Cross-Language Frame Semantics Transfer in Bilingual Corpora

Roberto Basili[1], Diego De Cao[1], Danilo Croce[1], Bonaventura Coppola[2], and Alessandro Moschitti[2]

[1] Dept. of Computer Science,
University of Roma Tor Vergata, Roma, Italy
{basili,croce,decao}@info.uniroma2.it
[2] University of Trento, Italy
{coppola, moschitti}@disi.unitn.it

Abstract. Recent work on the transfer of semantic information across languages has been recently applied to the development of resources annotated with Frame information for different non-English European languages. These works are based on the assumption that parallel corpora annotated for English can be used to transfer the semantic information to the other target languages. In this paper, a robust method based on a statistical machine translation step augmented with simple rule-based post-processing is presented. It alleviates problems related to preprocessing errors and the complex optimization required by syntax-dependent models of the cross-lingual mapping. Different alignment strategies are here investigated against the Europarl corpus. Results suggest that the quality of the derived annotations is surprisingly good and well suited for training semantic role labeling systems.

1 Motivation

The availability of large scale semantic lexicons, such as Framenet ([1]), has allowed the adoption of a vaste family of learning paradigms in the automation of semantic parsing. Building on the so called *frame* semantic model, the Berkeley FrameNet project [1] has developed a frame-semantic lexicon for the core vocabulary of English since 1997. As defined in [2], a frame is a conceptual structure modeling a prototypical situation. A frame is evoked in texts through the occurrence of its lexical units (LU), i.e. predicate words (verbs, nouns, or adjectives) that linguistically expresses the situation of the frame. Each frame also specifies the participants and properties of the situation it describes, the so called frame elements (FEs), that are the Frame Semantics instantiation of semantic roles. For example the frame CATEGORIZATION has lexical units such as: *categorize,classify,classification,regard*. Semantic roles shared by these predicates, are the COGNIZER (i.e. the person who performs the categorization act), the ITEM construed or treated, the CATEGORY (i.e. the class which the item is considered a member of) and CRITERIA. Semantic Role Labeling (SRL) is the task of automatic labeling individual predicates togheter with their major roles (i.e. frame elements) as they are grammatically realized in input sentences. It has been a popular task since the availability of the PropBank and Framenet annotated corpora [3], the seminal work of [4] and

A. Gelbukh (Ed.): CICLing 2009, LNCS 5449, pp. 332–345, 2009.

the successful CoNLL evaluation campaigns [5]. Statistical machine learning methods, ranging from joint probabilistic models to support vector machines, have been largely adopted to provide accurate labeling, although inherently dependent on the availability of large scale annotated resources.

It has been observed that the so called resulting *resource scarcity problem* affects a large number of languages for which such annotated corpora are not available [6]. Recent works thus explored the possibility of the cross-linguistic transfer of semantic information over bilingual corpora in the development of resources annotated with frame information for different European languages ([7,6,8]). As SRL on English texts can rely on extensive resources, the English portion of a bilingual corpus can be labelled with a significant accuracy: the cross-language transfer of predicate and role information is an appealing process aiming to produce large scale information in a relatively cheap way. The approach discussed by Sebastian Pado focused on methods for the cross-lingual induction of frame semantic information aiming at creating frame and role annotations for new languages. Based on Framenet, as a source of semantic information, it has been influential on later attempts, as for example in [7,8]. The main aspects of this work are the neat separation between alignment at the level of predicates (usually single words) and the level of roles. The first problem is tackled in [6] by relying on distributional models of lexical association that allow to estimate when a given lexical unit is in fact expressing a predicate (frame). This supported a light approach to the predicate alignment task with significant accuracy. The second problem is approached through the syntactic alignment of constituents that are *role bearing phrases*, i.e. that express sentential roles of the target predicates. These methods allow to rely on the linguistic information encoded in the syntactic bracketing and alleviate word alignment errors. Results are characterized by higher-precision projections even over noisy input data, typically produced by shallow parsing techniques (e.g. chunking).

The key problem of these classes of approaches is the complexity in devising the suitable statistical models that optimize the transfer accuracy. They have to account for word level alignments, syntactic constituency in both languages, the symmetry of the semantic role alignment relation that feed the model estimation and for the optimization process. In [6] different models are studied and several model selection strategies are presented. The best reported models are based on full parses for both languages that compensate against noisy word alignments. However, these are also shown to be sensible to the parse errors, that are quite common. As errors cumulate across complex preprocessing stages, one of the major limitation of the semantic transfer approaches is their sensitivity to noise in basic preprocessing steps, that may critically deteriorate the overall quality of the transfer outcome. Robust transfer methods of English annotated sentences within a bilingual corpus should avoid complex alignment models to determine more shallow and reusable approaches to semi-supervised SRL. The aim of this paper is the investigation of an architecture based on a controlled, yet scalable, statistical machine translation process. It exploits the conceptual parallelism provided by Framenet and a distributional model of frame instance parallelism between sentences, that guarantees a controlled input to the later translations steps. It also employs a unified semantic transfer model for predicate and roles. The result is a light process for semantic transfer in a bilingual corpus. In Section 2, the overview and details of the proposed

process are discussed, while the experimental evaluation on a bilingual English-Italian corpus is discussed in Section 3.

