Experiences with a multilingual ontology-based lexicon for news filtering

Weigand, H.; Hoppenbrouwers, S.

Published in:
Proceedings DEXA '98

Publication date:
1998

Citation for published version (APA):
Weigand, H., & Hoppenbrouwers, S. (1998). Experiences with a multilingual ontology-based lexicon for news filtering. In R. R. Wagner (Ed.), *Proceedings DEXA '98* (pp. 160-165). IEEE Press.

Experiences with a Multilingual Ontology-based Lexicon for News Filtering

Hans Weigand and Stijn Hoppenbrouwers

Infolab
Tilburg University
PO Box 90153
5000 LE Tilburg, The Netherlands
email: weigand@kub.nl

Abstract

Ontologies as part of a lexicon are important for many NLP tasks. In this paper, we draw on experiences gained in an ESPRIT project called TREVI, which concerns news filtering and enrichment and includes a English/Spanish multilingual Lexicon. We sketch the way the Lexicon, including an ontology, is constructed, and present the methodology used for adding domain ontologies. In general, the approach is lexicon-driven in the sense that ontology (semantics) and lexicon (word forms) are developed in tandem.

1 Introduction

Ontologies have been subject of investigation in AI for several years. Loosely speaking, an ontology is a database describing the concepts in the world or some domain, some of their properties, and how the concepts relate to each other. An ontology is often organized as a classification hierarchy. Ontologies should be distinguished from domain models that are application-specific: they are intended to be used and reused in many applications, and therefore should be kept minimal. A typical example is an ontology of the temporal domain (days, months, years, etc.) that could be used in all applications in which reasoning with time is an issue.

Ontologies describe concepts, not the way these concepts are expressed in words in a natural language. Therefore it is usually assumed that the ontology is language-independent. There are some problems with this assumption (cf. [12], [2]).

In the first place, there is the philosophical point that it is not possible for people to step outside their linguistic setting. Concepts are shaped through communication between members of a linguistic community, so it is unnatural to disconnect ontology from language.

As can be illustrated by numerous examples, there are concepts that occur in one language and not in another, and languages differ in the way they categorize concepts in the lexicon. Related to this point, it is obvious that we cannot talk about concepts without a representation. We can *distinguish* between the word and the concept, and say, for example, that two words denote the same concept, but words (or even artificially created formal predicates) are indispensable as handles.

There are also some practical problems with language-independent ontologies. If we want to have a broad scope rather than a specialist domain, it seems that the best way to start is with available machine-readable semantic dictionaries, such as WordNet [8], although it is clear that such a dictionary has a language bias (in this case, English).

Another point is the necessary distinction between (shared) concepts and the means by which they can be expressed. In the Cyc Project, for example, there is not always a clear separation between knowledge of English word forms and knowledge of concepts [6], resulting in *ad-hoc* solutions and a lack of genericity.

The best approach seems to develop the ontology and the lexicon(s) in parallel, and to stay aware of interlingual differences and the distinction between words and concepts.

Below, we will give a short overview of an NLP project our group is involved in, in which ontologies and lexicons are developed. Next, we will describe our approach to the tasks assigned to us.

2 Background

The current paper is based on experiences in the ESPRIT project TREVI. In this section, we give an overview of the project and the main problems that de multilingual ontology is supposed to solve in this project.

2.1 TREVI: News filtering and enrichment

TREVI is an ESPRIT project (#23311) started in January 1997. Tilburg University is responsible for the TREVI Lexicon Management System and its contents. The project is managed by ITACA (Rome, Italy).

The TREVI Project (Text Retrieval and Enrichment for Vital Information) aims at offering a solution to the problem of *information overflow*, i.e. the difficulty experienced both by both small and large companies and in extracting useful information from large amounts of data coming from the numerous electronic textual information services available at local or global level (Internet, proprietary networks, subscription services, World Wide Web, etc.).

