

[Schmitz, 2006] Schmitz, P. (2006). Inducing ontology from flickr tags. In *Collaborative Web Tagging Workshop at World Wide Web Conference, Edinburgh, Scotland*, pages 210–214.

[Stamou et al., 2002] Stamou, S., A., N., Hoppenbrouwers, J., Saiz-Noeda, M., and Christodoulakis, D. (2002). EuroTerm. Extending EWN using both the expand and merge model. In *1st International Wordnet Conference, Mysore, India*.

[Strube and Ponzetto, 2006] Strube, M. and Ponzetto, S. (2006). WikiRelate! Computing semantic relatedness using Wikipedia. In *Proceedings of The 21th National Conference on Artificial Intelligence, Boston, Massachusetts*, volume 21, page 1419–1424.

[Sussna, 1993] Sussna, M. (1993). Word sense disambiguation for free-text indexing using a massive semantic network. In *Proceedings of the second international conference on Information and knowledge management, Washington, DC, USA*, pages 67–74.

[Tufis et al., 2004] Tufis, D., Cristea, D., and Stamou, S. (2004). BalkaNet: aims, methods, results and perspectives. A general overview. *Science and Technology*, 7(1-2):9–43.

[Voorhees, 1993] Voorhees, E. (1993). Using wordnet to disambiguate word senses for text retrieval. In *Proceedings of the 16th Annual ACM SIGIR Conference on Research and Development in Information Retrieval, Pittsburgh, Pennsylvania.*, pages 171–180.

[Vossen, 1998] Vossen, P. (1998). Eurowordnet a multilingual database with lexical semantic networks. *Computational Linguistics*, 25(4).

[Witten and Milne, 2008] Witten, D. and Milne, D. (2008). An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence, Chicago, USA*, pages 25–30.

[Xu et al., 2009] Xu, Y., Jones, G., and Wang, B. (2009). Query dependent pseudo-relevance feedback based on Wikipedia. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, Boston, Massachussetts*, pages 59–66.

[Yan et al., 2009] Yan, Y., Okazaki, N., Matsuo, Y., Yang, Z., and Ishizuka, M. (2009). Unsupervised relation extraction by mining Wikipedia texts using information from the Web. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Suntec, Singapore*, pages 1021–1029.

[Yarowsky, 1995] Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of 33th Annual Meeting of the Association for Computational Linguistics, Cambridge, Massachusetts, USA*.

[Yasuda and Sumita, 2008] Yasuda, K. and Sumita, E. (2008). Method for building sentence-aligned corpus from Wikipedia. *Wikipedia and Artificial Intelligence: An Evolving Synergy*, pages 64–66.

[Ye et al., 2009] Ye, Z., Huang, X., and Lin, H. (2009). A graph-based approach to mining multilingual word associations from Wikipedia. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, Boston, Massachusetts, USA*, pages 690–691.

[Zhang et al., 2006] Zhang, L., Wu, X., and Yu, Y. (2006). Emergent semantics from folksonomies: A quantitative study. *Journal on Data Semantics VI*, pages 168–186.

Appendix A

Complementary Tables and Figures

	Wikipedia		WordNet	
	#sens/#with docs	**#clustered docs**	**#sens/#with docs**	**#clustered docs**
amazon	19/3	10	4/1	1
apple	11/2	7	2/2	1
camel	10/4	8	1/1	3
cell	17/3	7	7/1	4
columbia	57/4	6	5/2	4
cream	6/4	6	3/1	1
foreigner	9/3	6	2/0	0
fox	139/4	7	7/0	0
genesis	33/3	6	2/1	2
jaguar	21/2	6	1/1	4
oasis	26/2	3	2/0	0
pioneer	35/2	5	2/0	0
police	15/2	9	1/1	7
puma	18/1	8	1/0	0
rainbow	43/4	5	2/1	2
shell	19/2	7	10/0	0
skin	25/3	7	6/1	5
sun	54/2	6	5/1	3
tesla	8/4	7	2/1	4
thunder	16/4	5	3/2	1
total	6/3	6	2/0	0
traffic	13/3	8	4/1	6
trapeze	5/3	4	1/1	2
triumph	23/3	5	2/0	0
yes	12/2	3	1/0	0
Total	**640/72**	**157**	**78/19**	**50**

