Lexical Databases as a Base for Broad Coverage Ontologies

Pavel Smrz

Faculty of Informatics, Masaryk University
Botanicka 68a
60200 Brno, Czech Republic
smrz@fi.muni.cz

Abstract

The paper deals with current lexical databases that are seen as a basis for broad-coverage general-purpose ontologies. Various extensions and refinements of existing multilingual lexical knowledge bases are proposed with the aim of improving the capabilities of these resources. The main goal lies in the effort to gain a better lexical knowledge representation, which is crucial to coping with the requirements of the Semantic Web. The final section discusses the question of how lexical knowledge bases can be shared and combined. It presents the designed and implemented system WOMANISER that is able to merge independently developed parts of ontologies, check inconsistencies and report errors. The paper ends with the future directions of this research.

Introduction

Underpinning the current work is the belief that the Semantic Web will be able to transform the current World Wide Web into an environment with a clear semantics understandable by computers as well as humans. The present-day research suggests that these aims would be accomplished by a careful design of ontologies in the area of specialized, domain-specific ontologies. On the other hand, domain-independent, general-purpose ontologies present much more serious obstacles. This situation can be also reflected by the popularity of "bottom-up" and "middle-out" strategies (Niles 2001) in building middle level domain ontologies and lower-level application ontologies – it is much easier to design such an ontology from scratch.

It is clear that to be able to enter the semantic web era successfully, the problems of building general-purpose ontologies must be solved. Only the domain-independent ontologies can actually provide a base that would replace current favorable search engines like Google, AltaVista etc. with a comparable coverage.

The design of a broad-coverage general-purpose ontology is extremely labor-intensive when prepared from scratch or derived by merging existing resources. It is always difficult to find a wide agreement or solve all the inconsistencies of different ontologies. The development of the IEEE Standard Upper Ontology (SUO) and the hot debate about 3D or 4D orientation of the emerging ontology on the SUO mailing list demonstrates this problem.

However, the SUO is intended as the very beginning of general-purpose standard ontologies and will contain about 2000 terms only. The progress in the development of adjoining, large, general-purpose standard ontologies proved to be much slower. The most promising approach here is the effort to clean-up, refine and merge the existing resources – WordNet (http://www.cogsci.princeton.edu/~wn/), HowNet (http://www.keenage.com/zhiwang/e_zhiwang.html), CoreLex (http://www.cs.brandies.edu/~paulb/CoreLex/overview.html), the available part of Cyc (http://www.cyc.com/), etc. These databases are known under different names – ontologies, semantic networks, lexical knowledge bases, ... and their primary objective was often very different from providing a standard ontology (modeling the human mental lexicon in the case of WordNet, regular polysemy in CoreLex, etc.) This also accounts for the above-mentioned complexity of the standardization process.

The European project "Intelligent Knowledge Fusion" (http://www.nomos.it/html_files/ikf.html) that aims at the development of a general reference ontology linked to lexical resources such as WordNet, seems to pass to the most advanced point in the refinement of standard ontologies. The methodology of formal relations and formal properties is surely fruitful. However, the question is whether the current state of existing lexical knowledge bases such as WordNet, provides an adequate starting point for knowledge representation.

The cited lexical resources have often been criticized from different points of view and several improvements have been suggested previously. However, these proposals, to the best of our knowledge, have not yet led to a draft of a new conceptual structure of lexical knowledge bases.

This paper proposes various refinements of current multilingual lexical knowledge bases thereby taking the first step to a new structure of these lexical resources. The essential issue here is the multilinguality of knowledge bases. We believe that it is one of the key components of the wide applicability of the ontologically based semantic web. Our considerations are based on the experience gained in our participation on the EuroWordNet project (parallel wordnets for eight European languages – http://www.hum. uva.nl/~ewn/) and the recently started

Balkanet project (four other languages of Balkan countries – http://www. ceid.upatras.gr/Balkanet/). Our research is therefore oriented to the WordNet database and its clones but we believe it is easily applicable to other multi-lingual lexical databases.

