Cross-lingual Alignment of FrameNet Annotations through Hidden Markov Models

Paolo Annesi and Roberto Basili

University of Roma Tor Vergata, Roma, Italy
{annesi,basili}@info.uniroma2.it

Abstract. The development of annotated resources in the area of frame semantics has been crucial to the development of robust systems for shallow semantic parsing. Resource-poor languages have shown a significant delay due to the lack of sufficient training data. Recent works proposed to exploit parallel corpora in order to automatically transfer the semantic information available for English to other target languages. In this paper, an approach based on Hidden Markov Models is proposed to support the automatic semantic transfer and use an aligned bilingual corpus to develop large scale annotated data sets. As this method relies just on lexical alignment of sentence pairs, it is robust against preprocessing errors and does not require complex optimization, like syntax-dependent models for accurate cross-lingual mapping. The experimental evaluation over an English-Italian corpus is successful, achieving 86% of accuracy on average, and improves on the state of the art methods for the same task.

1 Introduction

In the studies on *frame semantics*[1], the development of tools targeted to languages for which annotated corpora, such as FrameNet [2], are not available has a limited and slower development. Machine learning methods, making use of annotated *resources* to train statistical learning NLP tools, cannot be optimized in an effective manner [3]. For this reason parallel or aligned corpora are particularly interesting. Annotations for resource-poor languages, such as Italian, are projected from the texts aligned with a second language like English.

In [4], several projection and transfer algorithms are proposed for acquiring monolingual tools from aligned multilingual resources. The study in [5] estimates the degree of syntactic parallelism in dependency relations between English and Chinese. Nevertheless direct correspondence is often too restrictive and syntactic projection yields good enough annotations to train a dependency parser. A bilingual parser that comes with a word translation model is proposed in [6]. In the frame semantics research, Chinese FrameNet is built up in [7] by mapping English FrameNet entries to concepts listed in HowNet[1], an on-line ontology for Chinese, however without exploiting parallel texts.

[1] http://www.keenage.com

Recent work 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 ([8,3,9]). In [3] an annotation projection by inducing FrameNet semantic roles from parallel corpora is presented, where investigation on whether semantic correspondences can be established between the two languages is discussed. The presented methods automatically induce semantic role annotations for a target language whereas a general framework for semantic projection that can incorporate different knowledge sources is introduced. This work distinguishes predicates alignment from roles alignment, relying on distributional models of lexical association for the first task and on the linguistic information encoded in the syntactic bracketing for the latter one. Results are characterized by higher-precision projections even over noisy input data, typically produced by shallow parsing techniques (e.g. chunking). These approaches have a significant complexity in devising the suitable statistical models that optimize the transfer accuracy. Moreover, they can be effectively used to develop Semantic Role Labeling (SRL) systems in a resource poor language. SRL is first applied to English texts and this makes it possible to label the English portion of a bilingual corpus with a significant accuracy. The large volumes of information can be thus derived, in a relatively cheap way, through cross-language transfer of predicate and role information. A method that avoids complex alignment models to determine more shallow and reusable approaches to semi-supervised SRL has been presented in [10]. It defines a robust transfer method of English annotated sentences within a bilingual corpus. This work 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. Even if this approach provides a simple process for semantic transfer, it is based on heuristic rules about word alignments and role segmentation.

The aim of this paper is to investigate a more robust method based on statistical principles, namely Hidden Markov Models (HMMs), aiming to map the semantic transfer problem into a sequence labeling task. The objective is to carry out semantic role transfer between aligned texts in a bilingual corpus with a very high accuracy. In sections 2, we discuss the markov model adopted in this study by providing the overview and the formal definitions of the proposed process. The experimental evaluation on a bilingual English-Italian corpus is discussed in Section 3.