Table A.1: Search results for the first 10 documents and for the Band nouns

	Wikipedia		WordNet	
	#sens/#with docs	**#clustered docs**	**#sens/#with docs**	**#clustered docs**
amazon	19/5	24	4/2	2
apple	11/2	18	2/2	1
camel	10/4	11	1/1	5
cell	17/4	19	7/1	11
columbia	57/6	12	5/4	10
cream	6/4	12	3/1	1
foreigner	9/4	17	2/2	2
fox	139/6	14	7/1	1
genesis	33/4	10	2/1	2
jaguar	21/3	16	1/1	9
oasis	26/4	8	2/0	0
pioneer	35/5	13	2/1	2
police	15/2	23	1/1	20
puma	18/2	17	1/1	1
rainbow	43/6	9	2/1	3
shell	19/5	18	10/1	3
skin	25/6	19	6/2	9
sun	54/10	16	5/1	4
tesla	8/6	20	2/1	15
thunder	16/7	12	3/2	2
total	6/3	7	2/0	0
traffic	13/3	21	4/1	19
trapeze	5/3	13	1/1	9
triumph	23/4	12	2/0	0
yes	12/3	7	1/0	0
Total	**640/111**	**368**	**78/29**	**131**

Table A.2: Search results for the first 25 documents and for the Band nouns

	Wikipedia		WordNet	
	#sens/#with docs	**#clustered docs**	**#sens/#with docs**	**#clustered docs**
amazon	19/5	41	4/2	2
apple	11/2	39	2/2	6
camel	10/6	26	1/1	17
cell	17/5	39	7/2	30
columbia	57/7	17	5/4	15
cream	6/5	18	3/3	5
foreigner	9/5	33	2/2	8
fox	139/8	26	7/1	1
genesis	33/6	17	2/1	4
jaguar	21/6	33	1/1	13
oasis	26/6	12	2/0	0
pioneer	35/5	16	2/1	2
police	15/2	42	1/1	37
puma	18/5	27	1/1	6
rainbow	43/9	17	2/1	6
shell	19/5	37	10/2	7
skin	25/6	38	6/2	25
sun	54/15	27	5/1	7
tesla	8/6	37	2/1	29
thunder	16/7	17	3/2	3
total	6/3	12	2/0	0
traffic	13/3	38	4/1	34
trapeze	5/3	29	1/1	22
triumph	23/5	19	2/0	0
yes	12/4	12	1/1	2
Total	**640/139**	**669**	**78/34**	**281**

Table A.3: Search results for the first 50 documents and for the Band nouns

	Wikipedia		**WordNet**	
	#sens/#with docs	**#clustered docs**	**#sens/#with docs**	**#clustered docs**
amazon	19/7	60	4/2	3
apple	11/2	56	2/2	8
camel	10/6	38	1/1	23
cell	17/6	56	7/3	47
columbia	57/12	24	5/4	21
cream	6/5	20	3/3	6
foreigner	9/5	50	2/2	15
fox	139/10	38	7/1	1
genesis	33/6	20	2/1	6
jaguar	21/6	49	1/1	17
oasis	26/7	16	2/0	0
pioneer	35/6	19	2/1	2
police	15/2	66	1/1	60
puma	18/10	42	1/1	8
rainbow	43/9	23	2/1	11
shell	19/6	55	10/2	7
skin	25/6	56	6/3	36
sun	54/17	33	5/1	9
tesla	8/6	58	2/1	42
thunder	16/7	19	3/3	4
total	6/3	13	2/0	0
traffic	13/6	62	4/1	53
trapeze	5/3	45	1/1	34
triumph	23/7	33	2/0	0
yes	12/4	13	1/1	2
Total	**640/164**	**964**	**78/37**	**415**

Table A.4: Search results for the first 75 documents and for the Band nouns

	Wikipedia		WordNet	
	#sens/#with docs	#clustered docs	#sens/#with docs	#clustered docs
amazon	19/7	91	4/2	5
apple	11/2	73	2/2	12
camel	10/6	51	1/1	29
cell	17/6	82	7/3	70
columbia	57/12	30	5/5	28
cream	6/5	26	3/3	8
foreigner	9/6	76	2/2	28
fox	139/12	61	7/1	3
genesis	33/6	33	2/1	11
jaguar	21/7	74	1/1	20
oasis	26/7	21	2/1	1
pioneer	35/8	25	2/1	7
police	15/3	99	1/1	91
puma	18/10	63	1/1	14
rainbow	43/9	26	2/1	13
shell	19/6	72	10/2	10
skin	25/6	77	6/3	49
sun	54/19	46	5/1	15
tesla	8/6	86	2/1	62
thunder	16/7	21	3/3	6
total	6/3	16	2/0	0
traffic	13/6	91	4/1	78
trapeze	5/3	51	1/1	37
triumph	23/8	53	2/0	0
yes	12/4	14	1/1	2
Total	**640/174**	**1358 (.54)**	**78/39**	**599 (0.24)**