The rest of the paper combines several types of lexical knowledge information that seem to be most beneficial in the process of improvement and extension of lexical knowledge resources. The final goal of our research is to propose a broadly applicable conceptual structure of lexical knowledge bases that, accompanied with well-structured ontologies, will provide the optimal starting point for knowledge understanding and inference.

Lexical Database Refinement

Hierarchical Relations

Most existing lexical knowledge bases and ontologies define hierarchical relations as their key component. We believe, along with Gangemi, Guarino, and Oltramari (2001), that the basic type of this hierarchical relation, known as hyper-hyponymic relation in the context of WordNet, should be divided into elaborated types of different relations. WordNet mixed "instance-of" and "is-a" relations that play different roles in the process of knowledge understanding. We also propose the separation of "taxonymy" from "true hyponymy". Cruse (2000) notes that true hyponymy is a transitive relation, but there are several cases of the taxonymy relation where transitivity seems to break down:

A car-seat is a type of seat.
A seat is a type of furniture.
A car-seat is not a type of furniture.

In contrast to the other proposals, however, we cannot see a crisp dichotomy between the above-mentioned cases. Therefore, we propose integrating multi-value, or even fuzzy logic to the labeling of particular types of relations. This will reflect test criteria as rigidity, identity, dependence (Gangemi et al. 2001) as well as e.g. subtle distinctions of natural stability (Cruse 2000). Currently, we are working with five fuzzy quantifiers only – necessary, expected, possible, unexpected, and impossible, but this set can be extended as needed.

The existing lexical knowledge bases usually expect a fixed, tree-like conceptual structure. Although this can work in knowledge bases dedicated to a special purpose, wide coverage ontologies ask for a more flexible approach. It implies at least "multi-parent" relations (multiple inheritance). In such cases, the conceptual hierarchy does not form a tree structure but a more complex, acyclic structure (DAG). It is already present in some of today's lexical databases but multiple inheritance is not used extensively (particularly due to the difficulty of dealing with multiple inheritance).

Again, our proposal goes even further, claiming the need for dynamic, "on-the-fly" generation of hierarchy. This approach shows to be applicable especially for the non-rigid relations which are not fully covered in OntoClean, for example. This flexible hierarchy can be arrived at based on the attributes assigned to each concept. It allows the instant rebuilding of the hierarchy when a new attribute is added to a particular concept. It also conforms to our strong belief that the correct structure of lexical knowledge bases can be found by an in-depth exploration of distinctive features of related concepts. The distinctive semantic features allow parallel hierarchies to be generated, putting together "female gender" ("lioness" and "tigress"), allowing the differences between "book" as a physical object and "book" as a text to be captured, etc.

A similar mechanism to that of Distinctive Features has been incorporated into our lexical knowledge base to enable domain labels to be captured. Lexical resources such as WordNet usually define a hierarchy of concepts that tries to reflect inherent properties of concepts while omitting the usage of concepts in the same domain. Thus, it is not possible to find information about the connection between "a doctor" and "a hospital" in the WordNet database, for example. However, such information showed to be very useful for many natural language processing applications. It helps the knowledge understanding and inference. Domain labels often cross the usual ontological categorization boundaries so that they call for the same handling as the distinctive features described above.

Polysemy

WordNet-like semantic networks have been often criticized from the word-sense granularity point of view. The word-sense distinguishing examples are usually taken from the homonymous words, "bank" being the most frequently cited. However, homonymy is relative rare on this level and much more widespread polysemy needs very careful exploration. As a case study we are trying to turn Hanks' meaning potentials (Hanks 2002) into an application. Hanks declares that there are no meanings in dictionaries, only "meaning potentials". He states that: "Meaning potentials are composed of components that are not necessarily mutually compatible, since it is not necessarily the case that all components of a word's meaning potential are activated every time it is used to make a meaning."