2 An Hidden Markov Model of the semantic transfer

The semantic transfer task consists in mapping the individual segments of an English sentence expressing semantic roles, i.e. target predicates or Frame Elements [2], into their *aligned* counterparts as found within the corresponding Italian sentence. In Fig.1 an example of a semantic transfer task is shown. In

this case the predicate (i.e. the *Lexical Unit* (LU), also called the *target* hereafter, for the frame EVENT) is *happen*. The semantics of the sentence also defines the TIME role, through the segment *after 2006*. The semantic transfer task here is to associate "*happen*" with the verb "*accadrá*" and "*after 2006*" with the fragment "*dopo il 2006*" in the Italian sentence. Given a parallel corpus with the English component labeled according to semantic roles, we aim at detecting, for each segment in an English sentence expressing the role X, the substring in the corresponding Italian sentence that exactly define X.

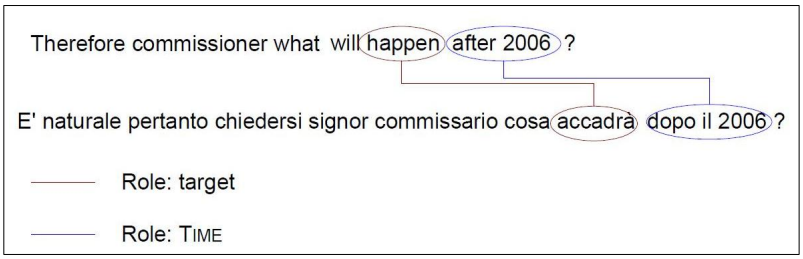

Fig. 1. Cross-language transfer: an example

For this reason, we will assume hereafter that the English component is always labeled through a SRL software that supplies all the (valid) semantic roles according to FrameNet database.

Let us define an English sentence f as $es_f = (e_1 \dots e_m)$, i.e. a sequence of words e_j, and the corresponding set of indexes j as $EnI = \{1, \dots, m\}$. Analogously, we define an Italian sentence f as the sequence of words $is_f = (w_1 \dots w_n)$, and the set of indexes as $ItI = \{1, \dots, n\}$. Giza [11] is a publicly available machine translation tool based on HMM word alignment models that provides the alignments between individual Italian and English word pairs: these are produced by Giza according to translation probabilities as estimated across an entire bilingual parallel corpus. Notice how all the English words related by Giza with an Italian word can be seen as the emissions of the latter word, given the corpus. Hereafter, the set of the emissions for each $i-th$ Italian word is defined as $E_i = \{e_1, \dots, e_n\}$ whereas every e_j is a possible translation of the Italian word w_i. Now, in the perspective of the semantic role transfer task sketched above, every Italian word can be characterized by one of the following *three* states:

1. it is *inside* the semantic role, as it translates one or more English words that are part of the role, as in the case of *dopo* in the Fig. 1 example
2. It appears *before* any other Italian words translating any part of the semantic role, i.e. it is *out of the role on its left* like *commissario*
3. It appears *after* any other Italian words translating any part of the semantic role, i.e. it is *out of the role on its right* like the token "?".

In this view, the semantic role transfer problem can be always mapped into a sequence labeling task, as for every role in an English sentence we need to

tag individual words in the Italian sentence with the three labels corresponding to the above states. In Figure 1, the English words $\{after, 2006\}$ compose the substring of es that defines the TIME semantic role, namely α: this substring will be hereafter denoted by $es(\alpha)$. In analogy with the English case, we will denote by $is(\alpha)$ the analogous substring of the Italian sentence is that expresses the same role α.