Table A.5: Search results for the first 100 documents and for the Band nouns

	Wikipedia		WordNet	
	#sens/#with docs	#clustered docs	#sens/#with docs	#clustered docs
argument	23/4	6	7/5	5
arm	8/0	0	6/0	0
atmosphere	13/4	8	6/1	1
bank	19/2	8	10/1	7
degree	22/6	6	7/2	4
difference	5/1	2	5/1	2
disc	12/1	1	4/2	1
image	29/4	6	9/1	3
paper	14/1	7	7/1	7
party	14/4	9	5/2	8
performance	8/2	6	5/1	2
plan	19/7	10	3/2	4
shelter	19/5	8	5/1	5
sort	5/2	4	4/3	4
source	32/5	8	9/1	1
Total	**242/48**	**89**	**92/24**	**54**

Table A.6: Search results for the first 10 documents and for Senseval nouns

	Wikipedia		WordNet	
	#sens/#with docs	**#clustered docs**	**#sens/#with docs**	**#clustered docs**
argument	23/7	14	7/5	8
arm	8/1	2	6/1	2
atmosphere	13/7	16	6/4	6
bank	19/2	21	10/1	20
degree	22/6	15	7/2	8
difference	5/2	6	5/3	6
disc	12/5	9	4/3	5
image	29/6	14	9/1	7
paper	14/4	20	7/2	18
party	14/5	22	5/2	20
performance	8/4	17	5/3	10
plan	19/8	21	3/3	14
shelter	19/7	17	5/4	11
sort	5/2	16	4/3	15
source	32/6	12	9/2	6
Total	**242/72**	**222**	**92/39**	**156**

Table A.7: Search results for the first 25 documents and for Senseval nouns

	Wikipedia		WordNet	
	#sens/#with docs	**#clustered docs**	**#sens/#with docs**	**#clustered docs**
argument	23/11	33	7/5	16
arm	8/1	4	6/1	4
atmosphere	13/7	26	6/4	13
bank	19/2	44	10/1	39
degree	22/7	36	7/2	22
difference	5/2	16	5/3	20
disc	12/6	23	4/3	13
image	29/6	21	9/1	14
paper	14/5	38	7/3	33
party	14/5	42	5/2	37
performance	8/4	25	5/4	16
plan	19/8	42	3/3	35
shelter	19/10	32	5/5	21
sort	5/2	34	4/3	35
source	32/6	16	9/3	14
Total	**242/82**	**432**	**92/43**	**332**

Table A.8: Search results for the first 50 documents and for Senseval nouns

	Wikipedia		WordNet	
	#sens/#with docs	**#clustered docs**	**#sens/#with docs**	**#clustered docs**
argument	23/14	49	7/6	19
arm	8/4	9	6/2	7
atmosphere	13/7	37	6/5	21
bank	19/3	68	10/2	58
degree	22/9	54	7/3	32
difference	5/2	22	5/3	28
disc	12/7	38	4/3	23
image	29/6	35	9/1	28
paper	14/5	57	7/3	48
party	14/5	56	5/2	51
performance	8/4	39	5/5	33
plan	19/9	59	3/3	51
shelter	19/10	51	5/5	32
sort	5/3	50	4/3	55
source	32/7	22	9/3	18
Totals	**242/95**	**646**	**92/49**	**504**

Table A.9: Search results for the first 75 documents and for Senseval nouns

	Wikipedia		WordNet	
	#sens/#with docs	**#clustered docs**	**#sens/#with docs**	**#clustered docs**
argument	23/14	64	7/6	25
arm	8/4	15	6/3	14
atmosphere	13/8	56	6/5	33
bank	19/3	86	10/2	77
degree	22/10	77	7/4	49
difference	5/2	31	5/4	40
disc	12/7	53	4/3	34
image	29/6	50	9/1	40
paper	14/6	71	7/3	60
party	14/5	75	5/2	70
performance	8/4	51	5/5	47
plan	19/10	80	3/3	67
shelter	19/10	70	5/5	38
sort	5/3	63	4/3	77
source	32/8	35	9/3	25
Total	**242/100**	**877 (0.59)**	**92/52**	**696 (0.46)**