The key result of the TREVI Project will be a set of software tools (the TREVI Toolkit) representing a substantial improvement in the flexible management of distributed textual information sources. The TREVI Toolkit will not rely on simple text-based search tools, but rather combines concept-based search and active data mining techniques to enrich online input text streams. It provides indexation, abstraction, smart correlation with data and knowledge sources, compilation into electronic publication formats, and subscription capability on the results through communication services (for example, HTML document servers on the WWW).

The languages supported by TREVI are Spanish and English. It has been decided to build one multilingual Lexicon Management System (LMS) with a shared (language-independent) semantic part (the concept base) and separate language-dependent morphosyntactic parts (the lexicons). The LMS has to cover both domain-independent lexical semantics and (a limited number of) specific domains related to the demonstrator cases.

The main functions of the LMS are (1) support of the parsing process and (2) support of the subject identification. For the former, the lexicons are the most important, whereas for the latter task, the concept base is crucial.

Subject identification provides the basis for the matching of documents with user profiles that have been set up using the same concept base, which implies the computation of the semantic distance between the document representation and the user profile (in terms of concepts).

2.2 Word sense disambiguation

A central problem in TREVI, as in concept-level NLP in general, is word sense disambiguation. Subject matching on the basis of concepts requires that the concepts are first identified correctly. However, words are often ambiguous. The basic problem is then to select the combination of word senses in a sentence or text that best fits the overall meaning and context of the sentence. Word sense ambiguities can be classified in three types [7]:

1. One sense fits the context, and others are anomalous;
2. Two or more senses are acceptable, but one is better
3. All senses are anomalous, but one must be chosen nevertheless

For example, consider the following Spanish sentence: *Fuentes financieras consultadas cifraron* ... (Financial sources that were consulted estimated ...) The word *fuente* has three senses: source, fountain, or dish. In this case, the first meaning is the correct one. This can be derived by looking at the selection restrictions on the verbs: a fountain or dish cannot be consulted.

For word sense disambiguation, it is essential to have information about selection restrictions (frame role restrictions). Concepts must be organized in a taxonomy so that constraints can be stated concisely. For example, the verb *sell* takes a person or organization as agent. The taxonomy should include that a bank is an organization, so that *bank* can fit in the agent role of *sell*.

What is also important is that the number of concepts in the ontology is kept low. For example, the word *bank* has different senses. In WordNet, it has the senses “financial institution” and “bank building” (among others). It is clear that these senses are closely related; the latter can be viewed as a projection of the first (cf. section 5). Therefore, it is doubtful whether the difference is important enough for the application at hand to maintain it: if the general TREVI user is interested in banks, he should not get messages about river banks, but messages about bank buildings (for example, “a new bank was opened in Bilbao yesterday”) seem to be close enough to include them.

3 Multilingual ontologies

Ontologies built so far tend to remain small. However, when the ontology is supposed to support NLP-related tasks such as news filtering, a different approach is needed. In this section, we describe our approach and the relationship between our ontology and WordNet.

3.1 Ontologies: a lexicon-driven approach

Since the LMS has the task to support the parsing process in TREVI, it should ideally provide full lexical semantics for a substantial vocabulary (say, 100,000 word forms). Since this is not feasible in view of the limited resources available to us, a more subtle approach is needed. This approach makes a distinction between three sets of words (and their underlying concepts): the core, the crowd, and the chosen.

1. The *core* contains basic concepts and the most common words which express them: not only very general categories like CONCRETE or ABSTRACT but

also basic classifications (natural kinds), such as ANIMAL, PERSON, COUNTRY, and also HORSE, MAKE, EAT, etc (cf. [9] [11]). Basic level concepts represent cognitively most salient categories, whereas non-basic concepts are specializations or generalizations of basic-level categories.