2 Cross-Language Transfer of Frame Semantics in Aligned Corpora

Reusing semantically annotated texts in English within bilingual corpora implies the ability of transferring semantic information from the source language sentences to the target ones, as a form of translation of semantic units (i.e. predicates and roles) from one language to the other. The specific semantic transfer problem can not be seen as a pure translation process. The presence of relatively free translations in bilingual corpora in fact does not allow to track and recover all semantic phenomena in the target sentences. Moreover, as the sentence in the target language is already available, proceeding through a translation from scratch is not even required. A more specific definition is thus necessary.

Given a bilingual corpus in English and in a second target language T (e.g. Italian), the semantic transfer needs first to select sentence pairs (s_E, s_T) that effectively realize a specific frame f, and then provide the frame annotations for f in the target language sentence s_T. This process may proceed by labeling the English sentence s_E through an existing highly-performant SRL system, deriving multiple translation possibilities of English segments in s_E through statistical MT tools, and then building the best available semantic annotations within the target language sentence s_T. While related work on this process (including [6,8]) is generally based on complex syntactic models, our aim is to define a method relatively independent on the syntactic constraints on the two languages, in order to support a larger scale approach. The proposed process is depicted in Fig. 1. It combines a statistical translation tool (i.e. Moses) and a sentence selection model. This latter allows to decide which sentence pairs in the aligned corpus are effective realizations of frames. Statistical machine translation here is used to collect translation candidates for semantic information: every annotated role in the English portion of the corpus gives rise here to segments whose partial translations are available in terms of phrase translation pairs (PT pairs in Figure 1) from the corpus ([9]). These are thus post processed to get the suitable role boundaries in the target sentences (Semantic alignment step in Fig. 1).

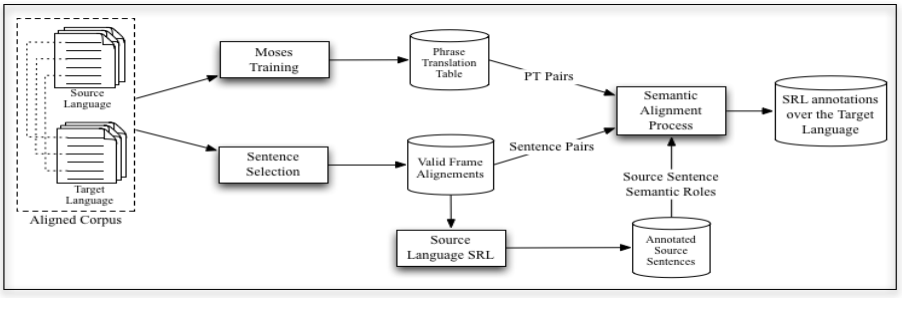

Fig. 1. The semantic transfer workflow

2.1 Cross-Language Predicate Level Alignment of Sentences

In a bilingual corpus, the parallelism of roles is conditional on the so-called *frame instance* parallelism ([6]): unless the frame expressed by two sentences is the same, the roles cannot be observed in parallel. The starting point of the semantic transfer approach is thus the selection of suitable sentences pairs as candidate expressions of frames. The underlying aligned corpus provides sentence paris (s_E, s_I) where Frame information about target predicates and roles (hereafter semantic elements) are both expressed in English and Italian. An aligned sentence pair (s_E, s_I) is a *valid* example of a frame f if both sentences express the specific semantic information related to f, i.e. exhibit conceptual and instance parallelism about f. We are interested to valid sentence alignments where the given frame f is known to manifest. The knowledge of predicate words of f (i.e. lexical units, $LU(f)$) in both languages is thus a starting point[1]. A pair (s_E, s_I) represents a *potentially valid frame alignment* for f iff $\exists p_E \in LU_E(f)$ and $\exists p_I \in LU_I(f)$ such that $p_E \in s_E$ and $p_I \in s_I$, where p_E or p_I are predicate words for f. However, this constraint is not sufficient as lexical units can be ambiguous so that not all valid frame alignments capture the same corresponding unique frame. In order for a pair to support the transfer of the semantic elements, the sentences must be known as expressions of the **same** frame f. For example in the sentence pair

s_E: *I will make his statement in English*
s_I: *Intendo farlo citando il suo intervento in inglese*

the verb *make* is not a predicate of the MANIFACTURE frame, although both *make* and *fare* are legal lexical units for the MANIFACTURE in both languages.

What it is needed here is a suitable model of valid frame alignments (s_E, s_I), that guarantees that a frame is expressed in s_E and s_I. At this aim we define the following function, called *pair frame relevance*, pf_rel:

$$pf\_rel((s_E, s_I), f) = \Gamma(\sigma_E(s_E, f), \sigma_I(s_I, f)) \quad (1)$$

where $\sigma_E(s_E, f)$ and $\sigma_I(s_I, f)$ measure the relevance of s_E and s_I respectively for f, and $\Gamma(.)$ is a composition function, such as the product or the linear combination.