Table A.10: Search results for the first 100 documents and for Senseval nouns

	#coincidences	**#predictions**	**precision**
amazon	79	91	0.87
apple	61	70	0.87
argument	23	60	0.38
arm	4	13	0.30
atmosphere	22	54	0.41
bank	83	84	0.99
camel	41	51	0.80
cell	74	81	0.91
columbia	20	28	0.71
cream	9	23	0.39
degree	27	75	0.36
difference	4	29	0.14
disc	13	52	0.25
foreigner	44	71	0.62
fox	36	60	0.60
genesis	25	31	0.81
image	2	31	0.06
jaguar	26	71	0.37
oasis	13	19	0.68
paper	35	66	0.53
party	47	72	0.65
performance	14	50	0.28
pioneer	15	24	0.62
plan	2	77	0.03
police	91	98	0.93
puma	41	52	0.79
rainbow	15	25	0.60
shell	24	70	0.34
shelter	19	63	0.30
skin	49	77	0.64
sort	57	59	0.97
source	28	33	0.85
sun	24	42	0.57
tesla	71	86	0.83
thunder	2	17	0.12
total	14	16	0.88
traffic	76	87	0.87
trapeze	36	49	0.73
triumph	8	37	0.22
yes	9	14	0.64
Total	**1280**	**2108**	**.60**

Table A.11: WSD experiments: TiMBL-core results

	#coincidences	#predictions	precision
amazon	30	91	0.33
apple	60	70	0.86
argument	36	60	0.60
arm	1	13	0.08
atmosphere	7	54	0.13
bank	83	84	0.99
camel	11	51	0.22
cell	42	81	0.52
columbia	17	28	0.61
cream	4	23	0.17
degree	2	75	0.03
difference	4	29	0.14
disc	36	52	0.69
foreigner	1	71	0.01
fox	7	60	0.12
genesis	18	31	0.58
image	9	31	0.29
jaguar	45	71	0.63
oasis	2	19	0.11
paper	36	66	0.55
party	34	72	0.47
performance	20	50	0.40
pioneer	18	24	0.75
plan	7	77	0.09
police	84	98	0.86
puma	27	52	0.52
rainbow	16	25	0.64
shell	23	70	0.33
shelter	26	63	0.41
skin	50	77	0.65
sort	57	59	0.97
source	24	33	0.73
sun	23	42	0.55
tesla	21	86	0.24
thunder	9	17	0.53
total	10	16	0.62
traffic	74	87	0.85
trapeze	35	49	0.71
triumph	28	37	0.76
yes	9	14	0.64
Total	**1046**	**2108**	**.50**

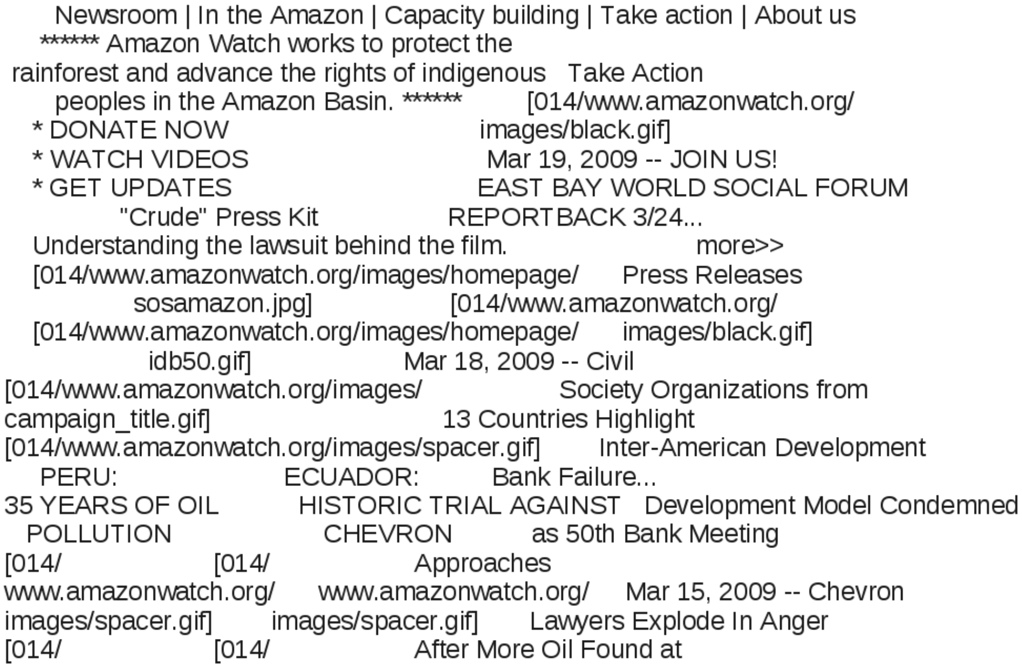
Table A.12: WSD experiments: TiMBL-inlinks results