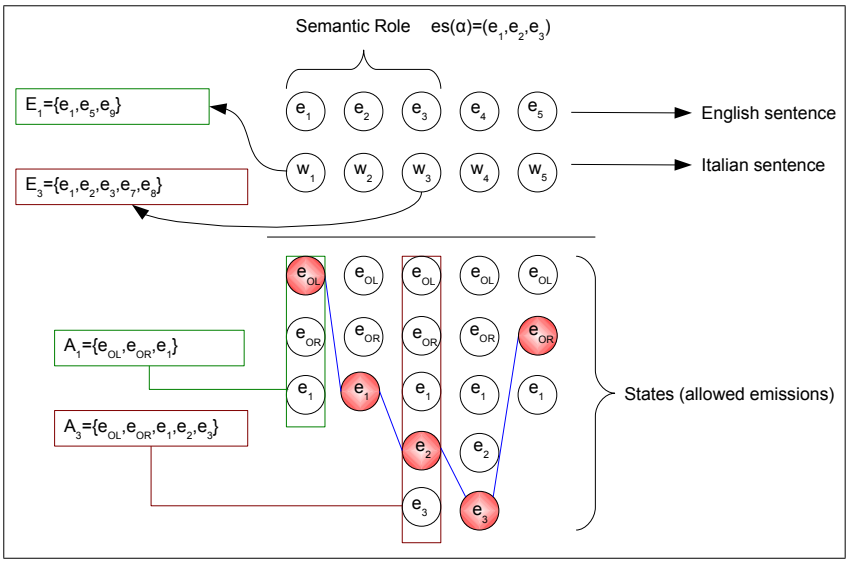

Fig. 2. Cross-language transfer: states and resolution

Notice that for the translations E_i of an Italian word w_i to be acceptable for any role α they must also appear in the segment $es(\alpha)$. In the example of Fig. 2, the set of words E_i must belong to the set[2] $E_i \cap es(\alpha)$, that defines the useful potential translations of the word w_i for the segment corresponding to the semantic role α. Notice how, in a generic sentence pair (es, is), every translation maps an Italian word $w_i \in is$ into one of its valid translations. The members of the set in (**??**) can be seen thus possible as state labels referring to individual English words e_k, whenever these latter appear in the English segment $es(\alpha)$ expressing a role α.

On the contrary, whenever an Italian word is not involved in the role α, i.e. it appears *before* or *after* the segment $is(\alpha)$, we will use the alternative tags e_{OL} and e_{OR}. These latter define that the i-th Italian word w_i does not belong to the targeted semantic role $is(\alpha)$. The set of valid labels for every Italian word

[2] It should noticed here that the sequence $es(\alpha)$ is in fact used as a set, with an odd but still understandable use of the corresponding membership function.

w_i are thus defined as $A_i = (E_i \cap es(\alpha)) \cup \{e_{\text{OL}}, e_{\text{OR}}\}$. An example of these labels is reported in Figure 2 as columns under every Italian word w_i.

Let us introduce the function $\theta(i)$ that, given is and $es(\alpha)$, couples each Italian word $w_i \in is$ with an English word $e_j \in A_i$, i.e. a possible translation or the special labels e_{OL} or e_{OR}. This function can be defined as follow

$$\theta(i) = j \quad \text{with} \quad e_j \in A_i \tag{1}$$

In Fig. 2, every i-th state is related with an emission in A_i. In this example the Italian word w_1 has its own set of emissions consisting of the English words $\{e_1, e_5, e_9\}$. Notice that as e_5 and e_9 do not belong to the English role subsequence es(α), they are not included in set A_1, that consist only of e_1, e_{OL} and e_{OR} indeed. The darker states define the resolution path that retrieves the Italian semantic role that is the words sequence (w_2, w_3, w_4). On the contrary the words w_1 and w_5 are labeled as outside the role on the left and on the right respectively.

The selection of the state sequence as the best labeling for a role $es(\alpha)$ is a *decoding task*: it requires to associate probabilities to all the possible transfer functions $\theta(\cdot)$, so that a transfer can be more likely than another one. Every state $e_j \in A_i$ is tied with the observation of the Italian word w_i: it represents the specific j-th translation as shown in the English sentence es. States $e_j \in A_i$ establish the correspondence between word pairs (w_i, e_j): for all the words $e_j \in es(\alpha)$ the state sequence provides the correspondence with the Italian segment $is(\alpha)$. The resulting HMM, given an Italian sentence is of length n, provides the most likely sequence of states $S = (e^1_{j_1}, e^2_{j_2}, \dots e^n_{j_n})$, whereas every $e^k_{j_k} \in A_k$: S identifies the Italian words w_i whose "translations" are inside the English segment $es(\alpha)$, i.e. $e_{j_i} \notin \{OL, OR\}$. Figure 2 reports an example where $es(\alpha) = (e_1, e_2, e_3)$ and the state sequence suggests $is(\alpha) = (w_2, w_3, w_4)$.