The relevance $\sigma(s, f)$ of sentences for a given frame f is approached here according to methods based on semantic spaces already applied to LU classification ([11]). Semantic spaces are first built from co-occurrence analysis of lexical units, and distance in the resulting space is used to measure the suitable frame for possibly unknown predicate words. The method is semi-supervised as known lexical units of a frame f are used as examples of regions of the semantic space in which f manifests. First, a clustering process is applied to the set of known lexical units (LU) for f[2]. The centroids of the derived clusters are then used as a representation of f: distances from centroids are used to detect the suitable frames for vectors of unknown predicate[3]. In essence, the distance from clusters of a frame f represents cues to suggests frames of novel words. As Latent

[1] In [10], a LSA-based method to compute lexical classification also for Italian is presented, and, accordingly, a lexicon of about 15,000 predicate words has been made available. This resource is used across all the experiments reported in this paper.

[2] The adopted clustering process called *qt-kMean* [12] has been applied to collect these regions.

[3] In [10,11], this process is also strengthen by the use of Wordnet synonymy information.

Semantic Analysis [13] is applied to the original space, for its duality property, sentences (i.e. pseudo documents) can be expressed in the same space of LUs[4] : similarity between sentences and frames can be thus computed in terms of a distance function. Details of this process are discussed in [10].

Given a raw source corpus (e.g. the two monolingual portions of the bilingual corpus) a corresponding semantic space can be built. Then, the *Sentence Frame relevance* $\sigma(s, f)$ of a sentence s for a frame f is defined by:

$$\sigma(s, f) = max(0, max_{C_f}\{sim(\boldsymbol{s}, \boldsymbol{c(C_f)})\}) \quad (2)$$

where C_f are clusters derived from the known LU's of the frame f in the semantic space, $\boldsymbol{c(C)}$ is the centroid of the cluster C, $\boldsymbol{s}$ denotes the representation of s in the semantic space and $sim(.,.)$ is the usual *cosine similarity* among vectors. Notice how only k dimensions characterize the semantic space after the application of the Singular Value Decomposition (SVD) [13]. When any two corpora in English and Italian are available, two different semantic spaces are defined, but comparable scores $\sigma(.,.)$ can be obtained. As a consequence Eq. 2 and 1 can be computed for any language pair. The ranking determined among valid sentence pairs by Eq. 1 allows to automatically select the pairs for which conceptual parallelism for f is realized with high confidence. Notice that both sentences are constrained so that reliable pairs can be selected, SRL can be applied to their English side and, finally, the predicate and role alignments step towards the target language can be applied.

2.2 Robust Cross-Lingual Alignment of Frame Annotations

The task of computing the correct cross-lingual alignment of semantic information, as made available by an automatic frame annotation system, consists in the detection of segments expressing the semantic information related to the target predicate and to all the frame elements, as they are realized in the target language sentence s_I in a pair (s_E, s_I). As the translation s_I is often not literal, we can not assume that s_I *always* expresses *all* the FE observed in the English sentence s_E. However, exceptions are fewer, and the full labeling of s_I can proceed as a search for the segments in s_I triggered by the individual semantic elements found in s_E. In the following, we will adopt this view: each alignment choice is tailored to detect the unique segment in s_I able to realize the same information as one source semantic element annotated in s_E. Semantic elements here include the target predicates (usually verb phrases or nominal predicates in s_E) or phrases expressing some frame elements (FE): these are thus *always* explictly realized as segments in s_E. In the example, s_E: *I think this is something we should study in the future*, the segment "[*think*]" realizes the predicate OPINION while ["*this is something we should study in the future*"] accounts for the realization of the CONTENT FE.

Given a valid frame alignment pair (s_E, s_I), a role α and its realization in s_E, namely $s_E(\alpha)$, the alignment task can be thus formalized as the function $SemAl()$ defined by:

$$SemAl((s_E, s_I), \alpha, s_E(\alpha)) = s_I(\alpha) \quad (3)$$

[4] Any sentence s is represented as the linear combination of the vectors built from its words t, i.e. $\boldsymbol{s} = \Sigma_{t \in s}\omega(t, s) \cdot \boldsymbol{t}$, where $\omega(t, s)$ is the usual $tf \times idf$ score. $\boldsymbol{s}$ is finally normalized in the semantic space, where $\boldsymbol{t}$ are computed.

$SemAl(.)$ computes the proper segment $s_I(\alpha)$ that realizes α in s_I. As described in Fig. 1, the function $SemAl()$ proceeds by first detecting all the possible translations pairs for subsegments produced by a statistical MT tool (i.e. Moses), and then by merging and expanding the set of potential translation choices.

