Knowledge-Based Multilingual Document Analysis

R. Basili† and **R. Catizone** ‡ and **L. Padro** ◇ and **M.T. Pazienza**†

G. Rigau◇ and **A. Setzer**‡ and **N. Webb**‡

F. Zanzotto†

† Dept. of Computer Science, Systems and Production
University of Rome, Tor Vergata
Via di Tor Vergata
00133 Roma, Italy
basili, pazienza, zanzotto@info.uniroma2.it

‡ Department of Computer Science
University of Sheffield
Regent Court, 211 Portobello Street
Sheffield S1 4DP, UK
R.Catizone, A.Setzer, N.Webb@dcs.shef.ac.uk
◇ Departament de Llenguatges i Sistemes Informatics

Universitat Politecnica de Catalunya
Centre de Recerca TALP
Jordi Girona Salgado 1-3
08034 Barcelona, Spain
l.padro, g.rigau@lsi.upc.es

Abstract

The growing availability of multilingual resources, like EuroWordnet, has recently inspired the development of large scale linguistic technologies, e.g. multilingual IE and Q&A, that were considered infeasible until a few years ago. In this paper a system for categorisation and automatic authoring of news streams in different languages is presented. In our system, a knowledge-based approach to Information Extraction is adopted as a support for hyperlinking. Authoring across documents in different languages is triggered by Named Entities and event recognition. The matching of events in texts is carried out by discourse processing driven by a large scale world model. This kind of multilingual analysis relies on a lexical knowledge base of nouns(i.e. the EuroWordnet Base Concepts) shared among English, Spanish and Italian lexicons. The impact of the design choices on the language independence and the possibilities it opens for automatic learning of the event hierarchy will be discussed.

1 Introduction

Modern information technologies are faced with the problem of selecting, filtering, linking and managing growing amounts of multilingual information to which access is usually critical. Our work is motivated by the linking of multilingual information in a wide range of domains. Although this problem appears to be directly related to the Information Retrieval task, we aimed to link articles, not in the broad sense of clustering documents related to the same topic, but rather more specifically linking particular pieces of information together from different documents. Furthermore, we found that IE research, although appropriate for our task, was not designed for the scale/variety of different domains that we needed to process. In general, creating the world model necessary for the addition of a new domain to an IE system is a time-consuming process. As such, we designed an IE system that could be semi-automatically and easily adapted to new domains - a process we will refer to as large scale IE. The key to creating new world models relied on incorporating large amounts of domain knowledge. As a result we selected EuroWordnet as our base knowledge source. EuroWordnet has the advantages of 1) providing the foundation for broad knowledge across many domains and 2) is multilingual in nature. In this paper, we will explain how our system works, how the knowledge base was incorporated and a discussion of other applications that could make use of the same technology.

2 The Application

In the 5th Framework NAMIC Project (News Agencies Multilingual Information Categorisation), the defined task of the system was to support the automatic authoring of multilingual news agencies texts where the chosen languages were English, Italian and Spanish. The goal was the Hypertextual linking of related articles in one language as well as related articles in the other project languages. One of the intermediate goals of NAMIC was to categorise incoming news articles, in one of the three target languages and use Natural Language Technology to derive an 'objective representation' of the events and agents contained within the news. This representation which is initially created once using representative news corpora is stored in a repository and accessed in the authoring process.

2.1 Automatic Authoring

Automatic Authoring is the task of automatically deriving a hypertextual structure from a set of available news articles (in three different languages English, Spanish and Italian in our case). This relies on the activity of event matching. Event matching is the process of selecting the relevant facts in a news article in terms of their general type (e.g. selling or buying companies, winning a football match), their participants and their related roles (e.g. the company sold or the winning football team) Authoring is the activity of generating links between news articles according to relationships established among facts detected in the previous phase.

For instance, a company acquisition can be referred to in one (or more) news items as:

- *Intel, the world's largest chipmaker, bought a unit of Danish cable maker NKT that designs high-speed computer chips used in products that direct traffic across the internet and corporate networks.*
- *The giant chip maker Intel said it acquired the closely held ICP Vortex Computersysteme, a German maker of systems for storing data on computer networks, to enhance its array of data-storage products.*
- *Intel ha acquistato Xircom inc. per 748 milioni di dollari.*
- *Le dichiarazioni della Microsoft, infatti, sono state precedute da un certo fermento, dovuto all'interesse verso Linux di grandi ditte quali Corel, Compaq e non ultima Intel (che ha acquistato quote della Red Hat) ...*

The hypothesis underlying Authoring is that all the above news items deal with facts in the same area of interest to a potential class of readers. They should be thus linked and links should suggest to the user that the underlying motivation is that they all refer to *Intel acquisitions*.