The starting point for the TREVI concept base is the set of concepts in WordNet connected to very frequent (lemma count in COBUILD corpus) and familiar words (high number of senses). This set was then analyzed and augmented by hand to ensure language-independence and minimality. On the basis of some experiments with LDOCE, Vossen (1995) has estimated the number of basic-level (nominal) concepts at about 11,000. We expect the core to contain about 3000 nominal concepts. Around 1,500 basic level action concepts will also be included. The conceptual core will be language-independent, but the concepts are linked to word forms in both English and Spanish. In the future, we hope to connect it to other European languages as well.

2. The *crowd* contains only words (carrying various kinds of morphosyntactic information), without any links to the concept base. The crowd is meant to optimally support the syntactic parsing process and hence contains a large number of word forms (including fnction words) in order to minimize encounters with unknown words. It does not lend support to sense disambiguation, as this is reserved for words/concepts part of the ”core” and ”chosen”.

   It is possible in principle to use sources such as WordNet to assign conceptual links even to (part of) the crowd, though within the scope of the TREVI project this was decided against as it would be without immediate use.

3. The *chosen* lexicon and concept base contain elaborate conceptual information for specific domains (the ones relevant for the project). Domain concepts are linked to the core. The structure of domain ontologies is described below.

The advantages of splitting core and crowd are the following: (1) the core provides us with a comprehensible (cognitively relevant) way of structuring the concept set; while (2) language-dependent nuances are not excluded, nor analyzed deeper than strictly necessary for the requirements of the application. When in the future, lexicons for more languages become available (and are required in the system), they can be added to the LMS with minimal effort: we require only the core to be shared. Note that the core is not fixed, and the addition of another language may in principle prompt conceptual extensions.

Also note that we do not analyze all concepts in the same depth. In any case, since there is no principle boundary between ontological and general-world (encyclopedic) knowledge, the boundary will be arbitrary and *situated* (in the sense of [7]).

3.2 Problems with WordNet

WordNet is a useful resource of word senses and is currently applied in several research projects. However, there are some problems:

Overdifferentiation For the more common words, WordNet typically distinguishes more senses then traditional dictionaries. For example, for the verb ”to charge” WordNet gives no less than 24 different senses. In a standard dictionary, around five main senses are given. Many of the senses in WordNet are in fact special usages of main senses. Overdifferentiation is a problem, since it makes it harder to identify the correct sense in a given input text when there is more choice.

Inconsistency There are numerous errors. For example, UK is classified as a kingdom but the Netherlands are not (while in fact it is a monarchy too).

Relevance WordNet contains some very peculiar bits of information. For example, battles are connected with the country where they occurred with a PART OF relation, e.g. the Battle of Maldon is PART OF England. This is a debatable relational link, which also raises the question of where to draw the line between lexical and encyclopedic information.

Incompleteness Some word groups (for example, for biological classifications) are very elaborate, but other domains are underdeveloped. For example, there seems to be no system in the city names and river names that are incorporated.

Separation of verbs and nouns WordNet follows the principle that related verbs and nouns refer to different concepts. However, many verbs have direct or indirect nominalizations; the WordNet approach leads to a duplication of information. In the TREVI approach, concepts are neutral with respect to part-of-speech; a concept can have a verbal or nominal expression, or even both.

Lack of frame semantics WordNet gives only rudimentary information as far as the subcategorization and semantic roles are concerned.

For these reasons, we use WordNet as a resource but develop a separate, “cleaner” ontology for our own purposes.

4 Advanced Domain Analysis

Ontologies, like terminologies in the past, are typically thought of as taxonomic structures. However, in our experience with domain modelling so far, taxonomies should in fact be put less central, as they often are much more arbitrary than the analyst wants to acknowledge. In TREVI, we take an approach in which taxonomies are only secondary.