We are trying to specify meaning potentials for the concepts that are newly included into the hierarchy. The preliminary results in this area suggest that the method of meaning potentials can dramatically reduce the number of different senses of polysemous words, especially in the case of common figurative meanings that usually impair the clarity and apparentness of lexical knowledge bases. On the other hand, problems with connecting the newly described concepts to the existing conceptual hierarchy are much more overwhelming.

Another serious problem arises when one tries to capture certain properties of concepts that are domain dependent or based on expert knowledge. These difficulties are well known even in the field of fully examined animal classification – dolphins and whales are not fishes which does not accord with common-sense intuition. These reasons led us to the incorporation of "degree of expertise" level tags that are able to differentiate between the basic intuition and the expert categorization of concepts. The same procedure has helped us to solve the need of time labels for concepts shifting in time (the basic WordNet ontology we are working with has the 3D orientation).

Multilinguality of Lexical Resources

The Semantic Web cannot be successful without tackling issues of multilinguality. Our experience from previous projects suggests that coping with the problems of lexical resources in more than one language presents new not insignificant obstacles to the development of a clear ontology based knowledge base. The most important issues will be demonstrated by Czech-English examples here but collaborations with other teams show that many of these difficulties have common grounds and are present in many other languages.

The first and the often-discussed problem of multilingual lexical resources touches on lexical gaps. Multilingual ontologies are usually designed in a gradual manner when one language serves as a base and the others map own concepts onto the base ones. However, there are base terms that are not lexicalized in the target language – so called lexical gaps. The English term "condiment" that serves as a hyperonym for "mustard", "seasoning" etc. in the WordNet database has no direct equivalent in Czech, for example.

There are two possible ways of dealing with such a situation. Either a new artificial concept is generated in the language where no direct correspondence with a base concept exists. The new concept does not match any lexicalized terms, consisting only of a definition that describes its meaning. The other possibility lies in the implementation of a special mechanism that handles lexical gaps. In the EuroWordNet project, these situations have been solved by the special kind of links – "the nearest hyperonym", for example. Then, non-lexicalized concepts are not needed and the lexical database can remain homogenous.

The complementary problem to the above-mentioned one occurs when one base concept finds its equivalents in more than one concept in the other language. An example, discussed also in the context of the EuroWordNet project (Vossen 1998), is the English term "wall" that is translated as "die Mauer" or "die Wand" in German. Again, the mechanism that interlinks the concepts between different language systems has to take such cases into consideration and provide means of dealing with them.

If we allow lexical gaps in the conceptual hierarchy of one language, the situation implies an inexact match between the hierarchies in particular languages (some terms present in one language can be "skipped" in the other language system). However, it must be guaranteed that the transitive hypero-hyponymic relation between concepts is language independent, that the relation between two concepts in a monolingual knowledge base will not be reversed in another language. This consistency check is provided by the WOMANISER system that is described in the following section.

Another interesting problem that arises when dealing with multilingual lexical resources is the question of incorporating morphology. The design of English lexical databases usually does not need to take this issue into consideration. However, the inflective and especially derivative morphology plays an essential role in all Slavic languages (e.g. Czech) or agglutinative languages (e.g. Turkish). Here, the standard morphological patterns are very productive and considerable portions of the base ontology can be generated automatically. It is also the reason that led us to incorporate the Czech morphological analyzer into our lexical knowledge base.

Consistency Checking

A typical way of preparing data for lexical knowledge bases is to put together a team of experts, who will build the complete database. However, there are multilingual projects that require a different approach, namely a co-operation of more teams that will work on a shared resource. The key question is then whether there are applications that can support such work or, even, whether it is possible to design and develop software tools that can help manage the coordination of the collaborating teams.

The communication between partners in such a situation can take one of the following forms:

1. Each group works independently on their own independent copy of all data. From time to time, all databases are sent to the appointed coordinator or supervisor who is responsible for the task. A serious disadvantage of this procedure is that all the progress made by one of the participating teams cannot be checked or at least reviewed by another group. This is possible only at the set moment when the supervisor checks all the databases and sends the results to all participants. Moreover, the supervisor checking all the data manually is rather laborious work.