The Statistical Translation Step. Recently, MOSES ([9]), an open-source toolkit for statistical machine translation (SMT) has been released, exploiting the idea of factored translation models and confusion network decoding. It performs highly flexible phrase-level translation with respect to other traditional SMT models. Some of its key advantages are the exploitation of constraints (and resources) from different linguistic levels that are thus factored within a unique translation model. In [14], factored models on the Europarl corpus, [15], are shown to outperform standard phrase-based models, both in terms of automatic scores (gains of up to 2% BLEU) as well as grammatical coherence. In our work, the open source Moses system [9] has been used on the English-Italian aligned portion of the Europarl corpus [15]. During training, Moses produces translation models over phrase structures that are stored as phrase translation (PT) tables.

In this work, translation refers to the ability of cross-language mapping of individual semantic elements. The translation from English is thus not "blind" but guided by the expectations raised by the available sentence in the target language. Instead of relying on the automatic translation, it is possible to analyze only the partial translations of s_E that in fact appear in the target sentence s_I. In this case, simple phrase level translations are more useful, as they represent translations of partial elements from which the detection of the entire targeted role is enabled. Phrase-level alignments among the individual source sentences are made available as translation tables of English and Italian phrases (including singleton words). In the sentence "*I regard the proposed charter of fundamental rights as an opportunity to bring the european union closer to the people*" the following segment,

s_E: *as an opportunity to bring the European Union closer to the people.*

represents the role CATEGORY for the underlying CATEGORIZATION frame, introduced by the verb *regard*. The corresponding segment in the Italian counterpart is:

s_I: *come un'opportunita' per avvicinare l'Unione Europea ai cittadini.*

An excerpt of the phrase alignments provided by the Moses phrase translation (PT) table, acquired on the Europarl corpus, is shown in Fig. 2.(A-B). Notice how word pairs, e.g. *(closer, avvicinare, 0.06)*, *(closer, unione,0.00001)* in (A), are characterized by very low probabilities due to the relatively free translation: here the verb phrase "*bring closer*" is expressed by a single Italian verb *avvicinare*, and the translation mapping can not be more precise.

By extending the pairwise word alignments, Moses accounts for phrase-level alignments with probabilities. Moses phrase translation tables define all segments s_E that have a translation included in s_I, whereas, for a single semantic element, all its parts that have partial translations in s_I can be found, as shown in Fig. 2.(B). The output of the statistical alignment phase is thus a set of segment pairs (es^i, is^j) weighted according to a probability, describing a generally many to many mapping between an English semantic element and some is^j segments in s_I. Pairs include: word pairs as well as pairs where the English source is covered by a longer Italian segment (i.e.

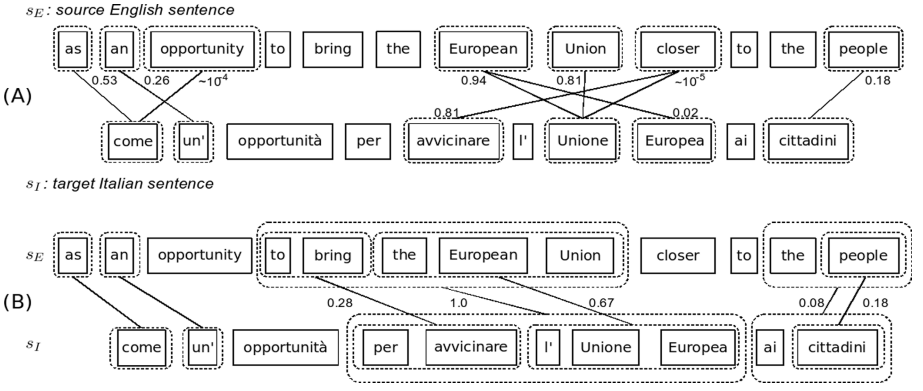

Fig. 2. An example of Moses alignments

$length(es^i) > length(is^j)$) or viceversa. A first *basic* algorithm for the function $SemAl((s_E, s_I), \alpha, s_E(\alpha))$ can be made dependent just on the Moses translation table. In this simple case, used hereafter as a baseline, the result $s_I(\alpha)$ is defined as the Italian segment is^j such that it exactly covers the English semantic element, i.e. such that it is translated from es^i with $es^i = s_E(\alpha)$.

In the example of Figure 2.(B), if a role α (e.g. THEME) characterize es^i=["*the European Union*"], the baseline alignment would result in ["*l'Unione Europea*"]. Unfortunately, in most cases, perfect matches are not made available as we will also see in section 3: roles are often realized in long segments, i.e. the targeted $s_E(\alpha)$, for which only partial segments es^i are translated. Further processing steps are thus needed to make a final decision about the best alignment of $s_E(\alpha)$ in s_I.