3 The NAMIC Architecture

The NAMIC system uses a modularised IE architecture whose principal components, used to create the IE repository, are morpho-syntactic analysis, categorisation and semantic analysis. During Morpho-Syntactic analysis, a modular and lexicalised shallow morpho-syntactic parser (Basili et al., 2000b), provides the extraction of dependency graphs from source sentences. Ambiguity is controlled by part-of-speech tagging and domain verb-subcategorisation frames that guide the dependency recognition phase. It is within the semantic analysis, which relies on the output of this parser, that objects in the text, and their relationships to key events are captured. This process is explained in more detail in 4. In the next two sections, we will elaborate on the IE engine. For a full description of the NAMIC Architecture see (Basili et al., 2001).

3.1 LaSIE

In NAMIC, we have integrated a key part of the Information Extraction system called LaSIE (Large-scale Information Extraction system, (Humphreys et al., 1998)). Specifically, we have taken the Named Entity Matcher and the Discourse Processor from the overall architecture of LaSIE. The roles of each of these modules is outlined below.

3.1.1 Named Entity Matcher

The Named Entity (NE) Matcher finds named entities (*persons*, *organisations*, *locations*, and *dates*, in our case) through a secondary phase of parsing which uses a NE grammar and a set of gazetteer lists. It takes as input parsed text from the first phase of parsing and the NE grammar which contains rules for finding a predefined set of named entities and a set of gazetteer lists containing proper nouns. The NE Matcher returns the text with the Named Entities marked. The NE grammar contains rules for coreferring abbreviations as well as different ways of expressing the same named entity such as Dr. Smith, John Smith and Mr. Smith occurring in the same article.

3.1.2 Discourse Processor

The Discourse Processor module translates the semantic representation produced by the parser into a representation of instances, their ontological classes and their attributes, in the XI knowledge representation language (Gaizauskas and Humphreys, 1996).

XI allows a straightforward definition of cross-classification hierarchies, the association of arbitrary attributes with classes or instances, and a simple mechanism to inherit attributes from classes or instances higher in the hierarchy.

The semantic representation produced by the parser for a single sentence is processed by adding its instances, together with their attributes, to the discourse model which has been constructed for a text.

Following the addition of the instances mentioned in the current sentence, together with any presuppositions that they inherit, the coreference algorithm is applied to attempt to resolve, or in fact merge, each of the newly added instances with instances currently in the discourse model.

The merging of instances involves the removal of the least specific instance (i.e. the highest in the ontology) and the addition of all its attributes to the other instance. This results in a single instance with more than one realisation attribute, which corresponds to a single entity mentioned more than once in the text, i.e. a coreference.

The mechanism described here is an extremely powerful tool for accomplishing the IE task, however, in common with all knowledge-based approaches, and as highlighted in the introduction to this paper, the significant overhead in terms of development and deployment is in the creation of the world model representation.

4 Large-Scale World Model Acquisition

The traditional limitations of a knowledge-based information extraction system such as LaSIE have been the need to hand-code information for the world model - specifically relating to the event structure of the domain. This is also valid for NAMIC. To aid the development of the world model, a semi-automatic boot-strapping process has been developed, which creates the event type component of the world model. To us, event descriptions can be categorised as a set of regularly occurring verbs within our domain, complete with their subcategorisation information.

4.1 Event Hierarchy

The domain verbs can be selected according to statistical techniques and are, for the moment, subjected to hand pruning. Once a list of verbs has been extracted, subcategorisation patterns can be generated automatically using a combination of weakly supervised example-driven machine learning algorithms. There are mainly three induction steps. First, syntactic properties are derived for each verb, expressing the major subcategorisation information underlying those verbal senses which are more important in the domain. Then, in a second phase, verb usage examples are used to induce the semantic properties of nouns in argumental positions. This information relates to selectional constraints, independently assigned to each verb subcategorisation pattern. Thus, different verb senses are derived, able to describe the main properties of the domain events (e.g. *Companies acquire companies*). In a third and final phase event types are derived by grouping verbs according to their syntactic-semantic similarities. Here, shared properties are used to generalise from the lexical level, and generate verbal groups expressing specific semantic (and thus conceptual) aspects. These types are then fed into the event hierarchy as required for their straightforward application within the target IE scenario.