In accordance with principles of Object-Oriented Analysis (e.g. [5]), *actions* (or events) are taken as central. In OO, an object type is determined by the actions, or methods, that it can perform, not on the basis of its structure. The actions, commonly expressed by action verbs, correspond to practices in the domain that do not change very much over time. On the other hand, many of the terms that label the agents and objects involved can and will be changed often. So whether temporary workers are called employees or not, is something that can be changed over night. But the actions that they perform, and the actions the organization performs on them, remain much more stable.

In principle, the domain analysis methodology we want to adopt follows the following steps: modeling of actions, resulting in an Action Ontology; analysis of terms, resulting in an Object Ontology; analysis of derived terms, resulting in a Terminology. However, for specific applications is possible to skip some phases of analysis if the application it supports does not call for it. For example: in TREVI, highly domain specific action frames are extracted automatically (using the ARIOSTO-LEX system [1]), rendering further action analysis superfluous. Another determining factor is the stability of the domain: when the domain is stable, corresponding to well-established practices, the Object Ontology will also have grown stable and hence an action analysis is less relevant. For these reasons, action modelling has been applied in TREVI in some small cases only (uptill now).

4.1 Action modelling

Action modeling takes the following course. First we try and find the relevant actions in the domain, starting with the action verbs encountered in texts (explicit or hidden, as in nominalizations). For each action, we determine the role or frame structure (agent, patient, etc.) and the selection restrictions [12], which should be filled by basic concepts of the core ontology. They should not be role names (so *person* instead of *employee*).

The total set of essential actions and the roles makes up a conceptual network comparable to an Entity Relationship diagram but better (linguistically) motivated. We call it the Action Ontology. The actions are categorized according to prototypical event structures, such as TRANSFER, TRANSFORM, TRANSPORT and ACT ON [4]. No further conceptual information is defined at this point.

4.2 Term analysis

We analyze a list of terms one by one. A term is a simple or compound Noun Phrase expressing some kind of entity type. Nominalizations (disguised actions) and reified properties are not taken into account: the former have been treated by the Action Model, the latter will be treated in step (3). Also, we do not include instances here, such as country names, but they will be included in (a special section of) the LMS. For polysemous terms, the procedure is applied to each sense (although it should be attempted to unify different senses under one prototype wherever possible).

A distinction is made between the following term classes:

names: terms that express basic level concepts. At this stage, we define for each name a concept frame containing prototypical information, such as a TELIC role (KNIFE telic CUT), PART-OF and CAUSE relations. There may be more complex specific roles expressed by an action. Prototype information seems most appropriate to names, because they correspond to rich concepts. For the other term classes below, the relevance of prototype information remains to be considered.

roles: terms that express a role of an entity in some action, for example, EMPLOYEE and EMPLOYER. Roles can be defined in terms of the action model, for example: an EMPLOYER isa (PERSON or ORGANIZATION) who does EMPLOY a PERSON (in this case, the term is a superordinate and a role), or DIRECTOR isa PERSON who does DIRECT an ORGANIZATION (in this case, the term is is a role and a subordinate). Roles are fully defined by the relationship they express (in this case, EMPLOY and DIRECT, respectively). Since these relationships are made explicit in the Action Model, we can simplify the Term Model by sorting out the roles.

superordinates/generalizations: terms that express a specific property (or capability) that is shared by a number of (basic) concepts. An example is VEHICLE as a generalization of CAR, SHIP etc, or LIQUID as a generalization of WATER, WINE, BLOOD etc). Generalizations are specified by means of the property - this is often based on an action, like ”transport” or ”flow” - they express, and this property is then inherited to its hyponyms.

subordinates/specializations: terms that classify basic level concepts according to some property. For example, PERSON can be specialized to MAN and WOMAN according to the property GENDER. Specializations are specified by means of the basic level concept they start from and the properties that they express. Note that roles are treated separately. Subordinates inherit properties of the basic concept they are attached to, but such prototypical information can be overruled.

component/part: terms that are defined by means of a PART-OF relation to a basic (from the point of view of aggregation) concept. For example, NOSE is defined as PART-OF a PERSON.

group: terms that are defined as an aggregation of basic concepts. For example, TEAM is defined as a GROUP-OF PERSON.