2.1 States, emissions and transitions for semantic transfer

In order to develop our Markov model of a semantic transfer task, let us discuss it through an example. Given the Italian sentence "*É naturale pertanto chiedersi signor commissario cosa accadrá dopo il 2006?*" and the corresponding English one "*Therefore commissioner what it will happen after 2006?*", we want to transfer the semantic role of TIME represented by the words "*after 2006*". FrameNet define the English sentence as follow

Therefore commissioner what it will ***happen*** [*after 2006* TIME]?

We suppose that this correct labeling is proposed by a existing SRL software and every role label is given as input. In order to analyze the role TIME, Fig. 3 shows how es(TIME) influences the set of possible states A_i for every Italian word.

In Table 2.1, the emissions supplies by Giza for each Italian word are shown. Therefore each Italian word w_i is associated to a set of English emissions in A_i. Notice how even if many less likely alignments have been neglected in Table 2.1,

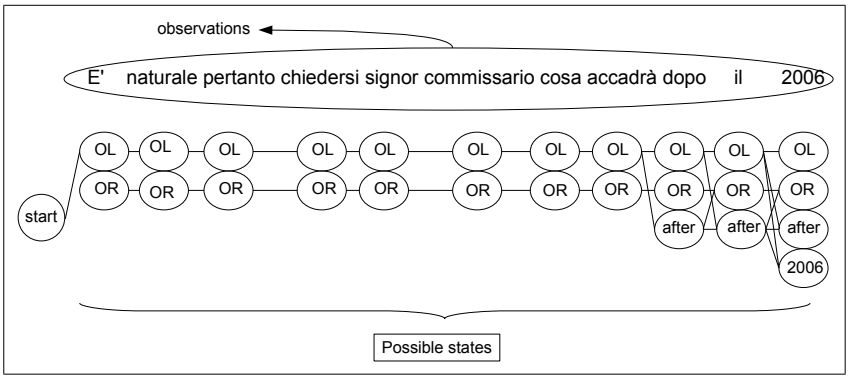

Fig. 3. States concerning the semantic transfer example and possible transition

the number of candidate translations supplied by Giza is still large. However, the set of the useful word alignments for the role TIME is restricted due to the parallelism between es(TIME) and is. In Fig. 3 all the possible states for this

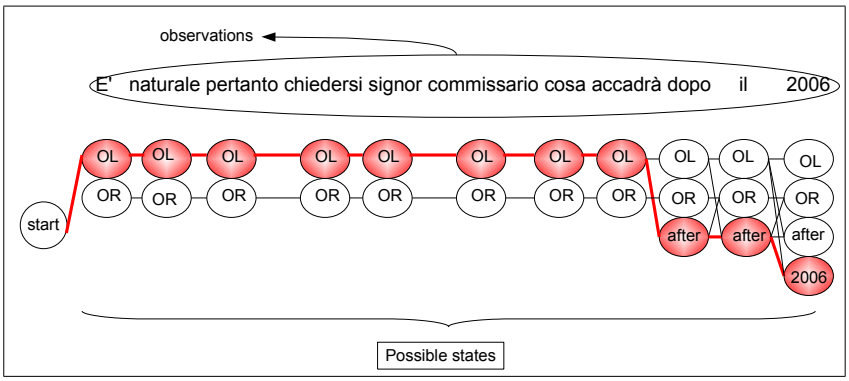

Fig. 4. Possible solutions and best solution

specific semantic transfer task are shown. States are those derived from all A_i sets for every observation i. In this way the available states for the first 8 observations are just e_{OL} and e_{OR}, since Giza align all the first 8 Italian words with English words not in the "*after 2006*" segment. Notice that some connections between states are not allowed. The *out right* state is not reachable from the *start* state, as we first have to pass trough an *out left* state or an English word emission state (the latter is not available in this particular case). The *out right* state can not reach an *out left* state obviously. Finally a state with a role English word can not be connected with an *out left* state.