The above example shows that the length (k) of the English segment, the length of the Italian segment and the Moses output probabilities are all cues that characterize the quality of (partial) translation pairs (es^i, is^j) for the semantic transfer of a role α. Three different strategies can thus be used:

- English segment length policy, e_length: by adopting k as a ranking criterion, translation segments related to longer subsequences of the targeted ones, i.e. $s_E(\alpha)$ are preferred and selected first.
- Italian segment length, or i_length: the longer Italian are here preferred, so that better translation segments correspond to longer is^j.
- $simpleprob$: the $simpleprob$ policy ranks higher the segments es^i that appear in translation pairs with higher probabilities

Robust Cross-Lingual Semantic Alignments. The general algorithm for computing the semantic alignment is triggered by a sentence pair, (s_E, s_I), a specific element (e.g. a role) α and the English segment expressing the role $s_E(\alpha)$. It proceeds through the following steps:

1. *Rank phase*. Rank all the Moses translation segments related to at least one word in $s_E(\alpha)$, according to one policy (e.g. e_length).

2. *Collect Phase*. Scan the translation pair table, from the best pair to the worse ones, and select candidates for all token in $s_E(\alpha)$ until the target English segment is not covered by at least one translation. In this phase, all the Italian segments that are translations of a yet uncovered English segment es^i in $s_E(\alpha)$ are selected
3. *Boundary Detection Phase*. Process all the collected italian segments and compute the best boundary, i.e. $s_I(\alpha)$. This is done by possibly merging adjacent Italian candidate segments, or filling gaps between non-adjacent ones.
4. *Post-Processing Phase*. Refine the computed boundaries by applying heuristics based on the entire sentence, i.e. according to the candidate solutions of all different semantic elements. A typical task in this phase is the pruning of potential overlaps between translations $s_I(\alpha)$ of different roles built in the *Boundary Detection Phase*.

Notice that the above general process is greedy. First, the targeted English segment $s_E(\alpha)$ is early used to prune irrelevant portions of the (English and Italian) sentences. Second, the selected policy determines the order by which individual translation pairs are collected. Given the above general strategy, different ranking models and the adoption (or skip) of the post processing step characterize different workflows. As the *Boundary Detection Phase* provides complete solutions $s_I(\alpha)$, it can be also retained as a final step, without applying any post processing.

Collect Phase. The algorithm that compute translation candidates for individual roles α is in Fig. 3. The operators $\sqcap$ compute here the common subsequences among the segment operands, while $A \setminus B$ denotes the sequence obtained by removing the segment B from A.

```
FUNCTION Select( α,              % The targeted role
                 s_E(α)          % The targeted segment
                 MosesPairs)     % Ranked pairs (es^i, is^j)
  CandidateSeq = ∅
  CurrTargetSeq = s_E(α)
  while (CurrTargetSeq ≠ ∅) AND (MosesPairs ≠ ∅) do
    (es^i, is^j) = POP(MosesPairs)
    Overlap = CurrTargetSeq ⊓ es^i
    if Overlap ≠ ∅ then
      CurrTargetSeq = CurrTargetSeq \ Overlap
      CandidateSeq = CandidateSeq ∪ {(es^i, is^j)}
    end if
  end while
  return (CandidateSeq)
```

Fig. 3. The Algorithm for the *Collect Phase*

Boundary Detection Phase. Once candidate translation pairs are selected and ranked according to a given policy, a solution is then built by merging adjacent Italian segments is^j. As some words (or segments) may not appear in the translation tables, merging may not produce effective subsequences of the Italian sentence. In this case potential gaps between the selected is^j are filled. In the example of Fig. 2, the available translation pairs, are first selected and then merged to cover new portions of the English role segment, $s_E(\alpha)$. In this case, [*per avvicinare l'Unione Europea*] is first merged with

[*ai cittadini*] as they are the best selected segments in the first step. Then [*come*] and [*un*] are also added and merged as they translate new tokens in $\alpha(s_E)$ (i.e. [*as*], [*an*]). Finally, the gap between [*come un*] and [*per avvicinare l'Unione Europea ai cittadini*], due to the missing translation for the Italian [*opportunita'*], is filled: the final output boundary is [*come un'opportunita' per avvicinare l'Unione Europea ai cittadini*] that in fact captures the entire CATEGORY role for the underlying CATEGORIZATION frame.

Post-Processing Phase. The *Boundary Detection* process applies independently to individual roles (or predicates) α. It is thus possible that the produced solutions for different roles include partially overlapping segments. However, when the solutions for all roles α are made available, possible inconsistencies can be detected and ambiguities solved. For example, violations to the planarity of the solution (i.e. overlaps between the different output role segments), can be forced by some adjustment. One typical case, often caused by grammatical movements of inner constituents of roles, is given by output segments for frame elements that, in the Italian syntax, also include the target, as in:

s_E: [*I*]$_{Cognizer}$ [*think*]$_{target}$ [*this is something we should study in the future*]$_{Content}$.
s_I: *Lo reputo un tema meritevole di essere approfondito in futuro*.