The total set of essential object concepts can be organized in the form of two hierarchies, or tree diagrams: a hyponymy hierarchy and a meronymy hierarchy. In such a hierarchy, the distinction between basic and non-basic is blurred. However, the distinction can still play a role, for example, in the presentation of query results. When a user wants to know what a SCHNAUTZER is, the reply should be that it is a kind of DOG, and not just a list of all hypernyms.

4.3 Terminology

The Terminology is defined here as the set of derived or analytical terms. Analytical terms express some property determined by some definition. For example, AGE is defined as the number of years after the BIRTH-event; NET INCOME is defined as the GROSS SALARY minus taxes, where GROSS SALARY in turn is defined as the MONEY EARNED-BY PERSON. Somehow a measure function must be provided for each property. This can be done qualitatively by giving the positive and negative antonymes (OLD/YOUNG), or by a specific measure function (e.g. MONEY AMOUNT).

5 Conclusion

In this paper, we provided an overview of the approach that we follow for the construction of a multilingual Lexicon Management System in context of the TREVI Project. In the LMS, ontologies and lexicons are developed in tandem. We described the way we set up a general concept base, including a core ontology parallel to a multilingual lexicon, and also how we want to set up in-depth ontologies for specific domains. Ultimately, the general concept base could be replaced by a (large) set of domain ontologies, but this will not be possible due to the limitations of this project.

The core of basic-level concepts has been linked to both Spanish and English lexicals. We are in the process of implementing a Lexicographers Workbench that supports the advanced domain analysis as described above. The tool can work with input extracted from text corpora but also with manual input, possibly derived from group decision support sessions. With the help of this tool, four domains will be analyzed in the course of 1998. The resulting domain lexicons will be used when the integrated TREVI system will be put to the test on actual news corpora.

Acknowledgments

The TREVI project is supported by the European Commission from 1997 till 1999. For more information: http://www.itaca.it/itaca/it/Trevihom.htm.

References

[1] R. Basili, M.T. Pazienza, P. Velardi, “An emperical symbolic approach to natural language processing,” in: *Artificial Intelligence*, 85 59-99, 1996.

[2] J. Bateman “Ontology construction and natural language,” in: N. Guarino, R. Poli (eds), *Proc. of the Int. Workshop on Formal Ontology in Conceptual Analysis and Knowledge Representation*, Padova, March 1994

[3] S.C. Dik, *The Theory of Functional Grammar,* Foris, Dordrecht, 1989.

[4] R. Jackendoff, *Semantics and Cognition*, MIT Press, Cambridge, 1983.

[5] G. Kristen, *Object-Orientation: The KISS-method: From Information Architecture to Information Systems,* Addison-Wesley, 1994.

[6] K. Mahesh, S. Nirenburg et al, “An Assessment of Cyc for Natural Language Processing” *MCCS-96-302*, New Mexico State University, 1996.

[7] K. Mahesh, “Ontology Development for Machine Translation: Ideology and Methodology” *MCCS-96-292*, New Mexico State University, 1996.

[8] G.A. Miller, “WordNet: A Lexical Database for English” *Communication of the ACM,* 38(11), Nov 1995.

[9] E. Rosch, “Classification of real-world objects: origins and representation in cognition”, in: P.N. Johnson-Laird and P.C. Wason (eds), *Thinking: readings in cognitive science*, Cambridge Univ Press, 1977.

[10] J. Sowa, *Conceptual Structures: Information Processing in Mind and Machine,* Addison-Wesley, 1984.

[11] P. Vossen, *Grammatical and Conceptual Individuation in the Lexicon*, Ph.D. Thesis, Univ of Amsterdam, 1995.

[12] H. Weigand, *Linguistically Motivated Principles of Knowledge base Systems,* Foris, Dordrecht, 1990.