Including deeper semantic information in the Lexical Markup Framework: a proposal

Isabel Segura Bedmar, José L. Martínez Fernández, Paloma Martínez

Department of Computer Science
University Carlos III of Madrid
Avda. Universidad 30, 28911 Leganés, Madrid
{isegura, jlmferna, pmf}@inf.uc3m.es

Abstract

The exploitation of various lexical resources is crucial for many complex Natural Language Processing (NLP) applications. These systems improve their results remarkably if a more intensive exploitation of the lexical and semantic resources is carried out. Therefore, optimizing the production, maintenance and extension of lexical resources is a crucial aspect impacting Natural Language Processing tasks, in particular those related to language understanding. The lexical resources have been built with extensive human effort over years of work, so it would be beneficial to enable the merging of these resources to form extensive global resources and also to define an standard way to interact with this kind of semantic repositories. Lexical Markup Framework (LMF) is a model, sponsored by the International Organization for Standardization, ISO, that provides a common standardized framework for the construction of NLP lexicons. This paper proposes an extension of the semantic part of the LMF metamodel. We hope this extension to improve the metamodel by the inclusion of semantic information considered useful in semantic interpretation of texts, as proved by research in Semantic Role Labeling processes.

Vključevanje globljih pomenskih informacij v LMF (Lexical Markup Framework/Okvir za leksikalno označevanje): predlog

Izkoriščanje različnih leksikalnih virov je ključno za mnogo celovitih aplikacij procesiranja naravnega jezika. Pri teh sistemih se rezultati občutno izboljšajo, če so leksikalni in pomenski viri učinkoviteje izrabljeni. Zatorej je optimiziranje priprave, vzdrževanja in širjenja leksikalnih virov ključno pri nalogah procesiranja naravnih jezikov, posebej tistih, povezanih z razumevanjem jezika. Leksikalni viri so bili zgrajeni z veliko človeškega truda v več letih, zato bi bilo koristno omogočiti njihovo združevanje in tako oblikovati obsežne globalne vire, prav tako pa določiti standardne pristope za delo s tovrstnimi pomenskimi zbirkami. LMF je model, ki ga podpira Mednarodna organizacija za standardizacijo (ISO) in v okviru katerega se pripravlja skupni standardizirani okvir za gradnjo leksikonov za procesiranje naravnega jezika. V članku predlagamo razširitev pomenskega dela metamodela LMF. Upamo, da bo ta razširitev izboljšala metamodel glede vključevanja pomenskih informacij, uporabnih pri pomenski interpretaciji besedila, kot se je to potrdilo pri raziskavi procesiranja oznak pomenskih vlog.

1. Introduction

The goals of a semantic parser are to identify the semantic relations between the words, and the construction of a structure allowing the interpretation of the meaning of the text (Shi and Mihalcea, 2005).

The identification of the semantic roles is a crucial part in the interpretation of texts (Gildea and Palmer, 2002), and therefore is important for information extraction and retrieval, question answering, natural language interfaces etc. (Hacioglu et al., 2003), (Melli et al., 2005).

In the last decade, the work in the information extraction research field has shifted from complex rule-based systems (Alshawi, 1992) to simpler finite-state or statistical systems such as (Hobbs et al., 1997) and (Miller et al., 1998). These systems have been used in the extraction of relations for specific semantic domains such as terrorist events in the framework of the DARPA Message Understanding Conferences. Other commercial systems have incorporated knowledge representation techniques traditionally used in IA, like frames or context-dependent templates.

Nowadays, the challenge is to be able to develop domain-independent systems or, at least, systems easily adjustable to any semantic domain. The semantic role labelling systems use lexical resources like VerbNet (Kipper, Dang and Palmer, 2000), PropBank (Kingsbury, Palmer and Marcus, 2002), or FrameNet (Baker, Fillmore, Lowe, 1998). The semantic role labelling systems improve their results remarkably if a more intensive exploitation of the lexical resources is carried out (Brharati, Venkatapathy and Reddy, 2005). These resources were built with extensive human effort over years of work. Hence, it would be beneficial to enable the merging of these resources to form extensive global resources.