Position i	**Italian word** $\mathbf{w}_i$	**English translations** $\mathbf{E}_i$
1	É	*was, this, is, that, therefore, ...*
2	naturale	*water, environmental, quite, natural, ecosystem, ...*
3	pertanto	*conclusion, changed, upon, elapsed, cautious, ...*
4	chiedersi	*know, request, now, days, think, asking, ...*
5	signor	*he, commissioner, echo, barroso, ...*
6	commissario	*frattini, dimas, chile, gentlemen, ...*
7	cosa	*topics, uphold, what, ...*
8	accadrá	*happen, supplied, go, prospects, ...*
9	dopo	*from, had,* ***after****, next, ...*
10	il	*when, basis, until,* ***after****, ...*
11	2006	***2006, after****, vienna, current, period, ...*

Table 1. Emissions supplied by Giza for the Italian words concerning the sentence in the example of Fig. 1. Note that the English segment es(TIME) is the substring, i.e. "*after 2006*"

In Fig. 4 all the possible solution paths are shown. The darker path is the selected most likely one. Our task is to derive the best function $\widehat{\theta}(i)$ in terms its overall probability among all the possible alternative $\theta(i)$. In this example the best path associates the input English substring "*after 2006*" with the Italian substring "*dopo il 2006*".

Using an Hidden Markov Model for the semantic role transfer task means to define a Bayes inference that consider all the possible state sequences given the observable emissions. Associating a probability to each transfer functions $\theta(i)$ we select the most likely sequence $\widehat{\theta}(i)$ that solve our transfer task as follows:

$$\widehat{\theta}(i) = \underset{\theta(i)}{\operatorname{argmax}} P\big(\theta(i)|es(\alpha), is\big) \tag{2}$$

By applying the Bayes rule to Eq. 2, we reduce it as follows:

$$\widehat{\theta}(i) = \underset{\theta(i)}{\operatorname{argmax}} P\big(is, es(\alpha) \mid \theta(i)\big)\ P\big(\theta(i)\big) \tag{3}$$

In Eq. 3, we distinguish two probabilities: the left one, $P\big(is, es(\alpha) \mid \theta(i)\big)$, i.e. the *emission* probability, and the right one, $P\big(\theta(i)\big)$, that is the *transition* probability. The emission probability is the probability that links a word $w_i \in is$ with its English counterpart through the selection of the state in A_i. The transition probability is the probability to cross a path between states, i.e. entering into a role and exiting correspondingly after having consumed some valid translations e_j. A first assumption about the emission probability is that the probability of an Italian word depends only on its own emissions. So we can retype this

probability as follow

$$P\big(is, es(\alpha) \mid \theta(i)\big) \approx \prod_{i=1}^{n} P(w_i \mid e_{\theta(i)}) \tag{4}$$

in which the emission probabilities do not depend on previous states in a path, so that the product of the emission probability can be used. A second assumption about the transition probability is that the state at step i only depends on the state $i-1$, so that the transition probability is given by

$$P\big(\theta(i)\big) \approx \prod_{i=2}^{n} P\big(\theta(i) \mid \theta(i-1)\big) \tag{5}$$

Finally replacing Equation 4 and 5 into Equation 3, we have

$$\widehat{\theta}(i) \approx \underset{\theta(i)}{\operatorname{argmax}} \prod_{i=1}^{n} P(w_i \mid e_{\theta(i)}) \prod_{i=2}^{n} P\big(\theta(i) \mid \theta(i-1)\big) \tag{6}$$

where the first one is the emission probability and the second one is the transition probability.

Estimating Emission probabilities. The emission probability expressed in Eq. 6 can be retyped using Bayes rule as:

$$P(w_i|e_{\theta(i)}) = \frac{P(e_{\theta(i)}|w_i)\ P(w_i)}{P(e_{\theta(i)})} \tag{7}$$

The probability $P(e_{\theta(i)}|w_i)$ defines the coupling between an Italian word w_i and an English one e_j supplied by the mapping $\theta(i)$. For $j \neq OL$ and $j \neq OR$ this probability is given by Giza. $P(w_i)$ defines the probability to extract w_i randomly from our corpus. Similarly $P(e_{\theta(i)})$ is the probability to extract $e_{\theta(i)}$ randomly from our corpus, that is the English word chosen by our transfer function. Given $P(w_i) = \frac{C(w_i)}{N_{it}}$ where $C(w_i)$ is the function that counts all the w_i occurrences in our corpus and N_{it} is the Italian corpus size, we define $P(w_i) = \frac{C(w_i)+1}{N_{it}+|D_{it}|}$ by applying a smoothing where $|D_{it}|$ is the size of the Italian vocabulary. Analogously, $P(e_{\theta(i)}) = \frac{C(e_{\theta(i)})+1}{N_{en}+|D_{en}|}$.
Equation 7 can be thus rewritten as