Here the subject of the predicate is not expressed in Italian and the pronoun Lo, corresponding to the determiner $this$, is prefixed to the predicate. Here, the *Boundary Detection* algorithm produces the following wrong span for the CONTENT role: [*Lo reputo un tema meritevole di essere approfondito in futuro*], i.e. the entire s_I sentence. The objectve of the post processing here is to superimpose planarity by discarding embedded solutions. The original solution for CONTENT is first segmented in the two portions, [*Lo*] and [*un, tema, meritevole, di, essere, approfondito, in, futuro*], by cutting out the predicate. Then, the correct right segment [*un, tema, meritevole, di, essere, approfondito, in, futuro*] is selected as it constitutes the longer solution. The final full annotation of the CATEGORIZATION frame in the example s_I is:

Lo [*reputo*]$_{Target,Cognizer}$ [*un tema meritevole di essere approfondito in futuro*]$_{Content}$.

that is in line with the semantic expectations provided by s_E[5].

3 Evaluation

There are mainly two different aspects of the proposed semantic transfer process worth of an in depth investigation. The first is the evaluation of the Sentence extraction step (Section 3.1) as determined by Equation 1. The second is the evaluation of accuracy of the overall semantic transfer, as reachable by the technique proposed in Section 2.2.

The computation of the ranking factor defined in Eq. 1 requires a vector representation for both the English and Italian sentences. As described in [10], the semantic space is derived through LSA, over the English and Italian components of the Europarl corpus

[5] The ellipsis of the agentive role, $Cognizer$, for the Italian sentence is here expressed through the multiple tags for the predicate word *reputo*. Notice that these multiple tags are not considered during the evaluation discussed in Section 3 and only the independently realised roles, i.e. the target in this case, are measured.

[15]. The vector components express occurrence of predicates in individual sentences (i.e. pseudo-documents), these latter used as features. The semantic space accounts for about 1 million sentences (i.e. 36 millions tokens), used as contexts for computing the co-occurrence vectors for individual words, including the targeted LUs. The SVD reduction with $k = 300$ allows to compute a 300-dimensional vectors for each word: sentences are accordingly represented by the linear combination of the vectors of their words. In all the experiments, the open source Moses system [9] has been used on the English-Italian aligned portion of the Europarl corpus [15]. Default settings are used in all the experiments.

For the evaluation of the semantic transfer accuracy, a gold standard, built from the aligned English-Italian component of the Europarl corpus, has been used. This gold standard, presented in [8], is made of 987 sentences in both languages English and Italian. The gold standard has not a complete alignment. As discussed in [8], only 61% of the sentences are annotated with the same frame, while only 82% have the same FEs in both languages. This is mainly due to the different versions of Framenet used for English (i.e. 1.1) and Italian (i.e. 1.3), as reported in [8]. As we are interested to the transfer achievable through automatic alignment of the source English annotations, we considered only the different FE alignments independently from the underlying Frame. As a consequence, the relevant test cases are only those FEs having the same label in both languages. In general this assumption does not cover all cases, but it gives a significant idea about the potential of the semantic transfer on a reasonable scale. In the gold standard, 1,727 and 1,730 frame elements were found respectively for the English and Italian component, where 881 were shared. In the 987 sentences, 984 target lexical units were aligned[6]. As the transfer of individual semantic elements proceeds from the English to the Italian sentences, we are interested in: (1) Perfect matches, i.e. the percentage of output Italian segments that are fully overlapping with the gold standard ones, (2) Partial Matches, i.e. the percentage of Italian segments with non empty intersection with the gold standard. Moreover, we also want to measure the quality of the computed approximation for each semantic element in terms of tokens. Thus we evaluate the token retrieval quality for all the translated source English roles against the Italian gold standard. The token retrieval task is measured according to the usual precision, recall and F-measure scheme: a token in $s_I(\alpha)$ is correct if it also part of the segment for α proposed by the oracle. False positives and negatives are given by tokens found only in $s_I(\alpha)$ or in the oracle respectively. These measures are a fine-grain evaluation of the overlaps between the solutions and the oracle.

3.1 Evaluating the Sentence Extraction Model

The evaluation of the sentence extraction accuracy is carried out by studying the probability distributions of the frame preference scores (Eq. 1), as computed over three sentence pair sets of similar cardinality (about 1,000 sentences). The first Control Set ($CS1$) includes sentence pairs where frame assignment is randomly applied: in this case, a randomly chosen frame f is selected for each pair and the scores $\sigma(s, f)$ are

[6] Three sentences have been neglected as for text encoding problems in the original gold standard.

