Semi-automatic Development of FrameNet for Italian

Sara Tonelli[1], Daniele Pighin[2], Claudio Giuliano[2], and Emanuele Pianta[2]

[1]Department of Language Science, Università di Venezia
[2]Fondazione Bruno Kessler, HLT Group, Povo (Trento)

1 Introduction

In this paper we introduce a joint project between Università di Venezia and Fondazione Bruno Kessler, Trento, for the semi-automatic development of FrameNet for Italian. The collaboration is aimed at investigating semi-automatic approaches to acquire FrameNet for new languages and at developing a paradigm that can be suitable for most European languages. The experiments we have been carried out so far go in three directions: 1) projection of frame information from English to Italian applying and comparing two rule-based algorithms 2) mapping between FrameNet and WordNet with a machine-learning approach 3) automatic assignment of sentences extracted from Wikipedia to FrameNet frames using a word-sense disambiguation system. In the following sections, we will briefly describe these three research directions and we will present the annotated resources that we have developed so far and that we plan to make available soon.

2 Frame information projection

Since other projects about the automatic transfer of frame information between languages have shown promising results (Padó and Lapata [9], Padó and Pitel [10]), we have decided to apply a similar approach to Italian. To this purpose, we developed and tested two projection algorithms that, given an English text annotated with frame information, and its Italian translation, project the annotated information from the source to the target text.

The first algorithm requires the target and the source text to be syntactically parsed and aligned at word level. Transferring the annotation of the frame-evoking

lexical unit is quite straightforward because it relies solely upon word alignment between an English lexical unit its Italian translation equivalent. On the contrary, frame element (FE) transfer is carried out at constituent level. Given an English constituent, annotated as FE, the algorithm extracts its semantic head, aligns it with the corresponding Italian head, then looks for the maximal syntactic projection of the Italian semantic head, and transfers the English FE annotation to such constituent. In this approach, the correct alignment of the head is enough to carry out FE transfer. However, this feature may also turn in a disadvantage, because if the semantic head is not aligned, there will be no transfer.

In order to cope with recall problems, we developed a second algorithm, where alignment between constituents for FE transfer is based on the best percentage of aligned words. In short, for every English constituent bearing a FE label, we align it to the Italian constituent that shares the highest number of aligned words (for more details, see Tonelli and Pianta [14]).

The two algorithms were tuned and tested on two different parallel corpora. The first one was an excerpt of 987 English and Italian sentences taken from the *Europarl* multilanguage parallel corpus [7]. The English side was manually enriched with frame-semantic information as described in Padó and Lapata [9] in the context of transfer experiments between English and German. The Italian corresponding sentences were also manually annotated with frame information in order to build a gold standard for the present experiments. This corpus was characterized by a high number of free translations and a limited set of frames, mostly related to the communication and the political scenarios.

The second corpus was built by manually translating in a controlled way 400 sentences from the Berkeley FrameNet corpus. The sentences were selected in order to maximize frame variability, with one different frame per sentence, and to reduce syntactic complexity. While the English side was already annotated in the framework of the Berkeley FrameNet project, we manually annotated the Italian side in order to build a second gold standard.

The evaluation of the two algorithms using two different gold standards highlighted a better performance of algorithm 2, but also proved that the features of the corpus on which the gold standard is based have a high impact on algorithm performance. For example, algorithm 2 scored an enhancement of 0.9 both in precision and in recall (0.66 vs. 0.75 and 0.40 vs. 0.49 respectively) if evaluated on the second gold standard.

In general, we noticed that the transfer approach can be applied to automatically annotate a corpus with frame information only if the parallel sentences don't present many free translations. Another important point is that the transfer performance improves when the syntactic complexity and the sentence length decrease. Besides, it is not straightforward to compare the transfer results to previous exper-

iments because the evaluation metrics used in the past are very different (see for example Basili et al. [1] and Padó [8]) and it may be worth to define a common evaluation framework.

3 WordNet – FrameNet mapping

A second research direction we have been investigating is the automatic extension and population of Italian frames exploiting existing resources. In particular, we propose to link English lexical units with WordNet synsets and then use *MultiWordNet*[1] [11] as a bridge to populate frames with lemmas from the corresponding Italian synsets. For example, if we consider the *rouse.v* lexical unit belonging to the CAUSE_TO_WAKE frame and we extract all WordNet synsets containing *rouse.v*, we should be able to assess that the synset with {*awaken, wake, waken, rouse, wake up, arouse*} best expresses the meaning of *rouse.v* in CAUSE_TO_WAKE and to discard {*bestir, rouse*}, {*rout out, drive out, force out, rouse*} and {*agitate, rouse, turn on, charge, commove, excite, charge up*}. Then, we could retrieve from *MultiWordNet* the Italian synset containing {*destare, svegliare*}, which is internally linked to {*awaken, wake, waken, rouse, wake up, arouse*}. In this way, we could automatically populate CAUSE_TO_WAKE with two Italian lexical units.