4.1.1 Acquisition of Subcategorisation Patterns

Each verb v is separately processed. First, each local context (extracted from sentences in the source corpus) is mapped into a feature vector describing:

- the verb v of each vector (i.e. the lexical head of the source clause);
- the different grammatical relationships (e.g. `Subj` and `Obj` for grammatical subject and objects respectively) as observed in the clause;
- the lexical items, usually nouns, occurring in specific grammatical positions, e.g. the subject Named Entity, in the clause.

Then, vectors are clustered according to the set of shared grammatical (not lexical) properties: *Only* the clauses showing the same relationships (e.g. all the `Subj`-$verb$-`Obj` triples) enter in the same subset V_i. Each cluster thus expresses a specific grammatical behaviour shared by several contexts (i.e. clauses) in the corpus. The shared properties in V_i characterise the cluster, as they are necessary and sufficient membership conditions for the grouped contexts.

As one context can enter in more than one cluster (as it can share all (or part) of its relations with the others), the inclusion property establishes a natural partial order among clusters. A cluster V_i is included in another cluster V_j if its set of properties is larger (i.e. $P_i \supset P_j$) but it is shown only by a subset of the contexts of the latter V_j. The larger the set of membership constraints is, the smaller the resulting cluster is. In this way, clusters are naturally organised into a lattice (called *Galois lattice*). Complete properties express for each cluster candidate subcategorisation patterns for the target verb v.

Finally, the lattice is traversed top-down and the search stops at the more important clusters (i.e. those showing a large set of members and characterised by linguistically appealing properties): they are retained and a lexicon of subcategorisation structures (i.e. grammatical patterns describing different usages of the same verb) is compiled for the target verb v. For example, `(buy, [Subj:X, Obj:Y])` can be used to describe the transitive usage of the verb buy. More details can be found in (Basili et al., 1997).

4.1.2 Corpus-driven Induction of Verb Selectional Restrictions

The lattice can be further refined to express semantic constraints over the syntactic patterns specified at the previous stage. A technique proposed in (Basili et al., 2000a) is adopted by deriving semantic constraints via synsets (i.e. synonymy sets) in the WordNet 1.6 base concepts (part of EuroWordNet). When a given lattice node expresses a set of syntactic properties, then this suggests:

- a set of grammatical relations necessary to express a given verb meaning, $R_1, ..., R_k$; and
- references to source corpus contexts C where the grammatical relations are realised in texts.

This information is used to generalise verb arguments. For each node/pattern, the nouns appearing in the same argumental position i (in at least one of the referred examples in the corpus) are grouped together to form a noun set N_i: a learning algorithm based on EuroWordNet derives the most informative EuroWordNet synset(s) for each argument, activated by the N_i members. Most informative synsets are those capable of (1) generalising as many nouns as possible in N_i, while (2) preserving their specific semantic properties. A metric based on conceptual density (Agirre and Rigau, 1995) is here employed to detect the promising, most specific generalisations $sem(N_i)$ of N_i. Then the derived sets for each argument $R_1, ..., R_k$ are used to generate the *minimal* set of semantic patterns $s_1, ..., s_k$ capable of "covering" all the examples in C, with $s_i \in sem(N_i) \quad \forall i$. The sequences express the most promising generalisations of examples C for the subcategorisation $R_1, ..., R_k$. As an example, `(buy, [Agent:Company,Object:Company])` expresses the knowledge required for matching sentences like "*Intel buys Vortex*". Full details on the above process can be found in (Basili et al., 2000a). Notice how `Company` is a base concept in EuroWordNet and it is *shared* among the three languages. It can thus be activated via the Inter-Lingual-Index from lexical items of any language. If included in the world model (as a concept in the `object` hierarchy), these base concepts play the role of a multilingual abstraction for the event constraints.

4.1.3 Induction of Domain event Types via Conceptual Clustering of Verb semantic Patterns

The final phase in the development of a large scale world model aims to link the event matching rules valid for one verb to the suitable event hierarchy nodes. The following semi-automatic process can be applied:

- First, a limited set of high level event types can be defined by studying the corpus and via knowledge engineering techniques (e.g. interactions with experts of the domain);
- then, semantic descriptions of verbs can be grouped automatically, according to the similarity among their corresponding patterns;
- finally, the obtained verb groups can be mapped to the high-level types, thus resulting in a flat hierarchy.