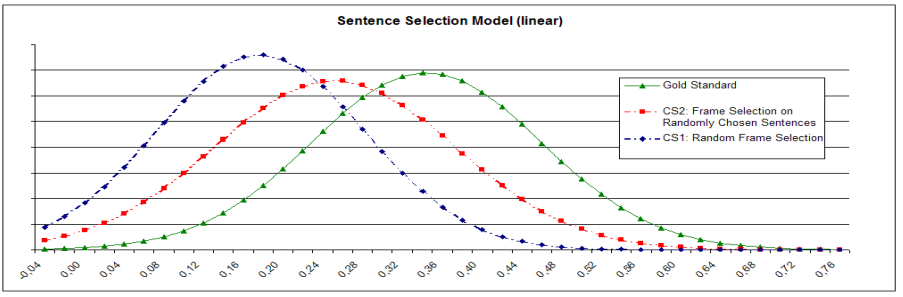

Fig. 4. Distributions of frame preference scores over the oracle and two reference Control Sets

used to compute Eq. 1 from the English and Italian sentences. A second Control Set ($CS2$) is obtained by selecting pairs for which the English sentence includes a known lexical unit of a frame f: such sentences and their Italian equivalent are then used to compute the $\sigma(s, f)$ scores in Eq. 1. Finally, the model in Eq. 1 is computed over the Oracle sentence pairs: here the frame f is known to be correct. In Fig. 4 the normal probability distributions $P(\Gamma = x)$ are reported for the three sets, where the composition function Γ is the linear combination of scores $\sigma(s, f)$ with equal weights (i.e. 0.5). As clearly indicated by the plots the mean values of the three distributions are significantly different. Increasing evidence given by higher semantic relevance scores Γ of sentence pairs corresponds to correct frames (as in the oracle). The difference between the first and the second Control Sets suggests that the knowledge about lexical units is important and it is well captured by the LSA similarity. When frame relevance holds for both languages (as implicitly true in the oracle, where the frame preference $\sigma(s, f)$ of a sentence is guaranteed to be correctly applied on both languages), the result is a strikingly higher score for the sentence pair (i.e. $\mu_{Oracle} \cong 0.36$ vs. $\mu_{CS1} \cong 0.18$). Evidence in Fig. 4 confirms that Eq. 1 allows to accurately rank sentence pairs as suitable representations for a given frame. This is useful for all the material to be annotated by an automatic process outside the gold standard, where a good conceptual (i.e. frame) parallelism is needed.

3.2 Evaluating the Overall Accuracy of the Semantic Transfer

The evaluation of the semantic transfer from English to Italian (i.e. the task described in Section 2.2) has been carried over the English-Italian gold standard. As for the mentioned mismatches between the adopted labeling for Italian and English data, tests are tailored to the subset of roles (i.e. targets and frame elements) that have the same label in both languages. The tested models are derived from the application of different ranking policies (e.g. e_length vs. $simpleprob$) as well as in the adoption of the post-processing phase ($+PP$ in table 1). The accuracy is evaluated independently over all semantic elements or just on roles (*FE only*). In this latter case, we simply neglect the targets in the accuracy computation. The baseline refers to the output of the basic algorithm defined in Section 2.2. It relies only on the Moses translations and refers to the best solution obtained through a direct look-up in the Moses PT tables.

Table 1. Accuracy of the role alignment task over the Gold Standard

Model	Perfect Matching (FE only)	Partial Matching (FE only)	Token Precision	Token Recall	Token F1
baseline	66.88% (28,37%)	72,78% (41,13%)	0.78 (0.59)	0,29 (0.14)	0.43 (0.23)
e_length	72,02% (39,48%)	90,98% (80,50%)	0.75 (0.71)	0.88 (0.85)	0.81 (0.78)
simpleprob	71,69% (38,77%)	**91,09**% (**80,73**%)	0.74 (0.70)	0.88 (0.85)	0.80 (0.77)
i_length	69,51% (34,04%)	89,56% (77,42%)	0.73 (0.69)	0.89 (0.86)	0.80 (0.77)
e_length (+PP)	**73,28**% (**42,20**%)	**89,94**% (**78,25**%)	**0.84** (**0.81**)	0.84 (0.81)	**0.84** (**0.81**)
simpleprob (+PP)	**73,28**% (**42,20**%)	89,83% (78,01%)	0.84 (0.80)	0.84 (0.81)	0.84 (0.81)
i_length (+PP)	70,92% (37,12%)	88,36% (74,82%)	0.82 (0.80)	**0.84** (**0.81**)	0.83 (0.79)

Table 1 reports the accuracy of perfect and partial matchings. Notice how the perfect matching corresponds to the usual SRL evaluation as applied to the labeling of the Italian test corpus: perfect matches here corresponds either to perfect boundary recognition and role classification. The last columns in Table 1 measure the gap in accuracy between *Perfect* and *Partial Matches*. Higher values in F1 suggest that tokens violating predicate and role boundaries are fewer.

As shown in table 1, the best model (i.e. $e\_length + PP$) achieves perfect matching for 42% of the Frame Elements (excluding target words) and 73% of all roles in the test sentences. Results for partial matching, according to the same approach reach percentage of respectively 78,25% and 89,94%. This shows that the proposed approach are almost everywhere able to find the correct core of individual semantic elements. Only few tokens violate boundaries, but most of the FE semantics is preserved. This is confirmed by the evaluation of tokens retrieval (see last three columns in Table 1), as a 81% of F1 is achieved only on the transfer of FEs. Notice how all the models are well above the baseline, obtained by relying just on Moses phrase translation pairs. This is particularly noticeable on FEs: notice that this is mainly due to the fact that targets are usually expressed by shorter segments, in general verbs, for which the higher frequencies in the Europarl allow Moses to produce more accurate translations. This is unfortunately no longer true for semantic roles, for which the baseline performs quite poorly, about 28% perfectly matched roles, with F1=0.23 at the token level.

