Frame information transfer from English to Italian

Sara Tonelli, Emanuele Pianta

FBK-irst, via Sommarive 18, 38050 Povo (TN), Italy
{satonelli, pianta}@fbk.eu

Abstract

We describe an automatic projection algorithm for transferring frame-semantic information from English to Italian texts, as a first sep towards the creation of Italian FrameNet. Projection of frame semantic information from English to other European languages has already been investigated for German, Swedish and French. With our work, we point out typical features of the Italian language as regards frame-semantic annotation, in particular we describe peculiarities of Italian that at the moment make the projection task more difficult than in the above-mentioned examples. Besides, we created a gold standard with 987 manually annotated sentences to evaluate the algorithm.

1. Towards Italian FrameNet

1.1. Introduction

FrameNet (Baker et al., 1998) is a lexical resource for English based on frame semantics and supported by corpus evidence, whose aim is to collect the range of semantic and syntactic combinatory possibilities of each word in each of its senses through annotation of example sentences. The conceptual model is based on three main elements:

Semantic frame: the conceptual structure that describes a particular type of situation, object or event and the participants involved in it

Lexical unit (LU): a word, a multiword or an idiomatic expression that evokes a frame

Frame element (FE): the semantic roles expressed by the syntactic dependents of the LU

The ongoing FrameNet project for English relies on manual annotation and contains 825 frames covering 6,100 fully annotated lexical units. Although this method provides a systematic and accurate approach to the frame annotation task, it is quite expensive and time-consuming and requires a large group of trained annotators. In fact, it took approximately eight years to develop a resource with partial coverage of the English language.

In order to create corpora with frame-semantic information for new languages, various approaches have been proposed to make the process run automatically. (Padó and Lapata, 2005; Padó and Pitel, 2007) and (Johansson and Nugues, 2006) describe an annotation transfer method that can be applied to parallel texts where the source corpus has been automatically annotated with a semantic role labeller trained on English FrameNet. (Johansson and Nugues, 2006) showed also that this kind of projection on large aligned corpora can be a preliminary step for developing a semantic role labeller for the target language.

Following the automatic projection approach, we plan to build the Italian FrameNet resource mostly relying on automatic procedures that can help reduce human effort. This pilot study aims at investigating semantic parallelism between English and Italian and at developing an algorithm for cross-lingual projection of frame information from annotated English texts to Italian translations.

In order to evaluate frame information projection from English to German, (Padó and Lapata, 2005) created a 987-sentence gold standard based on bitexts extracted from the Europarl corpus (Koehn, 2005). The English and the German side of the gold standard have been automatically annotated with part of speech and syntactic information and manually enriched with frame-semantic information. Since the Europarl corpus contains also the Italian translation of the texts used in the English-German gold standard, we decided to build an extended gold standard by manually annotating the Italian translations with frame information and to use it to test our frame information projection algorithm for the English-Italian language pair.

1.2. The projection algorithm

In the current state, the English-Italian projection algorithm requires that English text be parsed with the Collins' parser and that frame annotations make reference to syntactic constituents. The algorithm is based on four steps:

1 Automatic syntactic analysis of the Italian text

2 Automatic English-Italian alignment at word level

3 Automatic semantic head extraction for every annotated constituent in the English corpus side

4 Automatic projection of annotations from English to Italian constituents using aligned semantic heads as bridge

1.2.1. Italian corpus preparation

Italian texts are first parsed with Bikel's phrase-based statistical parser trained for Italian (Corazza et al., 2007)[1]. After that, the Italian sentences can be converted into XML-Tiger format and visualized as syntax trees with SALTO (Burchardt et al., 2006). The same tool was used in the English-German project to manually add frame-semantic information, which means that the output corpora will be fully compatible and have the same XML-structure.

[1]The parser developed by Corazza et al. obtained the best score in the EVALITA evaluation campaign for Italian NLP tools with 67.97 f-measure.