	#coincidences	**#predictions**	**precision**
amazon	52	91	0.57
apple	64	70	0.91
argument	38	60	0.63
arm	1	13	0.08
atmosphere	9	54	0.17
bank	83	84	0.99
camel	42	51	0.82
cell	46	81	0.57
columbia	20	28	0.71
cream	8	23	0.35
degree	6	75	0.08
difference	4	29	0.14
disc	37	52	0.71
foreigner	1	71	0.01
fox	12	60	0.20
genesis	28	31	0.90
image	9	31	0.29
jaguar	47	71	0.66
oasis	5	19	0.26
paper	35	66	0.53
party	38	72	0.53
performance	19	50	0.38
pioneer	18	24	0.75
plan	6	77	0.08
police	90	98	0.92
puma	38	52	0.73
rainbow	16	25	0.64
shell	24	70	0.34
shelter	27	63	0.43
skin	50	77	0.65
sort	58	59	0.98
source	24	33	0.73
sun	27	42	0.64
tesla	70	86	0.81
thunder	10	17	0.59
total	14	16	0.88
traffic	75	87	0.86
trapeze	36	49	0.73
triumph	29	37	0.78
yes	9	14	0.64
Total	**1225**	**2108**	**.58**

Table A.13: WSD experiments: TiMBL-all results

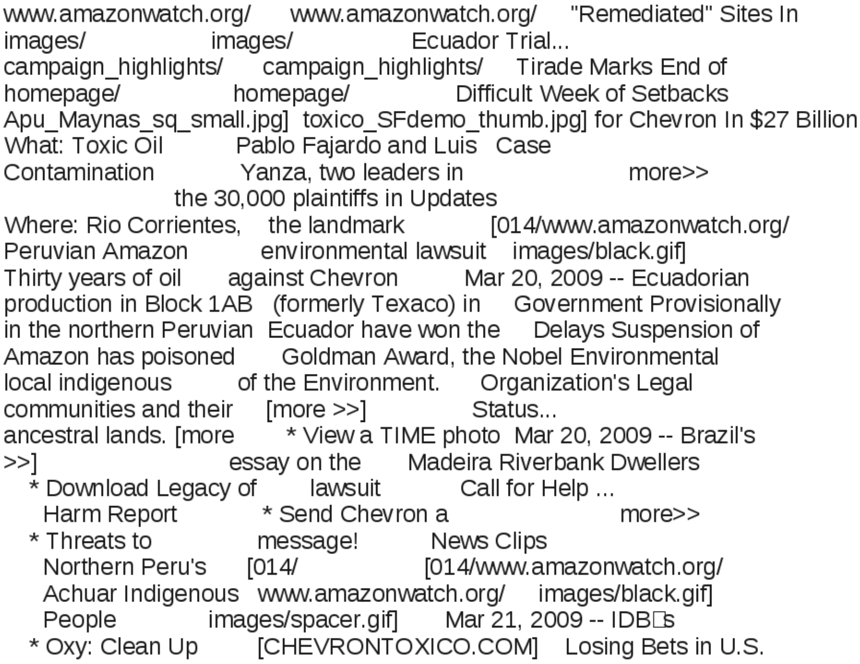
	#coincidences	#predictions	precision
amazon	78	91	0.86
apple	61	70	0.87
argument	22	60	0.37
arm	5	13	0.38
atmosphere	22	54	0.41
bank	83	84	0.99
camel	40	51	0.78
cell	73	81	0.90
columbia	17	28	0.61
cream	14	23	0.61
degree	29	75	0.39
difference	4	29	0.14
disc	24	52	0.46
foreigner	44	71	0.62
fox	34	60	0.57
genesis	23	31	0.74
image	10	31	0.32
jaguar	48	71	0.68
oasis	13	19	0.68
paper	35	66	0.53
party	48	72	0.67
performance	23	50	0.46
pioneer	14	24	0.58
plan	19	77	0.25
police	91	98	0.93
puma	42	52	0.81
rainbow	15	25	0.60
shell	36	70	0.51
shelter	38	63	0.60
skin	49	77	0.64
sort	57	59	0.97
source	27	33	0.82
sun	23	42	0.55
tesla	71	86	0.83
thunder	10	17	0.59
total	14	16	0.88
traffic	76	87	0.87
trapeze	36	49	0.73
triumph	24	37	0.65
yes	9	14	0.64
Totals	**1401**	**2108**	**.67**