An example of the target event hierarchy is given in figure 1.

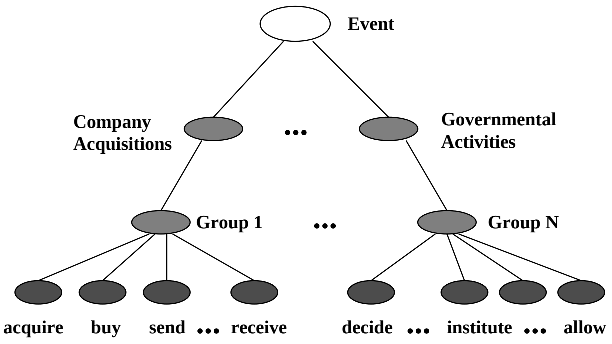

Figure 1: Top levels in the `event` hierarchy vs. verb clusters

Currently, a set of event types (8 main groupings in a financial domain ranging from "*Company Acquisitions*" and "*Company Assets*" to "*Regulation*") have been defined. Within the eight event groupings, we acquired more than 3000 lexicalisations of events. The clustering step has been approached with a technique similar to the Galois lattices, where feature vectors represent syntactic-semantic properties of the different verbs (i.e. pattern $s_i, ..., s_k$ derived in the previous phase). All verbs are considered[1] and the obtained clusters represent semantic abstractions valid for more than one verb. The following is an example of the grouping of the verbs *acquire* to *win*.

```
cluster(141,[acquire,buy,catch,
       contribute,earn,gain,hire,
       issue,obtain,offer,order,
       pay,reach,receive,refer,
       secure,sell, serve,trade,
       win]).
patt(141, [
 arg('Obj',
    ('measure quantity amount quantum'
      ,0),
     'abstraction '),
 arg('Subj',
    ('social_group',0),
    'entity something ')
   ]).
```

The above cluster expresses a conceptual property able to suggest a specific event subtype. Thus, manual mapping to the correct high-level concept ("*Company acquisition*" event type) is made possible and more intuitive. As semantic constraints in event types are given by base concepts, translations into Italian and Spanish rules (for example: `(acquistare, [Agent:Company,Object:Company])`) are possible. They inherit the same topological position in the `event` ontology. Accordingly, the world model has a structure (i.e. the main `object` and `event` hierarchies) which is essentially language independent. Only the lowest levels are representative of each language. Here, a language specific lexicalisation is required. The advantage is that most of the groups derived for English can be retained for other languages, and a simple translation suffices for most of the patterns. Lexicalisations are thus associated with the language independent abstractions (i.e. matching rules over parsed texts) which control the behaviour of instances of these events in the discourse processing.

The integrated adoption of EuroWordNet and the automatic acquisition/translation of verb rules is thus the key idea leading to a successful and quick development of the large scale IE component required for automatic authoring.

4.2 Object Hierarchy

In typical Information Extraction processing environments, the range of objects in the text is expected to be as limited and constrained as the event types. For example, when processing 'management succession' events (MUC-6, 1995), the object types are the obvious *person*, *location*, *organisation*, *time* and *date*. Intuitively however, if the need was to process the entire output of a news gathering organisation, it seems clear that we must be able to capture a much wider range of possible objects which interact with central events. Rather than attempt to acquire all of this object information from the corpus data, we instead chose to use an existing multilingual lexical resource, EuroWordNet.

4.2.1 EuroWordNet

EuroWordNet (Vossen, 1998) is a multilingual lexical knowledge (KB) base comprised of hierarchical representations of lexical items for several European languages (Dutch, Italian, Spanish, German, French, Czech and Estonian). The wordnets are structured in the same way as the English WordNet developed at Princeton (Miller, 1990) in terms of synsets (sets of synonymous words) with basic semantic relations between them.

In addition, the wordnets are linked to an Inter-Lingual-Index (ILI), based on the Princeton WordNet 1.5. (WordNet 1.6 is also connected to the ILI as another English WordNet (Daude et al., 2000)). Via this index, the languages are interconnected so that it is possible to go from concepts in one language to concepts in any other language having similar meaning. Such an index also gives access to a shared top-ontology and a subset of 1024 Base Concepts (BC). The Base Concepts provide a common semantic framework for all the languages, while language specific properties are maintained in the individual wordnets. The KB can be used, among others, for monolingual and cross-lingual information retrieval, which was demonstrated by (Gonzalo et al., 1998).