2. The second method is based on a centralized database. All data is stored on a central server, users connect to this server and their modifications are immediately accessible to all authorized parties. The advantage here is obvious immediately – the development can be fully synchronized at any time. But there are also drawbacks, especially the need for a full on-line connection to the central server with the exclusion of independent development. Furthermore, the services of the central database system should be available at any time in this case and the technologies providing the guaranteed services of a server might be too expensive even if one does not hanker after a mission-critical application.

3. The last possibility is a hybrid of the previous two approaches, which tries to gain both, the choice of the

independent distant work as well as the potential for synchronization at an arbitrary time point. Each group can work independently on their data. Every team is then responsible for synchronizing data by sending it to the server in a pre-determined format. The server, not the supervisor, is responsible for processing all possible automatic checks and for reporting any errors to the sender. Thus, the central database can be updated at any time and inconsistencies can be minimized.

The following text describes the tools that enable the third way of work. They were designed and implemented at the Faculty of Informatics, Masaryk University, Brno, and they will be used by other partners in the Balkanet project.

The first tool called VisDic (Visual Dictionary) has been described in detail in (Pavelek 2002) so that only the key feature, which is needed for understanding the function of the following tool, will be repeated here. VisDic records all modifications performed on the lexical database and generates change logs. These change logs called "journals" are text files, which describe how the local copy of the database has been modified – which records have been added, deleted or updated. The same set of operations should be performed on the data stored in the central server.

VisDic also checks many pre-defined conditions and it disallows the spread of inconsistencies to the edited database. However, if more teams modify data separately, only the server where all these data portions will be merged is able to check for consistency. The server-side process called WOMANISER (WOrdnet MANagement Information SERver) implements this feature by journal blending and data versioning.

The pre-defined checks take the form of triggers. Basically, there are two types of possible responses on a trigger fire. The strict ones are errors and the record that caused the violation of a condition will be rejected and reported to the sender or a defined set of users. The user can specify whether the rest of the sent batch of changes passed to the server should be rejected as well or the records should be checked and processed independently one after the other.

The second types of responses – warnings – are liberal, and can be reported in the same way as errors but they do not bring about the rejection of the respective record. There are multiple levels of these warnings and each user can specify which types of warnings from a particular level should be sent to him or her.

Let us summarize the possible responses that can occur when particular actions are performed. We will discuss the case of the interconnection between lexical resources in more than one language by means of common record identifiers – ILI (Inter Language Indices). The three basic operations are an insertion, deletion or modification of a record. All the manipulation with links is considered as the modification of the relevant record that contains a link to another record. Then, the possible responses are:

1. Responses to the operation of record insertion:

a. The modifying record includes a reference to the identifier of a record that is not presented in the server database. (It is likely that the referenced record has been deleted by some previous operation executed by another user.) Such a record will be rejected and the pre-defined type of the "dangling link" error will be sent back.

b. The definition of the record in the given modifying record includes a term or more than one term that are already present in the database (with the same "sense distinguishing identifiers"). Such a record will be rejected and the pre-defined "term duplicity" error will be sent back together with the recommended, next free "sense distinguishing identifiers" for respective terms causing the error.

c. The primary key of the modifying record is already used by another record. This type of error is usually related to the situation when two clients try to insert records for the same concepts at the same time, especially when the ILI identifiers are used directly as primary keys. Such a record will be rejected and the pre-defined "primary key unique constraint violation" error will be sent back together with the actual version of the record with the same identity.

d. The authorised users can also request that they be informed of all changes in the database carried out on the records that are referenced in actual modifying records. It plays a crucial role in the process of consistency checking. This information is provided in the form of various types of warnings; each type corresponds to a respective type of the possible modification performed on the referred records. "Referred record modified" warnings can be extended by precise information about the initial and the final state of the reported modification provided by the journal.