1.2.2. Alignment

The English-Italian corpus is aligned at word level with KNOWA (KNowledge-intensive Word Aligner) (Pianta and Bentivogli, 2004), a word aligner relying mostly on information contained in the Collins' bilingual dictionary, but also on a morphological analyzer and a multiword-recognizer. We chose KNOWA because with this language pair it outperforms GIZA++, in particular w.r.t. alignment of content words (85.5 precision vs. 53.2 of GIZA++ in the EuroCor task, which was carried out on a subset of English and Italian texts from Europarl as reported in (Pianta and Bentivogli, 2004)). This is important because the algorithm we propose relies on information projection between semantic heads, which are mostly content words.

1.2.3. Semantic head extraction

The best model for English-German projection is based on alignment at constituent level obtained through *word overlap similarity*, as described in (Padó and Lapata, 2005). We experimented a simpler strategy for constituent alignment which is based on *semantic heads* (see next section).
Annotations in the English side refer to syntactic constituents such as NP, VP, PP etc., which are maximal projection of a given lexical category. Any such constituent has only one semantic head, and we expect that its Italian translation be the semantic head of the Italian phrase corresponding to the English annotated constituent.
Since the English corpus is PoS-tagged and parsed with Collins' parser, we adapted his algorithm for syntactic head extraction to semantic head extraction. In case of discrepancies between syntactic and semantic heads, we give priority to semantic heads. For this purpose, we had to modifiy the priority list in the original head table. Besides, we added rules for subjectless sentences (SG) and basal NP nodes (NPB and NX), which were missing in the original head table.

1.2.4. Cross-lingual transfer

Frame information is conveyed by two different components: the frame itself is evoked by a lexical unit (*target*), whereas *frame elements* are usually expressed by more complex constituents. For this reason, the transfer of frame targets involves only a lexical unit, usually a verb, on both sides of the corpus, whereas a different transfer strategy is required for frame elements. After extracting the semantic head of the English constituent bearing frame element information, we get the Italian aligned semantic head, when available. Then, we find the highest syntactic projection of the Italian head compatible with the annotated English constituent. Finally we transfer annotation from the English maximal projection to the Italian constituent. We define a table of compatibility between English and Italian constituents, assessing for instance that NPs can correspond to either NPs or PPs.
Figure 1 shows an example of frame information transfer. The target element *beaten* was correctly aligned with *colpiti*, which is the literal translation of the source word. For this reason, frame annotation can be directly transferred from the English to the Italian lexical unit. As for the VICTIM frame element, in the first step we identify *children* as the semantic head of *women and children* (the head of coordinated structures is assumed to be the rightmost element in coordination). After matching *children* to *bambini*, we find the highest syntactic projection of the head compatible with the annotated English constituent, i.e. the uppermost NP. This strategy requires that only the head of the constituent be correctly aligned.

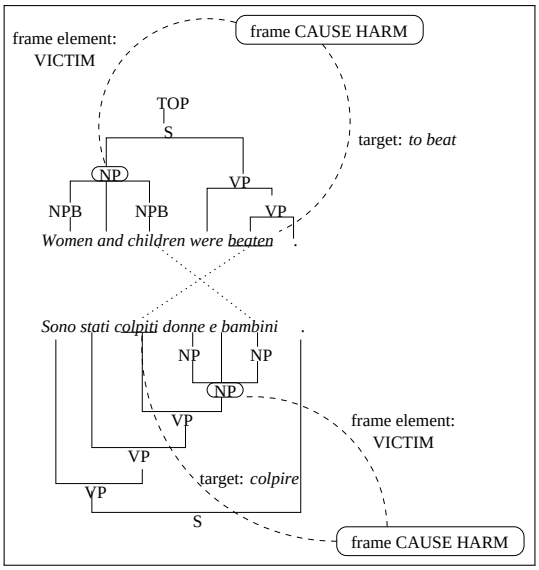

Figure 1: Example of frame information transfer

2. Evaluation

2.1. Gold standard

In order to evaluate our approach, we manually created a gold standard with the same 987 sentences used to build the English-German and the English-French gold standards. In this way, we contributed to building a parallel corpus of English, German, French and Italian sentences annotated with frame information. First, we extracted from the English-Italian Europarl corpus the 987 Italian sentences which represent the translation of English ones present in the other gold standards, then we parsed them with Bikel's phrase-based statistical parser. We manually corrected the resulting syntactic trees and converted them into XML Tiger format using the TigerRegistry conversion tool (Lezius, 2002). Finally, we annotated them with frame information using SALTO (Burchardt et al., 2006). Annotation was carried out on the basis of the online FrameNet version, which is the most up to date. Besides, we defined a new frame, HANDLING.