The creation of a standard on lexicons can be a useful aid for the construction and maintenance of the lexical resources, and for their integration into natural language processing systems. LMF (ISO 24613) is a model that provides a common standardized framework for the construction of NLP lexicons. The goals of LMF are: to provide a common model for the creation and use of lexical resources, to manage the exchange of data among these resources, and to enable the merging of large number of individual electronic resources to form extensive global electronic resources.

The aim of this paper consists to propose a set of improvements to LMF model in particular, in the semantic level, in order to improve the creation and the integration of lexical resources. The proposed extension is based on the analysis and study of research works in the fields of Semantic Role Labeling and lexical resources applications, like Information Retrieval, Information Extraction and Question Answering.

In section 2, the semantic roles and some related linguistics theories are treated. In section 3, the main lexical resources are described. In section 4, several previous works on standards for lexical resources are reviewed. In section 4, the model proposed by standard ISO 24613 is described. In section 5, our approach is developed and finally, in section 6, some conclusions are considered.

2. Semantic Roles

The semantic roles describe the semantic relation (non grammatical) that the arguments have with respect to the predicate of a sentence (usually a verb). Other terms used for their denomination are: thematic roles, semantic cases, thematic relations, semantic arguments, etc.

A semantic role describes an abstract function carried out by an element taking part in an action. This abstract function is defined regardless of the syntactic realizations that the element can acquire into a sentence. So, the semantic roles allow for the representation of generic actions, regardless of the language and the diverse grammar resources that a language offers to express the same action (Cook, 1989).

In the following sentences, the semantic roles of the predicate *to break* have different syntactic realizations:

[John $_{\text{Agent}}$] [broke $_{\text{V}}$] [the window$_{\text{Object}}$] with [the hammer $_{\text{Instrument}}$]

[The window $_{\text{Object}}$] [was broken $_{\text{V}}$] by [John $_{\text{Agent}}$]

[The hammer. $_{\text{Instrument}}$] [broke $_{\text{V}}$] [the window $_{\text{Object}}$]

In contrast to the syntactic level, where there is, more or less, agreement among the linguistic community about the syntactic components and their definition, the semantic level does not reach that degree of agreement when semantic roles and their characteristics must be stated.

The majority of abstract roles have been proposed by linguists as part of the *Linking* Theory (Levin and Rappaport, 1996) - the part of grammatical theory that describes the relationship between semantic roles and their syntactic realizations- which is more concerned with explaining generalizations across verbs in the syntactic realizations of their arguments.

The Proto-Role theory is most abstract and was proposed by (Valin, 1993), (Dowty, 1991). This theory has only two roles: Proto-Agent, Proto-Patient.

Fillmore (Fillmore, 1968) proposed a grammar of cases that classified the verbs according to the frames of cases or the necessary roles demanded by a verb. One of the essential elements of the model was a small set of roles universal, that is to say, generic enough to be valid for all the languages.

The more specific roles have been proposed by computer scientists, who are more concerned with the details of the realization of the arguments for specific verbs**.** For example, if a flight information system is considered, some specific roles could be: FROM_AIRPORT, TO_AIRPORT, DEPART_TIME, or verb-specific roles such as EATER and EATEN for the verb *eat*.

Finally, it is important to emphasize the difficulty in the identification of the semantic roles. The main reason is that there is no a direct mapping between the syntax and the semantics.

3. Review of the main linguistic resources containing semantic information

As already stated, the linguistic information is crucial in many NLP tasks. If semantic role labeling systems are considered, the use of this type of resources is essential.

FrameNet is based on the theory of semantic frames (Fillmore, 1976), where each frame corresponds to an interaction and its participants (roles). A frame has an appropriate name to describe the semantic relation defined by the semantic roles. The frame elements (roles) proposed by FrameNet are specific of each frame.

FrameNet includes corpus of annotated sentences with semantic roles. The corpus can be used to learn how to identify semantic relations starting with syntactic structures.

Its main disadvantage is that it does not define selection restrictions for semantic roles. In addition, the coverage of FrameNet (3040 verbs) and its scalability are seriously limited.

PropBank is a corpus in which verbs are annotated with semantic tags, including coarse-grained sense distinctions and predicate-argument structures. PropBank is based on the verbal classification introduced by Levin (Levin, 1993), that assumes there is a strong connection between syntax and semantic. The verbs are grouped together based on their syntactic behaviour and the resulting clusters are coherent from a semantic point of view as all verbs in one Levin class share the same semantic roles. The clusters are formed at a grammatical level according to diathesis alternation criteria. The arguments (roles) of PropBank are specific of each verb.

VerbNet is a verb lexicon providing detailed syntactic-semantic descriptions of Levin classes. As a result, the main hypothesis of VerbNet is that the syntactic frames of a verb are a direct reflection of the underlying semantic.

The main advantage of VerbNet is that it offers a hard generalization of the syntactic behavior of verbs. In addition, VerbNet provides selection restrictions for its roles. A selection restriction marks the semantic category to which the argument's header belongs to. Another remarkable advantage of VerbNet is that each verb entry is already linked to WordNet (Fellbaum, 1998), with a list of possible senses. In addition, it has a wider coverage than FrameNet (4159 verbs, as opposed to 3040 verbs of FrameNet; 2398 defined in both resources).

The main VerbNet drawback is that thematic roles are too generic to capture similar scenarios to those represented by semantic frames of FrameNet.

WordNet is a lexical database of nouns, verbs, adjetives and adverbs. Closed categories (prepositions, conjunctions, etc.) are not represented, as they are considered part of the syntactic knowledge, not of the semantic knowledge. The main disadvantage of WordNet is that it does not codify the syntactic behavior of the verbs.

The lexical resources are scarce but very valuable information. (Shi and Mihalcea, 2005) propose the integration of the lexical resources FrameNet, VerbNet and WordNet. Each of these resources encodes a different kind of knowledge and has its own advantages, so their combination can eventually result in a richer knowledge-base that could enable a more accurate and robust semantic parsing.

Few automatic methods for semantic classification exist, mainly due to the lack of resources with semantic information.

4. Standards for Lexical Resources

Several attempts have been made in the standardization of linguistic processes and resources. This section describes some of the main initiatives in this line.

GENELEX was a EUREKA project that had several aims and one of them was to design a global model to represent all kind of lexical information (for monolingual morphology, syntax and semantics, and multilingual correspondences), in a neutral mood, independent of applications and not directly linked to a particular theory. Furthermore, the project pursued to build adapted tools to create and maintain such lexicons. In addition, the effectively creation of large size lexical data in this model was considered.

EAGLES[1] (Expert Advisory Group on Language Engineering Standards) was an initiative of the European Commission, within DG XIII *Linguistic Research and Engineering* program, which aimed to accelerate the provision of standards for: very large-scale language resources (such as text corpora, computational lexicons and speech corpora); means of manipulating such knowledge, via computational linguistic formalisms, mark up languages and various software tools; means of assessing and evaluating resources, tools and products.

ISLE[2] (International Standards for Language Engineering) is both the name of a project and the name of an entire set of co-ordinated activities regarding the Human Language Technology (HLT) field. ISLE acted under the aegis of the EAGLES. The aim of ISLE was to develop HLT standards within an international framework, in the context of the EU-US International Research Cooperation initiative.

Its objectives were to support national projects, HLT RTD projects and the language technology industry in general by developing, disseminating and promoting de facto HLT standards and guidelines for language resources, tools and products.

MULTEXT provided specific guidance for the purposes of NLP and MT corpus-based research. MULTEXT tackled the definition of a software standard, an essential step toward reusability, and publishing the standard to enable future development by others.

PAROLE[3] was an EU funded project which aimed to build harmonized lexica and corpora in all languages of the Union. This allows multi-lingual links to be made at the same formal linguistic level (morphological, syntactic and semantic) and at the same level of descriptive granularity. The project took account of previous research into encoding lexica and corpora using standard, non-language specific formats

SIMPLE[3] was a project sponsored by the IV European Framework Program. This project represented the first attempt to develop wide-coverage semantic lexicons for a large number of languages, with a harmonized common model that encodes structured "semantic types" and semantic frames.

5. Lexical Markup Framework

The sub committee ISO-C37 elaborated a standard for the management of terminology (Terminology Markup FrameWork, ISO 16642), and later, decided to construct standards for natural language processing. ISO 24613, published under the name “Language resource management – Lexical markup framework”, provides a common model for the creation and use of lexical resources. In addition, the model makes it possible to manage the exchange of data among linguistic resources and to enable the merging of a large number of individual electronic resources to form extensive global resources.

The same specifications are to be used for both small and large lexicons. The descriptions range from morphology, syntax and semantic to translation. The range of targeted NLP applications is not restricted.

The LMF specification complies with the modeling principles of Unified Modeling Language, UML (Rumbaugh, Jacobson, and Booch. 2005) as defined by OMG[4].

LMF is composed of two components: a *core package* which describes the basic hierarchy of information in a lexical entry and some *extensions of the core package* that describe the reuse of the core components in conjunction with the additional components.

In Figure 2, the UML class diagram of the core package is presented. The class *Database* represents the entire resource and is a container for one or more lexicons. The class *Lexicon* is the container for all the lexical entries of the same language within the database.

The *Lexical Entry* is a container for managing the top level language components. As a consequence, the number of single words, multi-word expressions and affixes of the lexicon is equal to the number of lexical entries in a given lexicon. The Form and Sense classes are parts of the Lexical Entry. The Form consists of a text string that represents the word. The Sense disambiguates the meaning and context of a form. Therefore, the Lexical Entry manages the relationship between sets of related forms and their senses.

The current LMF extensions are described as UML packages (Figure 1).

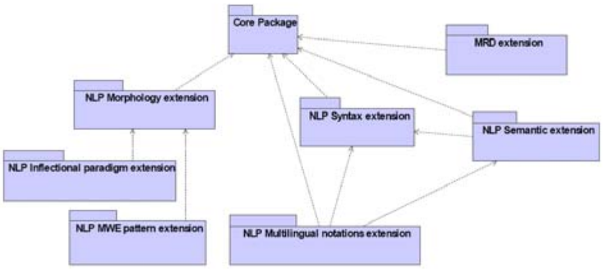
Core Package
MRD extension
NLP Morphology extension
NLP Syntax extension
NLP Semantic extension
NLP Inflectional paradigm extension
NLP MWE pattern extension
NLP Multilingual notations extension

Figure 1: Extensions of the core package UML

Creators of lexicons should select the subsets of the possible extensions that are relevant to their needs. All extensions conform to the LMF core model in the sense that some of the core package classes are extended. An extension cannot be used to represent lexical data regardless of the core package.

[1] http://www.ilc.cnr.it/EAGLES96/home.html

[2] http://www.mpi.nl/ISLE

[3] http://www.ub.es/gilcub/SIMPLE/simple.html

[4] www.omg.org

In Figure 3, the semantic extension of the model is represented. The purpose is to describe one sense and its relations with other senses belonging to the same language. LMF propose several descriptive mechanisms like synsets, predicates, relations or linkage with syntax. Due to the intricacies of syntax and semantics in most languages, the section on semantics comprises also the connection to syntax.

The most important classes shown in Figure 3 are *Sense*, *SemanticPredicate* and *SynSet*. The class *Sense* is described in the core package. *SemanticPredicate* is an element that describes an abstract meaning together with the association with Semantic Arguments (*SemanticArgument)*. A semantic predicate may be used to represent the common meaning between different senses that are not necessarily fully synonyms.

Synset links synonyms. *Synset* is an element that describes a common and shared meaning within the same language. Synset may link senses of two different lexical entries with the same part of speech.

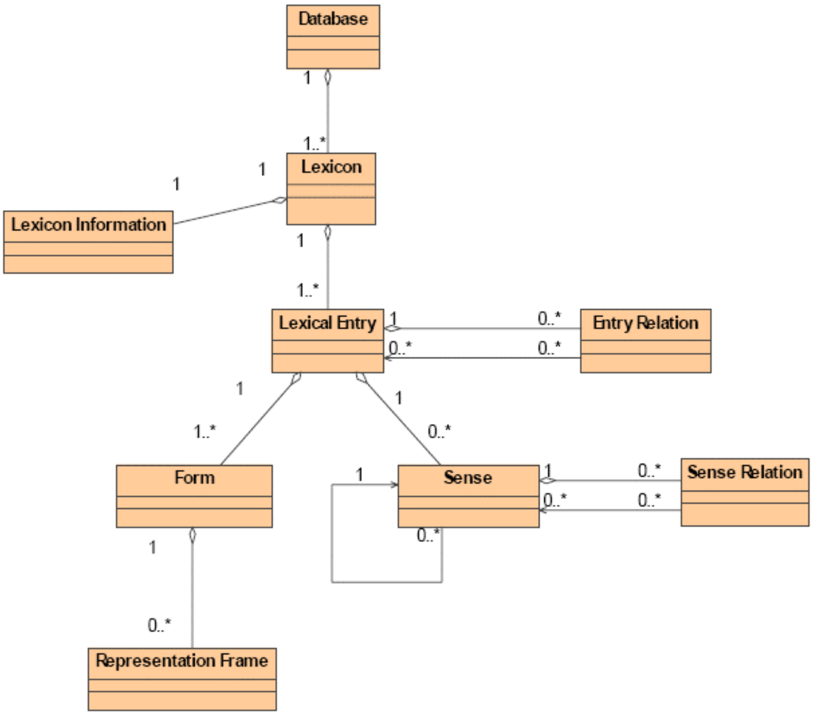

Figure 2. Core Package LMF

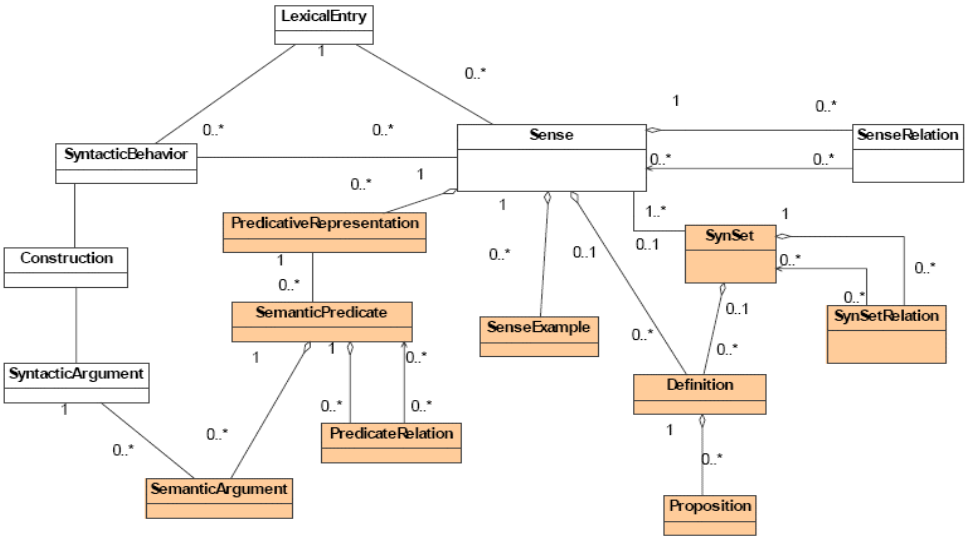

Figure 3: *Extension for Semantic LMF*

6. Proposed classes for the semantic part of LMF

In this section, we propose new classes for the semantic part of LMF. These new classes, denoted as *SemanticClass* and *SelectionalRestrictions* in Figure 4, correspond with the lexical features which have been found to be useful when studying application of lexical resources and semantic role labeling techniques. The final goal of this proposal is to provide the LMF model with the needed elements to comprise existing resources like VerbNet, PropBank, FrameNet and WordNet. These lexical resources, among other things, contain information about thematic roles, which is crucial when the semantic interpretation of texts is considered.

The semantic role labeling system that has obtained the best results until year 2005 was proposed by (Pradhan et al, 2005). The evaluation showed that the features *verb semantic class* and *verb sense* improved its performance. In a previous section, VerbNet and FrameNet resources have been described, both containing representations for groupings of lexical units regarding their semantic meaning. In PropBank or VerbNet, the focus is put on verbs, which can drive the semantic interpretation of a sentence, while FrameNet syntactic categories as nouns, adjectives and others are also considered. These semantic classes cannot be matched against the SynSet class proposed in Figure 3, because this class groups lexical entries with the same syntactic category. This is the reason to include *SemanticClass* in the metamodel. This SemanticClass would include this groups of senses, according to these resources. It is worth mentioning that the frames defined in VerbNet, FrameNet and others could be related with the class SemanticPredicate already defined in the LMF model but these frames can represent one or several semantic classes, depending on the lexical repository considered.

Furthermore, (Brharati, Venkatapathy and Reddy, 2005) showed that the sub-categorization frames help in predicting the semantic roles of the mandatory arguments, thus improving the overall performance. VerbNet defines for each verb class a set of thematic roles (*SemanticArgument)* and a set of syntactic frames (which can be included in the class *SemanticPredicate* defined in LMF) in which these roles are expressed. In addition, VerbNet defines selection restrictions (*Selectional Restrictions*) for the roles of each one of the classes (*+animate, +organization, +communication, +machina, +concrete, + +abstract, etc*). These restrictions are valuable information for determining which arguments correspond with the proper semantic roles. In Figure 4, the class *Selectional Restrictions* is included as an associative class between classes *SemanticClass* and *SemanticArgument* and it must take values from the *SynSet* class.

PropBank does not define semantic selection restrictions for its arguments, but these could be obtained easily, because PropBank and VerbNet are based in the same verbal classification (Giuglea and Moschitti, 2004). In this case, the semantic class of the head word can be useful to determine the correspondence between the syntactic components and the semantic arguments of PropBank. The head word of the noun phrase, and other lexical features, have generated good results in the classification task (Gildea and Jurafsky, 2002), (Pradhan et al., 2005), but these lexical features produce a large dispersion in the data, causing noise in the classification. In this case, it can be useful to use its semantic class (obtained from WordNet) with the purpose of reducing the noise in the classification.

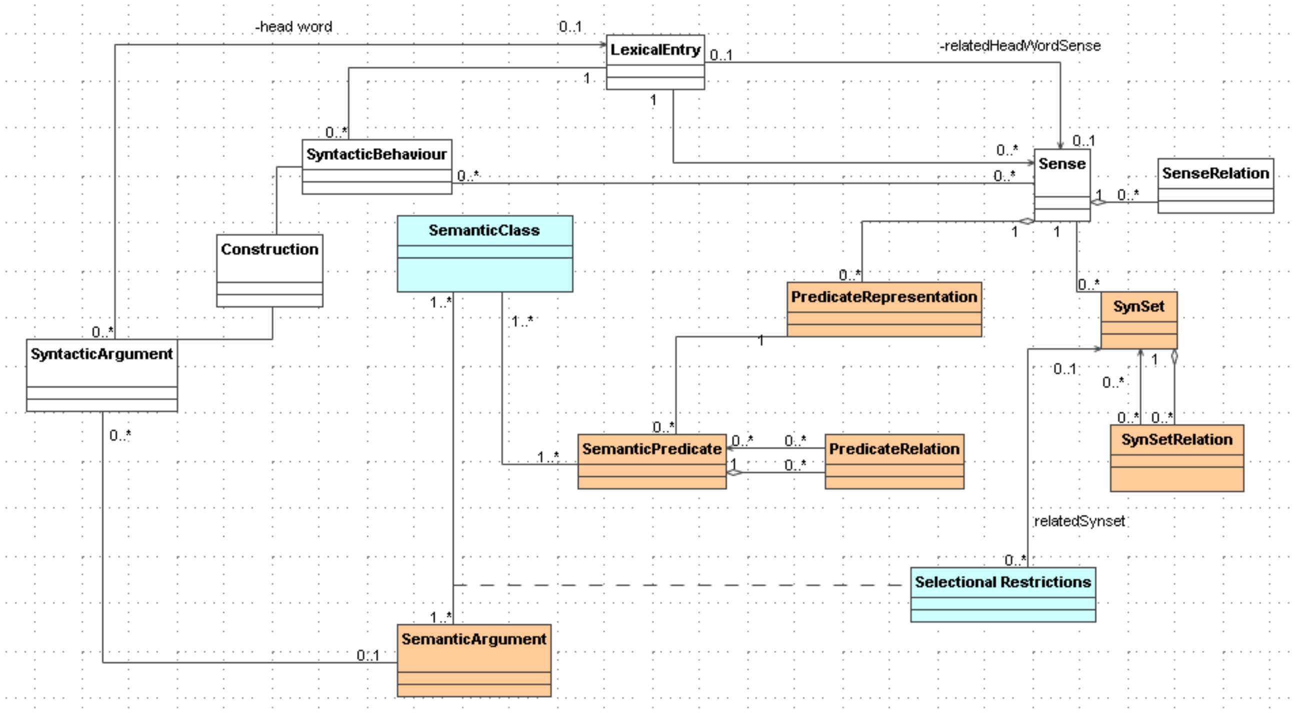

Figure 4. Proposed Semantic Extension of LMF model

In Figure 4, a relation *head word* exists for each noun phrase, so a relation is necessary between the classes *SyntacticArgument,* that represents a noun phrase, and *LexicalEntry,* that represents a word. Furthermore, a new relation *relatedHeadWordSense* is added to represent the semantic class of the head word appropriate for the syntactic argument.

Although specific resources like VerbNet, FrameNet or WordNet have been studied to propose the mentioned new set of classes, the identified elements can be considered generic enough to have a representation in the semantic extension of the LMF metamodel.

7. Conclusion

The availability of semantic information is a crucial issue in the interpretation of texts, and therefore it is important for many tasks related with Natural Language Processing such as Information Extraction, Question Answering or Information Retrieval.

Current lexical resources are small and expensive to produce and maintain. So, it is important to be able to combine them to construct resources with a wider coverage. The creation of a standard fixing the structure and interfaces to be provided by lexical repositories can be a useful aid in the construction and maintenance of these kind of resources and in their integration within Natural Language Processing applications.

In the present work, the LMF standard has been reviewed, and we have proposed several extensions for the semantic part of the LMF metamodel. These extensions are considered to be beneficial for systems where semantic interpretation of texts is pursued.

8. References

Alshawi, H., ed. 1992. The Core Language Engine. *Cambridge, MA: MIT Press.*

Baker, C. F., C. J. Fillmore, J. B. Lowe. 1998. The Berkeley Framenet Project. In *Proceedings of the International Conference on Computational Linguistics (COLING/ACEL-98),* páginas 86-90, Montreal.

Brharati, A., S. Venkatapathy, P. Reddy. 2005. Inferring semantic roles using sub-categorizacion frames and maximum entropy model. In *Proceeding of CoNLL'2005 Shared Task*.

Cook, W. A. 1989. Case Grammar Theory. *GEORGETOWN UNIVERSITY PRESS, WASHINGTON, D.C.*

Dowty, D. 1991. Thematic Proto-roles and Argument Selection. In *language*, 67.

EAGLES, 1996. Evaluation of Natural LanguageProcessing Systems. *Final Report, Center for Sprogteknologi, Copenhagen.*

Fellbaum, C. editor. 1998. WordNet: An Electronic Lexical Database. *Language, Speech and Communications. MIT Press, Cambridge, Massachusetts.*

Fillmore, C. J. 1968. The case for case. In *EmmonW. Bach and Robert T. Harms, editors, Universals in Linguistic Theory. Holt, Rinehart &Winston,* New York, páginas 1–88.

Fillmore, C. J. 1971. Some problems for case grammar. *In R. J. O'Brien, editor, 22nd annual Round Table. Linguistics: developments of the sixties – viewpoints of the seventies, volume 24 of Monograph Series on Language and Linguistics. Georgetown University Press,Washington D.C., páginas 35–56.*

Fillmore, C. J. 1976. Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: conference on the Origin and Development of Language and Speech*, volume 280, páginas 20-32.

Gildea, D. y D. Jurafsky. 2002. Automatic labeling of semantic roles. *Computacional Linguistics*, 28(3):245-288.

Gildea, D. y M. Palmer. 2002. The Necessity of Parsing for Predicate Argument Recognition. In *Proceedings of ACL 2002*, Philadelphia, USA.

Giuglea, A. M. y A. Moschitti. 2004. Knowledge Discovering using FrameNet, VerbNet and PropBank. In *Proceedings of the Workshop on Ontology and Knowledge Discovery at ECML 2004,* Pisa, Italia.

Hacioglu, K., S. Pradhan, W. Ward, J. Martin, D. Jurafsky. 2003. Shallow Semantic parsing using support vector machines. Technical Report TR-CSLR-2003-1, Center for Spoken Language Reserach, Boulder, Colorado.

Hobbs, J. R., D. Appelt, J. Bear, D. Israel, M. Kameyama, M. Stickel, M. Tyson. 1997. ``FASTUS: A Cascaded Finite-State Transducer for Extraction Information from Natural Language Text". In *Emmanuel Roche and Yves Schabes, editors,* Finite-State Language Processing, capítulo 13, páginas 383-406. MIT Press, Cambridge, Massachusetts, Londres.

ISO 24613 Language resource management – Lexical markup framework. ISO Geneva 2005.

Kingsbury, P., M. Palmer, M. Marcus. 2002. Adding semantic annotation to the Penn Treebank. In *Proceedings of the Human Language Technology Conference*. San Diego. CA

Kipper, K., H. T. Dang, M. Palmer. 2000. Class-Based Construction of a Verb Lexicon. In *AAAI-2000*, Austin TX.

Levin, B. 1993. English Verb Classes and Alternations: A Preliminary Investigation. The University of Chicago Press.

Levin, B. y M. Rappaport. 1996. From lexical semantics tu argument realization manuscript.

Melli, G., YangWang, Y. Liu, M. M. Kashani, Z. Shi, B. Gu, A. Sarkar, F.Popowich. Description of SQUASH, the SFU Question Answering Summary Handler for *the DUC-2005 Summarization Task.*

Pradhan, S., K. Hacioglu, V. Krugler, W. Ward, J. H. Martin, D. Jurafsky. 2005. Support Vector Learning for Semantic Argument Clasification. *Machine Learning,* 60, 11-39.

Rumbaugh, J., I. Jacobson, y G. Booch. 2005. The Unified Modeling language reference manual, *2ª ed, Addison Wesley* 2005.

Shi, L. y R. Mihalcea. 2005. Putting Pieces Together: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing, In *Proceedings of the Sixth International Conference on Intelligent Text Processing and Computational Linguistics*, Méjico.

Valin, V. y D. Robert. 1993. A synopsis of role and reference grammar. In *Robert D. Van Valin, editor, Advances in Role and Reference Grammar.* John Benjamins Publishing Company, Amsterdam, páginas 1–166.