4.2.2 EuroWordNet as the Object Ontology

The example rules shown in the previous section relate to Agents which conveniently belong to a class of Named Entities as would be easily recognised under the MUC competition rules (*person*, *company* and *location* for example). However, a majority of the rules extracted automatically from the corpus data involved other kinds of semantic classes of information which play key roles in the subcategorisation patterns of the verbs.

In order to be able to work with these patterns,

[1] Initial partitions according to the Levin classification (Levin, 1993) are adopted. A partition of the verbs is built for each of the Levin classes and conceptual clustering is applied internally to each group.

it was necessary to extend the number of semantic classes beyond the usual number of predefined classes, across a variety of languages.

Representing the entirety of EWN in our object hierarchy would be time consuming, and lead to inefficient processing times. Instead we took advantage of the Base Concepts (Rodriquez et al., 1998) within EWN, a set of approximately 1000 nodes, with hierarchical structure, that can be used to generalise the rest of the EWN hierarchy.

These Base Concepts represent a core set of common concepts to be covered for every language that has been defined in EWN. A concept is determined as important (and is therefore a base concept) if it is widely used, either directly or as a reference for other widely used concepts. Importance is reflected in the ability of a concept to function as an anchor to attach other concepts.

The hierarchical representation of the base concepts is added to the object hierarchy of the NAMIC world model. Additionally, a concept lookup function is added to the namematcher module of the NAMIC architecture. This lookup takes all common nouns in the input, and translates them into their respective EWN Base Concept codes.

This process was reversed in the event rule acquisition stage, so that each occurrence of a object in a rule was translated into a Base Concept code. This has two effects. Firstly, the rules become more generic, creating a more compact rule base. Secondly, given the nature of the inter-lingual index which connects EWN lexicons, the rules became language independent at the object level. Links between the lexicalisations of events are still required, and at present are hand-coded, but future development of the verb representations of WN might eliminate this.

In summary, this new, expanded WM covers both the domain specific events and a wide range of agents, and can be acquired largely automatically from corpus data, and used to process large amounts of text on a spectrum of domains by leveraging existing multilingual lexical resources.

5 Discussion and Future Work

The NAMIC system was created to provide an environment for automatic hypertextual authoring of multilingual news articles. In order to address that task, we created language processors in three languages (English, Italian and Spanish) which allows us to create a database of conceptually analysed text. The ability to analyse text in this way is vital for the authoring process, but is also applicable to a wide range of technologies, including Information Retrieval in general, and Question-Answering in particular.

Information Retrieval (Spark Jones and Willett, 1997; Rijsbergen, 1979), or document retrieval as it is in practice, is a well used, robust technology which allows users to access some subset of documents by means of a set of keywords. However, the retrieval of answers to questions by keywords, whilst easy to implement, suffers by their restrictive nature. For example, a keyword based retrieval mechanism would be unable to distinguish between the queries *who killed Lee Harvey Oswald?* and *who did Lee Harvey Oswald kill?*, operating as they do by reducing these queries to a bag of stemmed words. By accessing the kind of knowledge base that we created in the Namic project where events and their relations are explicitly represented, an IR system would be able to distinguish between the above two queries or any other queries that require this kind of data mining.

One possible future extension of the NAMIC scenario, is to move from only allowing users to *browse* through a space of connected articles to a system that supports journalists in the creation of news articles. State of the art techniques for searching, analysing, authoring and disseminating information in the news domain originating from diverse language sources are needed in order to support the working activities of authors (i.e. the journalists) within a complex environment for searching, elaborating and delivering news. The information so derived will enter the dissemination process (archives to the agencies and/or Web channels) and enhanced presentation to the user will be supported in a way that it can be readily understood, accepted, rejected or amended as necessary.

Reporters covering the early stages of a ”breaking” story rely on a format of questions. Typically, these questions include: What? Where? Who? When? But, although definitions of a news story include the originality of the event (”Something that happened today which did not happen yesterday”), coverage also relies on archives. Checks made in the potentially multilingual archives - increasingly comprised of digital resources - make up one of the most important phases in reporting. If such a search path can be imitated by a computer, this would greatly enhance the speed and accuracy of archive searches. For example, in the immediate aftermath of a crash involving a passenger airliner, a number of simple questions and answers may be addressed to the archive. Has this type of aircraft crashed before? If so, what happened? How many fatalities have there been in incidents involving this type of aircraft? Has there been a crash before at this airport? What are the main characteristics of this aircraft? What are those of the airport? Answers to these questions may prompt a series of subsidiary questions.