As a preliminary step to the evaluation of the frame projection algorithm, we analyzed frame parallelism and role parallelism between English and Italian gold standards, which we assume to be a prerequisite for accurate projection. Results are reported in Table 1 in comparison with the English-German and English-French gold standards described in (Padó and Pitel, 2007).

The difference between *frame parallelism* for English-

Language Pair	Frame parallelism	FE parallelism
Eng-Ita	0.61	0.82
Eng-Ger	0.71	0.91
Eng-Fr	0.69	0.88

Table 1: Comparison of frame and FE parallelism

German and English-Italian may depend on the fact that English and German are more closely related language pairs. Besides, annotation of the Italian gold standard was carried out consulting the latest FrameNet version with around 880 annotated frames, while the gold standards for other languages were all created with FrameNet 1.1, with around 520 frames. In fact, the Italian gold standard shows a higher frame variability with 158 frames, while the English gold standard contains only 83 frames, the German one has 73 frames and the French 121 frames. Furthermore, 28 frame instances were assigned to the new HANDLING frame.

According to the methodology introduced by (Padó and Pitel, 2007), two sentences are counted as having parallel FEs if they contain the same target and the same frame element regardless of the role span. We adopted this approach and calculated *FE parallelism* only for sentences that have parallel frame annotation and regardless of the role span, as reported in Table 1. This means that, even when the targets correspond, there is a 18% of frame elements that don't match. This value is higher than for the other language pairs mainly because of English frame elements which are missing in the Italian gold standard. Most of them correspond to null-subject pronouns, since the subject of a sentence in Italian can be left unexpressed. So, every time we find a role-bearing subject pronoun such as *I*, *you*, *they*, *we*, *he* or *she* in the English corpus, we can expect that no corresponding overt lexical item is found in the Italian translation, as shown in the example below:

Ex. JUDGMENT DIRECT ADDRESS frame

[I]$_{Speaker}$ *thank [you]*$_{Addressee}$ *[for your report]*$_{Reason.}$
$\emptyset$$_{Speaker}$ *[La]*$_{Addressee}$ *ringrazio [per la relazione]*$_{Reason.}$

The same can be observed for expletive *it*, that is never expressed in Italian. Furthermore, in the Europarl corpus a large proportion of text is composed by speeches in first person, which increases the number of subject personal pronouns. In general, we observed that about 15% of all English FEs correspond to an null-subject pronoun in the Italian gold standard.

Other factors that negatively affect frame element parallelism are free translations and different interpretations of the sentences given by English and Italian annotators. Annotators divergences involve in particular frame elements which are semantically similar, such as *Topic*/*Message* in the STATEMENT frame, *Agent*/*Cause* in the CAUSE_HARM frame or *Area*/*Path* in the MOTION frame.

A third cause of missing parallelism is the different version of FrameNet used in the English and the Italian annotation. In version 1.1, for example, the SCRUTINY frame had the *Standard* frame element, which was called *Enabled_situation* in version 1.3. The same happened to the LIKELIHOOD frame, where the *Event* frame element in version 1.1 was changed into *Hypothetical_event*.

The degree of parallelism between frames and frame elements in the English and Italian gold standards (0.61 for frame transfer, 0.82 for FE transfer) represents an upper-bound for recall in automatic projection experiments. This relatively low value shows that structural differences between source and target language and translation shifts strongly affect the frame projection task. We expect recall to be higher in case of a parallel corpus where the English sentences are literally translated into Italian. As for syntactic similarity, we think it could be improved only taking another language (e.g. a Romance language) as projection source for Italian.