4 Conclusions

Complex models for semantic cross-lingual transfer of Framenet information require highly performant parser and complex model optimization. In this paper a light, yet robust, semantic transfer method has been presented aiming to produce large scale frame semantic annotations over bilingual corpora. Although no direct comparison was made possible with respect to previous work (basically, for major differences in the adopted languages, measures and representations), the obtained results appear superior to previously proposed methods. A public distribution of the aligned material is foreseen for stimulating further comparative analysis. The adoption of unsupervised techniques for sentence selection as well as the poorer requirements of the semantic transfer approach

here proposed imply a larger applicability with more space for improvements. First of all, the approach is open to improvement through further grammatical analysis of the proposed alignments: chunking and parsing can be still applied to refine possibly wrong solutions and increase the token-level precision. Moreover, better statistical modeling of alignment preferences (through joint bayesian models) should be investigated to further improve the boundary detection step. The presented methodology has been currently applied to extend to current English-Italian gold standard of [8]. An existing SRL system, described in [16,17,18], has been used to annotate data outside the gold standard, i.e. about 20,831 sentences. As a result about 17,765 among the analysed sentences have been annotated in Italian with two or more roles. A relevant open issue is thus the evaluation of its impact on the learning of the current SVM-based SRL system for Italian. If the potential advantages in adopting a large scale (but noisy) training set with respect to smaller high-quality gold standards could be assessed, this would definitively open new perspectives on the use of bilingual corpora for a semi-supervised approach to SRL training.

Acknowledgments. The authors are thankful to the FBK group for granting the access to the gold standard developed by Emanuele Pianta and Sara Tonelli at FBK.

References

1. Baker, C.F., Fillmore, C.J., Lowe, J.B.: The Berkeley FrameNet project. In: Proc. of COLING-ACL 1998, pp. 86–90 (1998)
2. Fillmore, C.J.: Frames and the semantics of understanding. Quaderni di Semantica 4(2), 222–254 (1985)
3. Palmer, M., Gildea, D., Kingsbury, P.: The Proposition Bank: an Annotated Corpus of Semantic Roles. Computational Linguistics 31(1), 71–106 (2005)
4. Gildea, D., Jurafsky, D.: Automatic Labeling of Semantic Roles. Computational Linguistics 28(3), 245–288 (2002)
5. Carreras, X., Màrquez, L.: Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling. In: Proc. of CoNLL 2005, Ann Arbor, Michigan, pp. 152–164 (2005)
6. Padïo, S.: Cross-lingual annotation projection models for role-semantic information. PhD Thesis, Dissertation, Universität des Saarlandes, Saarbrücken, Germany (2007)
7. Padïo, S., Pitel, G.: Annotation prïecise du francais en sïemantique de roles par projection cross-linguistique. In: Proc. of TALN 2007, Toulouse, France (2007)
8. Tonelli, S., Pianta, E.: Frame information transfer from english to italian. In: Proc. of LREC Conference, Marrakech, Marocco (2008)
9. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open source toolkit for statistical machine translation. In: Annual Meeting of the Association for Computational Linguistics (ACL), Demonstration Session, Prague, Czech Republic (2007)
10. De Cao, D., Croce, D., Pennacchiotti, M., Basili, R.: Combining word sense and usage for modeling frame semantics. In: Proc. of The Symposium on Semantics in Systems for Text Processing (STEP 2008), Venice, Italy, September 22-24 (2008)
11. Roberto, B., De Cao, D., Pennacchiotti, M., Croce, D., Roth, M.: Automatic induction of framenet lexical units. In: Proc. of the 12th International Conference on Empirical Methods for NLP (EMNLP 2008), Honolulu, USA (2008)

12. Heyer, L., Kruglyak, S., Yooseph, S.: Exploring expression data: Identification and analysis of coexpressed genes. Genome Research (9), 1106–1115 (1999)
13. Landauer, T., Dumais, S.: A solution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. Psychological Review 104, 211–240 (1997)
14. Koehn, P., Hoang, H.: Factored translation models. In: Proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, Czech Republic, pp. 868–876 (2007)
15. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: Proc. of the MT Summit, Phuket, Thailand (2005)
16. Moschitti, A.: Making Tree Kernels Practical for Natural Language Learning. In: Proc. of EACL 2006, pp. 113–120 (2006)
17. Moschitti, A., Pighin, D., Basili, R.: Tree Kernels for Semantic Role Labeling. Computational Linguistics Special Issue on Semantic Role Labeling (3), 245–288 (2008)
18. Coppola, B., Moschitti, A., Pighin, D.: Generalized Framework for Syntax-based Relation Mining. In: Proceedings of the IEEE International Conference on Data Mining (ICDM 2008), Pisa, Italy (2008)