$$P(w_i|e_{\theta(i)}) = P(e_{\theta(i)}|w_i)\ \frac{C(w_i)+1}{C(e_{\theta(i)})+1}\ \frac{N_{en}+|D_{en}|}{N_{it}+|D_{it}|} \tag{8}$$

in which three emission probabilities, depending on the value of $\theta(i)$ are represented. When an Italian word w_i is part of the semantic role the emission probability of an English word is defined as

$$P(w_i|e_j) = P(e_j|w_i)\ \frac{C(w_i)+1}{C(e_j)+1}\ \frac{N_{en}+|D_{en}|}{N_{it}+|D_{it}|} \tag{9}$$

where $P(e_j|w_i)$ is given by Giza.

When an Italian word is outside a semantic role (i.e. $\theta(i) = OL$ or $\theta(i) = OR$) the corresponding emission is estimated as

$$P(w_i|e_{\text{OL}}) = \frac{\sum_{is} \sum_{\alpha \in is} \delta_{OL}(w_i, is, \alpha)}{\sum_{is} \sum_{\alpha \in is} \sum_{w_i \notin is(\alpha)} \delta_{OL}(w_i, is, \alpha)} \tag{10}$$

whereas the function $\delta_{OL}(w_i, is, \alpha)$ as well) is given by

$$\delta_{OL}(w_i, is, \alpha) = \begin{cases} 1 & \text{if } w_i \text{ is on the left of } is(\alpha) \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

Notice that $\delta_{OL}(w_i, is, \alpha)$ counts the occurrences of w_i on the left of a semantic role α and it has a counterpart in the function $\delta_{OR}(w_i, is, \alpha)$ that counts the right co-occurrences.

As for this kind of emission probabilities, we apply smoothing so that Eq. 10 becomes $P(w_i|e_{\text{OL}}) = \frac{\left(\sum_{is} \sum_{\alpha \in is} \delta_{OL}(w_i, is, \alpha)\right) + 1}{\left(\sum_{is} \sum_{\alpha \in is} \sum_{w_i \notin is(\alpha)} \delta_{OL}(w_i, is, \alpha)\right) + |D_{it}|}$
Finally, $P(w_i|e_{\text{OL}})$ has its obvious counterpart $P(w_i|e_{\text{OR}})$ for the words w_i on the right of any semantic role.

Estimating transition probabilities. The transition probability constraints the overall likelihood of a path through the markov model of a semantic transfer task. Every transition depends only on the current, i.e. k-th, state and on the next $k-1$-th state. The type of a state is defined by the attributes in A_i as defined in Section 1. A transition is determined by the choice of a mapping function $\theta(i)$ that decides how to map an incoming words w_i. $\theta(i)$ clearly depends on the word itself w_i as it characterizes the best possible translations of w_i in the targeted English sentence es. However computing a lexicalized estimate of the probability $P\big(\theta(i) \mid \theta(i-1)\big)$ is problematic as data sparseness would limit the suitable treatment of rare and unseen phenomena (i.e. unigrams and bigrams absent from the training corpus).

The model presented hereafter departs from the hypothesis of a lexical estimate and generalizes it according to the three macro labels (the syntactic states of being before, within or after a semantic role). This gives rise to a lower number of transition types between states and not words, that are depicted in Table 2. Note that the transitions that enter (or exit) in (from) a semantic role (i.e. from the OL state to a word) are only allowed **once** in a legal solution of the semantic transfer task. Other particular transitions are also not allowed, as for example the one from an "out right" position (OR) back to an "out left" one (OL), as it is not possible to restart the role tagging process when it has been already accomplished on the left. The remaining transitions are all allowed several times. The transition probability can be thus defined as follows:

$$P\big(\theta(i) \mid \theta(i-1)\big) = \frac{\frac{C^b_{i,i\text{-}1}}{C^b}}{\frac{C_{i\text{-}1}}{C^b+1}} = \frac{C^b_{i,i\text{-}1}}{C_{i\text{-}1}} \, \frac{C^b + 1}{C^b} \tag{12}$$