Table A.14: WSD experiments: TiMBL-core+freq results



```
      Newsroom | In the Amazon | Capacity building | Take action | About us
     ****** Amazon Watch works to protect the
  rainforest and advance the rights of indigenous   Take Action
      peoples in the Amazon Basin. ******        [014/www.amazonwatch.org/
    * DONATE NOW                              images/black.gif]
    * WATCH VIDEOS                             Mar 19, 2009 -- JOIN US!
    * GET UPDATES                             EAST BAY WORLD SOCIAL FORUM
            "Crude" Press Kit                 REPORTBACK 3/24...
    Understanding the lawsuit behind the film.                      more>>
    [014/www.amazonwatch.org/images/homepage/      Press Releases
              sosamazon.jpg]                  [014/www.amazonwatch.org/
    [014/www.amazonwatch.org/images/homepage/      images/black.gif]
                idb50.gif]                  Mar 18, 2009 -- Civil
 [014/www.amazonwatch.org/images/                  Society Organizations from
 campaign_title.gif]                             13 Countries Highlight
 [014/www.amazonwatch.org/images/spacer.gif]       Inter-American Development
     PERU:                    ECUADOR:          Bank Failure...
 35 YEARS OF OIL           HISTORIC TRIAL AGAINST   Development Model Condemned
   POLLUTION                  CHEVRON           as 50th Bank Meeting
 [014/                  [014/                   Approaches
 www.amazonwatch.org/      www.amazonwatch.org/     Mar 15, 2009 -- Chevron
 images/spacer.gif]        images/spacer.gif]       Lawyers Explode In Anger
 [014/                  [014/                   After More Oil Found at
```

Figure A.1: Textual information in the document 014 for Figure 4.23



```
www.amazonwatch.org/      www.amazonwatch.org/      "Remediated" Sites In
images/                   images/                   Ecuador Trial...
campaign_highlights/      campaign_highlights/      Tirade Marks End of
homepage/                 homepage/                 Difficult Week of Setbacks
Apu_Maynas_sq_small.jpg]  toxico_SFdemo_thumb.jpg] for Chevron In $27 Billion
What: Toxic Oil           Pablo Fajardo and Luis    Case
Contamination             Yanza, two leaders in                   more>>
                          the 30,000 plaintiffs in Updates
Where: Rio Corrientes,    the landmark              [014/www.amazonwatch.org/
Peruvian Amazon           environmental lawsuit     images/black.gif]
Thirty years of oil       against Chevron           Mar 20, 2009 -- Ecuadorian
production in Block 1AB   (formerly Texaco) in      Government Provisionally
in the northern Peruvian  Ecuador have won the      Delays Suspension of
Amazon has poisoned       Goldman Award, the Nobel Environmental
local indigenous          of the Environment.       Organization's Legal
communities and their     [more >>]                 Status...
ancestral lands. [more    * View a TIME photo  Mar 20, 2009 -- Brazil's
>>]                       essay on the              Madeira Riverbank Dwellers
   * Download Legacy of   lawsuit                   Call for Help ...
     Harm Report          * Send Chevron a                        more>>
   * Threats to           message!                  News Clips
     Northern Peru's      [014/                     [014/www.amazonwatch.org/
     Achuar Indigenous    www.amazonwatch.org/      images/black.gif]
     People               images/spacer.gif]        Mar 21, 2009 -- IDB□s
   * Oxy: Clean Up        [CHEVRONTOXICO.COM]       Losing Bets in U.S.
```

Figure A.2: Textual information in the document 014 for Figure 4.24