2.2. Frame projection evaluation

In order to evaluate both frame and FEs projection, we divided the corpus into a development set (300 sentences) and a testset (687 sentences). The development set has been used to tune the projection algorithm, while the testset has been used to evaluate the quality of the frame annotation resulting from the automatic projection of frame information from English to Italian.

The *coverage* of the *word alignment* process is in line with KNOWA performance on the EuroCor alignment task, and amounts to 65.1% on the whole corpus, 48.6% for content words and 64.0 for words listed in WordNet. If we only consider frame targets, 70% of the lexical units have been aligned.

	Precision	Recall	F-measure
Eng-Ita	0.71	0.50	0.59

Table 2: Frame projection evaluation

Wrong transfers depend mainly upon misalignments, structural differences between aligned sentences and translation shifts. Missing transfers depend upon missing translation equivalents in KNOWA dictionaries, for example *to breach* → *infrangere*. In some cases, they can also depend on free translations (ex. *legislation* → *proposta legislativa [proposal for a new law]*).

2.3. Frame Element projection evaluation

We carried out two different evaluations of FE projection. The first one is based on FE projection between aligned sentences with matching frames and considers a projection correct if the FE-bearing constituent span in the Italian output matches exactly the corresponding constituent with the same FE in the Italian gold standard. The second evaluation considers all frame elements in the testset, regardless if the source and the target sentences have a matching frame. We consider a FE projection correct if the same frame element is present both in the Italian gold standard and in the automatically annotated sentence, regardless of matching FE spans, and if the FE-bearing constituents in the two sentences have at least the same semantic head.

The first evaluation focuses on the performance of our projection algorithm. The second one aims at investigating to what extent our approach can be used to annotate a corpus as basis for building the Italian FrameNet.

2.3.1. Span exact match evaluation

In this evaluation we took into account only the frame elements in the bitext-sentences with matching frames, counting the exact span matches of the automatically annotated Italian frame elements against role annotation in the gold standard. This means that, for this kind of evaluation, the gold standard is reduced to 61% of the sentences in the testset, namely those who share the same frame in Italian and in English. We computed role projection precision and recall on the Italian sentences parsed with Bikel's parser as mentioned in Section 1.2.1 and on a corrected version of the parse trees, with manually revised constituent spans and nodes. Evaluation results are reported in the table below:

Input type	Precision	Recall	F-measure
Bikel's trees	0.48	0.39	0.43
Corrected trees	0.62	0.51	0.56

Table 3: Frame projection evaluation

The evaluation shows to what extent the approach could be improved if the parsers available for Italian performed better. In general terms, exact matches are mostly correct if they involve annotation projection between constituents that are the same in English and in Italian, while they tend to fail if they imply annotation projection between different constituents, for example a VP and a PP.

2.3.2. Head match evaluation

In this evaluation, we considered all frame elements in the automatically annotated Italian corpus against the Italian gold standard. Evaluation has been carried out by adopting three different criteria for assessing the match between automatic annotation and gold standard. In all three criteria, the automatically annotated FE matches the gold standard FE if they share at least the same semantic head. However, criterium 1 is more strict in that it requires that also the annotation of the corresponding targets match. Criterium 2 is somewhat looser in that it accepts matching FEs if the automatic annotation of the target word is right or *missing*. Criterium 3 considers correct all matching frame elements between automatic and manually annotated sentences regardless of whether the target has been annotated with the right frame. Results are reported in Table 4.