In order to carry out the mapping, we first developed a dataset by manually annotating 2,158 lexical unit - synset pairs as positive or negative examples. Then, we trained a binary classifier with the SVM optimizer SVM-Light [5] and polynomial kernels of different degrees. Although the mapping task is not new (see Johansson and Nugues [6] and Shi and Mihalcea [12]), we extracted a novel set of features that can cope with coverage problems of past mapping experiments. In particular, we exploited a stem overlap measure between WordNet glosses and LU definitions in FrameNet and we also took into account information about WordNet domains of the candidate synsets (for details about the features, see Tonelli and Pighin [15]).

We produced in this way a mapping between FrameNet frames and WordNet synsets, which we called *MapNet*, having 0.79 P, 0.57 R and 0.66 F1. Using *MultiWordNet* as a bridge, we could automatically acquire 6,429 Italian lexical units for 561 frames (precision evaluated on 15 complete frames: 0.88). Then, we further decided to exploit *MultiSemCor* [2], a parallel corpus of English and Italian sentences with synset annotation to acquire example sentences for the Italian FrameNet database. So, we labeled them with frame labels according to our mapping. This allowed us to acquire 23,872 Italian sentences.

[1] In *MultiWordNet*, every synset contains lemmas in different languages, included English and Italian

4 Sentence extraction from Wikipedia

The third research direction we have been investigating is the automatic acquisition of new example sentences and lexical units exploiting the huge amount of data available through Wikipedia. In particular, for every lexical unit in the English FrameNet, we apply a word sense disambiguation system [4] that, for a given pair frame - lexical unit *(F; l)*, identifies the Wikipage that best expresses the meaning of *l*. Then, we retrieve the Italian version of the linked Wikipage, if available, and extract all sentences in the Italian Wikipedia that contain a reference to that Wikipage.

The WSD system was trained using for every lexical unit *l* all sentences from Wikipedia where *l* is the anchor of an internal link. The set of pages anchored by *l* represents the senses of *l* in Wikipedia and the contexts, i.e. sentences where *l* appears, are used as labelled training examples. For example, the lexical unit *building.n* in the frame BUILDINGS is an anchor in 708 different sentences that point to 42 different Wikipedia pages.

After the training, the system can map a *(F; l)* pair with the Wikipedia page that best expresses the meaning of *l*. So, we retrieve the Italian version of that Wikipage and extract all Italian sentences pointing to it. For example, if we link `http://en.wikipedia.org/wiki/Court` to the JUDICIAL_BODY frame, we first retrieve the Italian version of the site `http://it.wikipedia.org/wiki/Tribunale`. Then, with a top-down strategy, we further extract all Italian sentences pointing to the *Tribunale* page and acquire as lexical units all words with an embedded reference to this concept, for example *tribunale* and *corte*. In this way, we can populate the JUDICIAL_BODY frame with the extracted lexical units and the retrieved sentences containing them.

For the moment, we have carried out our experiments starting from nominal lexical units in the English FrameNet that have no example sentences in the database. At the end of the mapping, we were able to extract 23,078 sentences from the Italian Wikipedia and assign them to 371 different frames. A preliminary evaluation on 1,000 randomly chosen sentences scored 0.69 accuracy.

5 The Italian FrameNet data so far

The resource we have been developing comprises some implemented algorithms and annotated text. The algorithms / systems are:

- Two transfer algorithms for cross-lingual projection of frame information
- One WordNet – FrameNet mapping system (can be exported for every language available in MultiWordNet)

- One sentence extraction system from Wikipedia (can be exported for every language available in Wikipedia)

The annotated data comprise:

- *Europarl* gold standard with 1,000 parallel sentences in English and Italian, parsed, aligned at word level, manually annotated with frame information. It has already been used as gold standard for automatic annotation experiments of Italian (see Basili et al. [1]).
- 400 sentences in English extracted from the Berkeley FrameNet database and translated into Italian, parsed (only Italian side), aligned at word level, manually annotated with frame information. This and the previous corpus together contain at least one lexical unit and one example sentence for every frame in the English FrameNet.
- 2,158 LU-synset pairs manually annotated as positive or negative examples; 5,162 LU-synset pairs automatically annotated and available for download at `http://danielepighin.net/cms/research/MapNet`.
- 23,872 Italian sentences from the *MultiSemCor* corpus, with PoS, lemma and synset information, automatically enriched with frame labels pointing to the synsets (`http://multisemcor.itc.it/`)
- 23,078 sentences from Italian Wikipedia with frame label (371 different frames). The dataset is easily extendible to all languages of Wikipedia. Also the number of annotated sentences can be largely and easily increased applying the Word Sense Disambiguation system described in 4 to all lexical units in the English FrameNet.

6 Conclusions and future work

In this work, we have presented three research directions in the framework of the semi-automatic development of Italian FrameNet. Besides, we have described the annotated data collected so far. The project is still ongoing and we plan to create a public website and make available at least the annotated parallel corpora as soon as possible. In the next step, we will also develop and test some strategies to semi-automatically validate the frame assignment for MultiSemCor and the Wikipedia sentences. Since our aim is to release a resource that is also in line with the FrameNet database standard, we plan to investigate some approaches to automatically acquire grammatical functions for the annotated parallel texts, and to convert our gold standards in the *Salto* format (see Burchardt et al. [3]) to the FrameNet Desktop standard.

References

[1] R. Basili, D. De Cao, D. Croce, B. Coppola, and A. Moschitti (2009) Cross-language frame semantics transfer in bilingual corpora. In *Proceedings of CICLing*. Springer-Verlag.

[2] L. Bentivogli and E. Pianta (2005) Exploiting Parallel Texts in the Creation of Multilingual Semantically Annotated Resources: The MultiSemCor Corpus. *Natural Language Engineering, Special Issue on Parallel Texts*, 11(03):247–261, September.

[3] A. Burchardt, K. Erk, A. Frank, A. Kowalski, S. Padó, and M. Pinkal (2006) Salto - a versatile multi-level annotation tool. In *Proceedings of LREC 2006*, pages 517–520, Genoa Italy.

[4] A. Gliozzo, C. Giuliano, and C. Strapparava (2005) Domain kernels for word sense disambiguation. In *Proceedings of the 43rd annual meeting of the Association for Computational Linguistics (ACL-05)*, pages 403–410, Ann Arbor, Michigan, June.

[5] T. Joachims (1999) Making large-scale support vector machine learning practical. In Bernhard Schölkopf, Christopher J. C. Burges, and Alexander J. Smola, editors, *Advances in kernel methods: support vector learning*, pages 169–184. MIT Press, Cambridge, MA, USA.

[6] R. Johansson and P. Nugues (2007) Using WordNet to extend FrameNet coverage. In Proc. of the Workshop on Building Frame-semantic Resources for Scandinavian and Baltic Languages, Tartu.

[7] P. Koehn (2005) Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of MT Summit*.

[8] S. Padó (2007) *Cross-Lingual Annotation Projection Models for Role-Semantic Information*. PhD thesis, Universität des Saarlandes.

[9] S. Padó and M. Lapata (2005) Cross-linguistic Projection of Role- Semantic Information. In *Proceedings of Human Language Technology Conference and EMNLP*, pages 859–866, Vancouver, Canada.

[10] S. Padó and G. Pitel (2007) Annotation précise du fraçais en sémantique de roles par projection cross-linguistique. In *Proceedings of TALN-07*, Toulouse, France.

[11] E. Pianta, L. Bentivogli, and C. Girardi (2002) MultiWordNet: developing an aligned multilingual database. In *First International Conference on Global WordNet*, pages 292–302, Mysore, India.

[12] L. Shi and R. Mihalcea (2005) Putting Pieces Together: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing. In *Proceedings of CICLing-05*, pages 100–111. Springer.

[13] S. Tonelli and C. Giuliano (2009) Wikipedia as frame information repository. In *Proceedings of Empirical Methods in Natural Language Processing*, Singapore, Malaysia.

[14] S. Tonelli and E. Pianta (2009) Three issues in cross-language frame information transfer. To appear in *Proceedings of Recent Advances in Natural Language Processing (RANLP-09)*, Borovets, Bulgaria.

[15] S. Tonelli and D. Pighin (2009) New features for FrameNet – WordNet Mapping. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, Boulder, CO, USA.