2. Responses to the operation of record deletion:

a. An attempt to delete a record that is not included in the database (the record has been probably deleted by a previous operation executed by another user). The operation will be rejected and the "record does not exist" error will be reported.

b. The deletion will succeed in all other situations if there are no pointers referring the record that is to be deleted. On the other hand, if the database contains references to the record, the type of reference pointing to the record controls the response. The database administrator can specify the intended behavior for each particular type of relation. The simplest approach is to delete all the relations referring to the deleted record automatically. A more elaborated procedure can "re-link" all referring records to the record referred to with the same type of relation by the deleted record. For example, all hyponyms of a deleted record become direct hyponyms of the actual hyperonym of the deleted record. The last possibility is to reject the record deletion if there are records referring to the deleted one in the database. The relevant error message will be "record referred".

3. Responses to the operation of record update (modification):

a. The first two types of responses are equivalent to those described in 1a and 1b, also with the same error messages reported.

b. The next type corresponds to 2a, the record that should be modified could not be found in the database. Again, the “record does not exist” error will be reported.

c. The other type of the response is similar to 1d. The authorised users can again request that they be informed of all the changes in the database carried out by previous operations either right on the actual record or on the records that are referenced in the actual modifying record. “Actual record modified” and “referred record modified” warnings are reported respectively with the possible additional information about the initial and the final state of the reported previous changes.

In addition to such basic warnings and errors, complex conditions can be defined and checked with the help of the system. For example, the vast majority of relations included in the database must not create cycles or even loops – a record must not refer to itself. This situation can, with difficulty, be detected and the modification, which would cause such an inconsistency, can be rejected. Another consistency check can guarantee that a hypero-hyponymic relation between two records in a monolingual database will not be reversed in the database of another language.

Conclusions and Future Directions

The refinements of lexical knowledge bases presented in this paper are applied to improve the quality of the Czech part of the multilingual lexical resource developed under the current Balkanet project. We also strongly believe that the WOMANISER system will be beneficial when employed to manage the coordination of the collaborating teams involved in the project.

There are still many open research problems related to the conceptual design of lexical resources. One of them concerns the attempts to integrate generative concepts to the structure of the knowledge base. Such integration usually calls for dynamic entities in the knowledge structure that can be implemented in the form of generative rules. The co-existence of these dynamic issues together with the much more static information in the standard knowledge base gives one of the directions for our research.

Other topics that will be tackled in our future explorations involve the effort to reduce the demanding work on ontology extensions. Our analysis (Pala and Smrz 2002) of the definitions contained in the standard Czech dictionary SSJC (Dictionary of Literary Czech Language) shows that the structure of the vast majority of these definitions matches a simple standard schema (e.g. genus proximum and distinguishers for nouns). Therefore, a (semi-) automatic procedure adjoining new terms to an existing hierarchy seems promising.

References

Cruse, A. 2000. *Meaning in Language (An Introduction to Semantics and Pragmatics)*. Oxford University Press.

Gangemi, A.; Guarino, N.; and Oltramari, A. 2001. Conceptual Analysis of Lexical Taxonomies: The Case of WordNet Top-Level. In Proceedings of FOIS-2001 (International Conference on Formal Ontology in Information Systems).

Gangemi, A.; Guarino, N.; Masolo, C. and C. Oltramari, A. 2001. Understanding Top-Level Ontological Distinctions. In Proceedings of IJCAI 2001 Workshop on Ontologies and Information Sharing.

Hanks, P. 2002. Lexical Analysis: Corpus, Computing, and Cognition. Ph.D. diss., Faculty of Informatics, Masaryk Univesity Brno, Czech Republic.

Niles, I.; and Pease, A. 2001. Origins of The IEEE Standard Upper Ontology. In Proceedings of IJCAI 2001 Workshop on Ontologies and Information Sharing.

Pala, K.; and Smrz, P. 2002. Glosses in WordNet 1.5 and Their Standardization (The Exercise for Balkanet). Paper Draft.

Pavelek, T. 2002. VisDic – A New Tool for WordNet Editing. In Proceedings of the First International Conference of the Global WordNet Association, Mysore, India.

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