	Precision	Recall	F-measure
Criterium 1	0.46	0.30	0.37
Criterium 2	0.57	0.37	0.45
Criterium 3	0.64	0.41	0.49

Table 4: FE projection evaluation

All approaches show a low recall, which is affected by the factors already mentioned in section 2.2 for frame transfer. In few cases, discrepancies between Italian and English frame elements depend on different interpretations given by the annotators to the aligned sentences. For example, *[Two million children] have been killed [in armed conflict]* is the literal translation of *[2.000.000 di bambini] sono rimasti uccisi [in conflitti armati]*. Despite this, the annotator of the English gold standard labeled *[in armed conflict]* as *Cause*, while in the Italian gold standard *[in conflitti armati]* bears the *Circumstances* role.
Type 3 shows that a number of sentences have common frame elements even if they don't share the same frame. This feature is particularly evident in Europarl, where most targets correspond to verbs of statement and of opinion. In fact, the four most frequent frames in the Italian gold standard (AWARENESS, OPINION, STATEMENT and QUESTIONING) have relevant frame elements in common. For instance, STATEMENT and QUESTIONING share the core frame elements *Speaker*, *Message* and *Topic*. In general, we believe that Evaluation type 2 can be seen as the most significant type in a realistic frame information projection task.

3. Conclusions and future work

In this paper, we argue that English-Italian projection of frame-semantic annotation can be a means of developing Italian FrameNet with reduced human effort. We observe sufficient semantic parallelism between English and Italian to map frame assignments, even if at present the projection task seems more suitable to speed up semiautomatic annotation than to convey fully automatic frame information transfer. In particular, we noticed that in the Italian corpus there are instances where the fundamental assumption of our projection approach, namely that word alignment can be interpreted as frame-semantic equivalence, fails. The same had been observed for French as well in (Padó, 2007). See for example the following instance:

RELIANCE frame
EnGold: *[We]*$_{Protagonist}$ *rely on [you]*$_{Intermediary}$ *[to help us realise that aim]*$_{Benefit.}$
ItaProjection: *Facciamo affidamento [sul vostro aiuto]*$_{Benefit}$ *per conseguire quell'obiettivo.*
ItaGold: *Facciamo affidamento [sul vostro aiuto]*$_{Intermediary}$ *[per conseguire quell'obiettivo]*$_{Benefit.}$
Lit. transl.: *We rely [on your help] [to realise that aim].*

The correct alignment *help - aiuto* leads to the transfer of the *Benefit* role to *[sul vostro aiuto]*, while the latter should bear the *Intermediary* role. This affects recall as well, because the *Intermediary* role remains unassigned.
In general, we observed that syntactic similarity between source and target language can improve the projection results. For this reason, we believe that frame projection between Romance languages may be worth investigating.

4. Aknowledgements

We would like to thank Sebastian Padó for the support and the information about English-German and English-French frame information transfer.

5. References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In Christian Boitet and Pete Whitelock, editors, *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics*, pages 86–90. Morgan Kaufmann Publishers.

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Pado, and Manfred Pinkal. 2006. Salto - a versatile multi-level annotation tool. In *Proceedings of LREC 2006*, pages 517–520, Genoa Italy.

Anna Corazza, Alberto Lavelli, and Giorgio Satta. 2007. Analisi sintattica-statistica basata su costituenti. *Intelligenza Artificiale*, (2):38–39.

Richard Johansson and Pierre Nugues. 2006. A framenet-based semantic role labeler for swedish. In *Proceedings of Coling/ACL 2006*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit*.

Wolfgang Lezius. 2002. TIGERSearch – Ein Suchwerkzeug für Baumbanken. In Stephan Busemann, editor, *Proceedings der 6. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 2002)*, pages 107–114, Saarbrücken.

Sebastian Padó and Mirella Lapata. 2005. Cross-lingual bootstrapping of semantic lexicons: The case of framenet. In *Proceedings of Proceedings of AAAI*.

Sebastian Padó and Guillaume Pitel. 2007. Annotation précise du français en sémantique de roles par projection cross-linguistique. In *Proceedings of TALN-07*, Toulouse, France.

Sebastian Padó. 2007. *Cross-Lingual Annotation Projection Models for Role-Semantic Information*. Ph.D. thesis, Universität des Saarlandes.

Emanuele Pianta and Luisa Bentivogli. 2004. KNOwledge intensive word alignment with knowa. In *Proceedings of Coling 2004*, pages 1086 – 1092.