The depth of interpretation which an experienced and educated journalist can bring to events cannot hope to be imitated by a computer, at least for some considerable time. However, what does seem possible is that a computerised assistant, a sort of electronic cub reporter, could assist the human journalist by finding and collating relevant archival materials in an intelligent fashion - i.e. without precise, low-level instruction from the journalist. This multilingual question-answering task would be aided by the development the proposed system.

In conclusion, we believe that the creation of a sophisticated knowledge base resource can benefit many Information Technology applications - IR and Question Answering to name two. We were able to create such a resource in the NAMIC project by implementing a scalable IE system containing a robust world model based on EuroWordnet. We feel that this kind of automatic resource building will play a significant part of future IT applications.

6 Acknowledgements

This research is funded by the European Union, grant number IST-1999-12392. We would also like to thank all of the partners in the NAMIC consortium especially Yorick Wilks.

References

E. Agirre and G. Rigau. 1995. A Proposal for Word Sense Disambiguation using Conceptual Distance. In *International Conference "Recent Advances in Natural Language Processing" RANLP'95*, Tzigov Chark, Bulgaria.

R. Basili, M.T. Pazienza, and M. Vindigni. 1997. Corpus-driven unsupervised learning of verb subcategorization frames. In M. Lenzerini, editor, *AI*IA 97: Advances in Artificial Intelligence, Lecture Notes in Artificial Intelligence n., 1321*. Springer Verlag, Berlin.

R. Basili, M.T. Pazienza, and M. Vindigni. 2000a. Corpus-driven learning of Event Recognition Rules. In *Proceedings of Machine Learning for Information Extraction workshop, held jointly with the ECAI2000*, Berlin, Germany.

R. Basili, M.T. Pazienza, and F.M. Zanzotto. 2000b. Customizable Modular Lexicalized Parsing. In *Proceedings of the 6th International Workshop on Parsing Technology, IWPT2000*, Trento, Italy.

R. Basili, R. Catizone, L. Padro, M.T. Pazienza, R. Rigau, A. Setzer, N. Webb, Y. Wilks, and F.M. Zanzotto. 2001. Multilingual Authoring: the NAMIC Approach. In *Proceedings of the Workshop on Human Language Technology and Knowledge Management (at ACL-EACL 2001)*, Toulouse, France.

J. Daude, L. Padro, and R. Rigau. 2000. Mapping WordNets using Structural Information. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics ACL'00*, Hong Kong, China.

R. Gaizauskas and K. Humphreys. 1996. XI: A Simple Prolog-based Language for Cross-Classification and Inheritance. In *Proceedings of the 6th International Conference on Artificial Intelligence: Methodologies, Systems, Applications (AIMSA96)*, pages 86–95.

J. Gonzalo, F. Verdejo, I. Chugur, and J. Cigarran. 1998. Indexing with WordNet Synsets can improve Text Retrieval. In *Proceedings of the COLING/ACL'98 Workshop on Usage of WordNet for NLP*, Montreal, Canada.

K. Humphreys, R. Gaizauskas, S. Azzam, C. Huyck, B. Mitchell, H. Cunningham, and Y. Wilks. 1998. University of Sheffield: Description of the LaSIE-II system as used for MUC-7. In *Proceedings of the Seventh Message Understanding Conferences (MUC-7)*. Morgan Kaufman. Available at http://www.saic.com.

B. Levin. 1993. *English Verb Classes and Alternations*. Chicago, Il.

G. Miller. 1990. Five Papers on WordNet. *International Journal of Lexicography*, 4(3).

MUC-6. 1995. Proceedings of the Sixth Message Understanding Conference (MUC-6). Morgan Kaufman. Available at http://www.saic.com.

C.J. Rijsbergen. 1979. *Information Retrieval*. Butterworths, London.

H. Rodriquez, S. Climent, P. Vossen, L. Bloksma, A. Roventini, F. Bertagna, A. Alonge, and W. Peters. 1998. The Top-Down Strategy for Building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology. *Special Issue on EuroWordNet. Computers and the Humanities*, 32(2-3):117–152.

K. Spark Jones and P. Willett, editors. 1997. *Readings in Information Retrieval*. Morgan Kaufmann, San Francisco, CA.

P. Vossen. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht.