Towards the Automatic Merging of Lexical Resources: Automatic Mapping

Muntsa Padró
IULA
Universitat Pompeu Fabra
Barcelona, Spain
muntsa.padro@upf.edu

Núria Bel
IULA
Universitat Pompeu Fabra
Barcelona, Spain
nuria.bel@upf.edu

Silvia Necsulescu
IULA
Universitat Pompeu Fabra
Barcelona, Spain
silvia.necsulescu@upf.edu

Abstract

Lexical Resources are a critical component for Natural Language Processing applications. However, the high cost of comparing and merging different resources has been a bottleneck to have richer resources with a broad range of potential uses for a significant number of languages. With the objective of reducing cost by eliminating human intervention, we present a new method for automating the merging of resources, with special emphasis in what we call the mapping step. This mapping step, which converts the resources into a common format that allows latter the merging, is usually performed with huge manual effort and thus makes the whole process very costly. Thus, we propose a method to perform this mapping fully automatically. To test our method, we have addressed the merging of two verb subcategorization frame lexica for Spanish, The results achieved, that almost replicate human work, demonstrate the feasibility of the approach.

1 Introduction

The production, updating, tuning and maintenance of Language Resources for Natural Language Processing is currently being considered as one of the most promising areas of advances for the full deployment of Language Technologies. The reason is that these resources that describe, in one way or another, the characteristics of a particular language are necessary for Language Technologies to work.

Although the re-use of existing resources such as WordNet (Fellbaum, 1998) in different applications has been a well known and successful case, it is not very frequent. The different technology or application requirements, or even the ignorance about the existence of other resources, has provoked the proliferation of different, unrelated resources that, if merged, could constitute a richer repository of information augmenting the number of potential uses. This is especially important for under-resourced languages, which normally suffer from the lack of broad coverage resources. The research reported in this paper was done in the context of the creation of a gold-standard for subcategorization frames of Spanish verbs to be used in lexical acquisition (Korhonen, 2002). We wanted to merge two hand-written, large scale Spanish lexica to obtain a new one that is richer and validated. Because subcategorization frames contain highly structured information, it was considered a good scenario for testing new lexical resource merging methods.

Several attempts at resource merging have been addressed and reported in the literature. Teufel (1995) and Chan & Wu (1999) were concerned with the merging of several source lexica for PoS tagging. The merging of more complex lexica has been addressed by Crouch and King (2005) who produced a Unified Lexicon with lexical entries for verbs based on their syntactic subcategorization in combination with their meaning, as described by WordNet, Cyc (Lenat, 1995) and VerbNet (Kipper et al., 2000).

In this context, a proposal such as the Lexical Markup Framework, LMF (Francopoulo et al. 2008) is an attempt to standardize the format of

Proceedings of Recent Advances in Natural Language Processing, pages 296–301,
Hissar, Bulgaria, 12-14 September 2011.

computational lexica as a way to avoid the complexities of merging lexica with different structures. But there is no particular facility to easy the mapping from non-standard into standard.

Molinero et al (2009) build a morphological and syntactic lexicon for Spanish (*Leffe*) by merging four different lexica. They convert these sources into the *Alexina* format which is compatible with LMF in order to merge them. Nevertheless, both the mapping to this common format and the merging of the resources is done using manually developed rules that need a deep knowledge of the lexica to be merged.

The research presented here is closely related to Necsulescu et al (2011), that presents a method to automatically merge lexica using graph unification mechanism. To do so, the lexica need to be represented as feature structures. Again, the conversion of the lexica into the common format (in this case a graph structure) is performed developing a set of manual rules.

Despite the undeniable achievements of the research just mentioned, most of it reports the need for a significant amount of human intervention to extract information of existing resources and to represent it in a way that can be compared with another lexicon, or towards proposed standards, such as the mentioned LMF. Thus, there is still room for improvement in reducing human intervention. This constituted the main challenge of the research reported in this paper: finding a method that can perform blind, but semantic preserving operations to allow for automatically merging two lexical resources, in this particular case two subcategorization frame (SCF) lexica for Spanish, as we did in Necsulescu et al. (2011).

In next section we introduce the proposed method for automatic mapping and merging of information. Section 3 presents the obtained results, and in section 4 we state the conclusions and the future work.

2 Merging Lexica

Basically, merging of lexica has two well defined steps (Crouch and King, 2005). In the first, because information about the same phenomenon can be expressed differently, the existing resources have to be mapped into a common format, which makes merging possible in a second step. While automation of the second step has already proved to be possible, human intervention is still critically needed for the first. In addition to the cost of manual work, note that the exercise is completely adhoc for the particular resources to be merged. The cost is what explains the lack of interest in merging existing resources, even though it is critically needed, especially for under-resourced languages. Any cost reduction will have a high impact in the actual re-use of resources.

Thus, our objective was to reduce human intervention in the first step by devising a blind, semantic preserving mapping algorithm that covers the extraction of the information and the conversion into a format that allows, later, the merging.

In our experiments, we wanted to merge two subcategorization lexica developed for rule-based grammars: the Spanish working lexicon of the Incyta Machine Translation system (Alonso, 2005) and the Spanish working lexicon of the Spanish Resource Grammar, SRG, (Marimon, 2010) developed for the LKB framework (Copestake, 2002). Note that different senses under the same lemma are not distinguished in these lexica, and thus, are not addressed in the research reported here[1]. SRG and Incyta lexica encode the same phenomena related to verbal complements, their role and categorical characteristics expressed as restrictions. SCFs in the SRG lexicon are formulated in terms of feature-attribute value pairs, so they have a graph structure. In the Incyta lexicon, SCFs are represented as a list of parenthesis with less structured internal information[2]. In both cases, a lemma can have more than one SCF, and it is indeed the most frequent case as we will see later. For more-details about these two lexica, see Necsulescu et al. (2011).

In order to approach current proposals for standard formats (Francopoulo et al. 2008; Ide & Bunt, 2010) that recommend graph-based and attribute-value formalisms, we choose to map Incyta information towards SRG format which was closer to the standard recommendations. The devised method was to find semantically equivalent pieces of information and to substitute the parenthetical list by the attribute-value equivalent matrix.

[1] These characteristics made it not advisable to use LMF where lemma and sense are the mandatory information for a lexical entry.

[2] Decorated lists, parenthetical or otherwise marked, have been a quite common way of representing SCF information, i.e. COMLEX, VERBNET among others.

2.1 Semantic Preserving Mapping

Our experiment to avoid manual intervention when converting the two lexica into a common format with a blind, semantic preserving method departs from the idea of Chan and Wu (1999) to compare information contained in the same entries of different lexica, looking for significant equivalences. However they were working only with part-of-speech tags, while we handle complex, structured information. Note that we need to automatically learn correspondences for both, labels (such as the label of a noun phrase) and structures (e.g. the representation of a prepositional phrase that is fulfilled by a clause phrase in indicative mode).

The basic requirement for the automatic mapping is to have a number of verbs encoded in both lexica to be compared. Then it is possible to assess that a piece of the code in lexicon A corresponds to a piece of code in lexicon B since a significant number of other verbs hold the same correspondence. Thus, when a correspondence is found, the relevant piece in A will be substituted by the piece in B, performing the conversion into the target format.

Since we wanted our method to not be informed by human knowledge of the lexica, in order to make it applicable to more than one lexicon, the first point to solve was how to compare SCF code with no available previous information about their internal semantics. The code in Incyta lexicon is as in example (1).

(1) (($SUBJ N1 N0 (FCP 0 INT) (MD-0 IND) (MD-INT SUB)) ($DOBJ N1))

Therefore, the information that had to be discovered was the following:

- Incyta lexicon marks each SCF as a list of parenthesis, where the first level of parenthesis indicates the list of complements.
- Each component of the list begins with an identifier followed, without necessarily any formal marker, by additional information about properties of the component in the form of tags. For example, in (1) above, direct object ($DOBJ) is fulfilled by a noun phrase (N1).
- Incyta marks disjunction as a simple sequence of tags. In (1), subject ($SUBJ) may be fulfilled by N1 (noun phrase) or N0 (clause phrase). Furthermore, properties of one of the elements in the disjunction are specified in one or more parenthesis following the tag, as it is the case of N0 in (1). The 3 parenthesis after N0 are in fact properties of its realization: it is a sentential complement (FCP) whose verb should appear in indicative (MD-0 IND) unless it is an interrogative clause (MD-INT SUB).

We devised an algorithm that could discover this internal structure in Incyta SCFs. Our algorithm first splits every SCF in all possible ways according to formal characteristics (complete parenthetical components for Incyta and complete attribute-value matrices for SRG) and looks for the most frequently repeated pieces along the whole lexicon, so it is assessed that a particular piece is a meaningful unit. Note that we wanted to discover minimal units in order to handle different information encoding granularity. If we would have mapped entire SCFs or large pieces of them, the system could substitute information in A with information in B, possibly missing a difference.

Note that when performing the mapping for small pieces we ensure that we save as much the information as possible in the original lexicon, but this also causes the mapping result to not be a complete SCF. Since the ultimate goal is merging the two lexica, it is in the merging step that the partial elements will obtain the missing parts.

To sum up, our algorithm does the following with the Incyta SCF code:

1. Split SCF into each parentheses that conforms the list (this is to find $SUBJ and $DOBJ in 1).
2. For each of these pieces, it considers the first element as its key, and recursively splits the following elements.
3. It detects the relationship among the different elements found inside the parentheses by assessing which of them always occur together. For (1), it will detect that FCP appears only when there is a N0, and that MD-0 appears only when we have seen (FCP 0). In this way, we will obtain the constituents of the parentheses grouped according to their dependency.

Once extracted the different parts of each Incyta SCF and joined the elements that are correlated, our algorithm does the mapping:

1. For each element extracted from the Incyta SCF, it creates a list of verbs that contain it. This list is represented as a binary vector whose element *i* is 1 if the verb in position *i* is in the list.

2. It splits the SRG graphs following the feature-value attributes and builds a binary vector with the verbs that contain each element.
3. For each Incyta SCF minimal unit, it assesses the similarity with each SRG unit comparing the two binary vectors using the Jaccard distance measure, especially suited for binary vectors and as in (Chan and Wu, 1999).
4. It chooses as the corresponding elements those that maximize similarity.

Once the corresponding elements have been extracted, a new feature structure is constructed substituting Incyta units with those from SRG and the actual merging with the SRG lexicon is done. Since the SCFs have a graph structure, we used a unification mechanism (NLTK, Bird 2006) to merge both lexica, lemma by lemma, as in Necsulescu et al. (2011). Thus, we obtained, totally automatically, a new lexicon that contains SCF information from both lexica.

3 Evaluation and Results

To evaluate the results of our automatic mapping algorithm, we used the resulting lexicon of Necsulescu et al (2011) work as our gold-standard. To create this lexicon, Necsulescu et al (2011) developed a manually built set of extraction rules that converted Incyta list-based SCF's into SRG-like feature structures. Once both dictionaries were reliably converted into the same format, they were merged by using unification, thus obtaining a richer lexicon that we have used as the gold-standard for the automatic mapping exercise.

In order to evaluate the quality of the automatic mapping step, we compared the lexicon resulting from the merging of the SRG and the automatically mapped Incyta lexicon with the gold-standard. This comparison was first carried out by looking for identical SCFs in the entries of every particular verb. However, the results of the automatic mapping are in some cases parts of SCFs, because of the piece splitting process. As said, merging adds the lacking information in numerous cases, but the Incyta SCFs that do not unify with any SRG SCF remain incomplete. Also, there are cases in which the manually converted frame has more information than the automatic one, but the SCFs resulting from the automating mapping subsumes the one in the gold-standard, so they may be considered correct, although incomplete. Thus, in a second measure, we also count these pieces that are compatible with SCFs in the gold-standard as a positive result.

The evaluation is done using traditional precision, recall and F1 measures for each verb and then we compute the mean of these measures over all the verbs. The results, shown in table 1, are near 88% of F1 even in the strict case of identical SCFs. If we compare SCFs that unify, the results are even more satisfactory.

	P	**R**	**F1**
A-identical	87,35%	88,02%	87,69%
B-compatible	92,35%	93,08%	92,72%

Table 1: Average results of the mapping exercise

In Figure 1 we can see the performance in terms of number of SCFs under a lemma that are the same in the gold-standard and in the merged lexicon. We also plot the ratio of verbs that have a particular number of SCFs or less. The verbs that have one or two SCFs (about 50% of the verbs) obtain high values, as it may be expected. Nevertheless, 95% of verbs (those with 11 or less SCFs per lemma) obtain at least F1=80% when counting strictly equal SCFs and F1 over 90% when counting unifying SCFs. Note that these figures are the lower threshold, since verbs with less SCFs have better results, as it can be seen in Figure 1. To summarize, we consider that the obtained precision and recall of all verbs, even those with more than two SCFs, are very satisfactory.

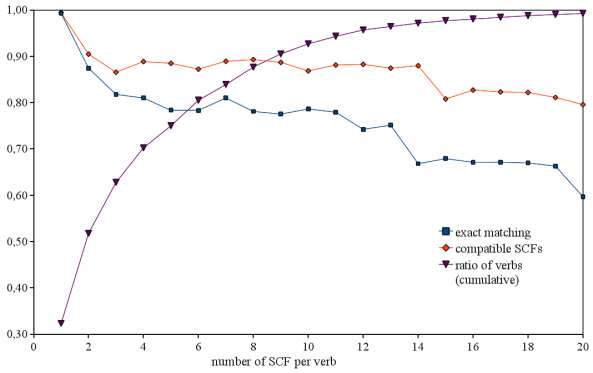

Figure 1: Average F1 and cumulative number of verbs with respect to the number of SCFs

As for the error analysis, the results revealed that some SCFs in the gold-standard are not in the automatically built lexicon. One case is SCFs with adverbial complements. Our algorithm maps adverbials onto PPs and the resulting SCF misses part

of the original information. Nevertheless, our algorithm correctly adds information when there are gaps in one of the dictionaries. It is able to learn correspondences such as "INT" (Incyta for interrogative clause) to "q" in SRG and to add this information when it is missed in a particular entry of the SRG lexicon but available in the Incyta entry.

4 Conclusions and Future Work

We have proposed a method to reduce human intervention in the merging of lexical resources. In order to unify different lexica, the resources need to be mapped into a comparable format. To reduce the cost of extracting and comparing the contents, we proposed a method to make the mapping automatically. We consider the results obtained very satisfactory. Our method rids the manual information extraction phase, which is the big bottleneck for the re-use and merging of language resources.

The strongest point of our method is that it can be applied without the need of knowing the structure nor the semantics of the lexica to be compared. This allows us to think our method can be extended to other types of Lexical Resources. The only requirement is that all resources to be merged contain some common data. Although further work is needed for assessing how much common data guarantees the same results, the current work is indicative of the feasibility of our approach.

It is important to note that the results presented here are obtained without using what Crouch and King (2005) call patch files. Automatic merging produces consistent errors that can be objects of further refinement. Thus, it is possible to devise specific patches that correct or add information in particular cases where either wrong or incomplete information is produced. It is future work to study the use of patch files to improve our method.

Acknowledgments

This project has been funded by the PANACEA project (EU-7FP-ITC-248064) and the CLARA project (EU-7FP-ITN-238405).

References

Juan Alberto Alonso, András Bocsák. 2005. Machine Translation for Catalan-Spanish. The Real Case for Productive MT; In Proceedings of the tenth Conference on European Association of Machine Translation (EAMT 2005), Budapest, Hungary.

Steven Bird. 2006. NLTK: the natural language toolkit. In Proceedings of the COLING/ACL on Interactive presentation sessions. Association for Computational Linguistics, Morristown, NJ, USA.

Ignacio Bosque and Violeta Demonte, Eds. (1999): Gramática descriptiva de la lengua española, R.A.E. - Espasa Calpe, Madrid.

Daniel K. Chan and Dekai Wu. 1999. Automatically Merging Lexicons that have Incompatible Part-of-Speech Categories. Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-99). Maryland.

Ann Copestake. 2002. Implementing Typed Feature Structure Grammars. CSLI Publications, CSLI lecture notes, number 110, Chicago.

Dick Crouch and Tracy H. King. 2005. Unifying lexical resources. Proceedings of Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes. Saarbruecken; Germany.

Christiane Fellbaum. 1998. WordNet: An Electronic Lexical Database. MIT Press.

Gil Francopoulo, Núria Bel, Monte George, Nicoletta Calzolari, Mandy Pet, and Claudia Soria. 2008. Multilingual resources for NLP in the lexical markup framework (LMF). Journal of Language Resources and Evaluation, 43 (1).

John Hughes, Clive Souter, and E. Atwell. 1995. Automatic Extraction of Tagset Mappings from Parallel-Annotated Corpora. Computation and Language.

Nancy Ide and Harry Bunt. 2010. Anatomy of Annotation Schemes: Mapping to GrAF. Proceedings of the Fourth Linguistic Annotation Workshop, ACL 2010

Daniel Jurafsky and James H. Martin. 2009. Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. 2nd edition. Prentice-Hall.

Karin Kipper, Hoa Trang Dang, and Martha Palmer. 2000. Class-based construction of a verb lexicon. In Proceedings of AAAI/IAAI.

Anna Korhonen. 2002. Subcategorization Acquisition. PhD thesis published as Technical Report UCAM-CL-TR-530. Computer Laboratory, University of Cambridge

Doug Lenat. 1995. Cyc: a large-scale investment in knowledge infrastructure. In CACM 38, n.11.

Christopher D. Manning and Hinrich Schütze. 1999. Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA, USA.

Montserrat Marimon. 2010. The Spanish Resource Grammar. Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10). Paris, France: European Language Resources Association (ELRA).

Miguel A. Molinero, Benoît Sagot and Nicolas Lionel. 2009. Building a morphological and syntactic lexicon by merging various linguistic resources. In Proceeding of 17th Nordic Conference on Computational Linguistics (NODALIDA-09), Odense, Danemar.

Monica Monachini, Nicoletta Calzolari, Khalid Choukri, Jochen Friedrich,Giulio Maltese, Michele Mammini, Jan Odijk & Marisa Ulivieri. 2006. Unified Lexicon and Unified Morphosyntactic Specifications for Written and Spoken Italian. In Calzolari et al. (eds.), LREC2006: 5th International Conference on Language Resources and Evaluation: Proceedings, pp. 1852-1857, Genoa, Italy.C.J.

Silvia Necsulescu, Núria Bel, Muntsa Padró, Montserrat Marimon and Eva Revilla: Towards the Automatic Merging of Language Resources. In Proceedings of WoLeR 2011. Ljubljana, Slovenia.

Pollard and I.A. Sag. 1994. Head-driven PhraseStructure Grammar. The University of Chicago Press, Chicago.

Simone Teufel. 1995. A Support Tool for Tagset Mapping. In EACL-Sigdat 95.