	$\mathbf{e_{i+1}}$	$\mathbf{e_{OL}}$	$\mathbf{e_{OR}}$
$\mathbf{e_i}$	+	0	1
$\mathbf{e_{OL}}$	1	+	0
$\mathbf{e_{OR}}$	0	0	+

Table 2. Transition probabilities between states. States $\mathbf{e}_i$ (with $i > 0$) characterize transitions from two pairs or Italian words both internal (i.e. members of the sequence) to $is(\alpha)$.

where the notation $C^b_{i,i\text{-}1} = C\big(\theta(i) \mid \theta(i-1)\big)$ is used for bigrams and the notation $C_{i\text{-}1} = C\big(\theta(i-1)\big)$ for unigrams.
respectively. The counts used in the estimates of Eq. 12 are summarized in Table 3.

	$\mathbf{e}_{i+1}$	$\mathbf{e_{OL}}$	$\mathbf{e_{OR}}$
$\mathbf{e_i}$	$C^b_{i,i+1}$	na	$C^b_{i,\mathrm{OR}}$
$\mathbf{e_{OL}}$	$C^b_{\mathrm{OL},i+1}$	$C^b_{\mathrm{OL,OL}}$	na
$\mathbf{e_{OR}}$	na	na	$C^b_{\mathrm{OR,OR}}$

Table 3. Counts involved in the different transition probabilities.

3 Evaluation

In this section the Markov model for the semantic transfer will be evaluated over an English-Italian bilingual parallel corpus. The sentences used for this purpose have been also employed in [10,12]. As reported in Table 5, it consists in a set of 984 sentences split into 788 training sentences and the remaining 196 ones used for testing. The bilingual corpus is an excerpt of the European parliament data [13], available online. More precisely the about 200 sentences employed in testing where annotated in English and Italian according to their own predicates and semantic roles. The sentences do not share all their annotations as they have been manually labeled according to different FrameNet versions. In Table 4 the number of semantic roles manually annotated in both the English and Italian sentences are shown. Basically, all the LUs (i.e. target predicates) are shared between the two corpus, while only half of the frame elements use the same labels. The statistical data used to build up the model are supplied by Giza and computed over the sentences used as the training corpus.

The emission probabilities are computed by Eq. 8 and can be divided in two main classes. The first one is the probability of the translation of an Italian word w_i into an English word e_k, that is part of the known targeted semantic role, $es(\alpha)$. It is estimated as in Eq. 9 in terms of the Giza probabilities. The second

Semantic Roles	English corpus	Italian corpus	In common
Lexical Units	998	984	984
Frame Elements	1713	1661	842
Total	2711	2645	1826

Table 4. Semantic roles in the bilingual corpus. The roles in common between English and Italian corpus are those for which labels are identical.

one is the probability of an Italian word w_i to be part of segments that are outside $es(\alpha)$. It is computed according to Eq. **??** by estimating counts over the training corpus, whereas observations about Italian words occurring on the left or on the right of a semantic role could be collected.

	Sentences	Semantic Roles	Targets	Frame Elements
Training	788	1438	788	650
Testing	196	388	196	192

Table 5. The training/test set splitting adopted for the training of the HMM.

The transition probabilities are not lexicalized and can be thus computed for every kind of transition within those depicted in Table 2. The model described allows only a unique solution, that is a set of one or more contiguous Italian words expressing a semantic role, highlighted between the two labels (OL) *out left* and (OR) *out right*. The system is evaluated according to the usual metrics of precision, recall and F-measure as computed over the involved semantic transfer information. First the partial or perfect matching of the individual role segments is computed: an output segment is partially (or fully) detected if it has a partial (or perfect) overlap with the corresponding segment defined in the oracle. Percentage is obtained as the ratio with respect the set of all targeted segments. Token-based measures are also considered. Token-recall or precision are obtained considering as individual decisions the tokens belonging to the targeted roles. A(n Italian) token labeled by a semantic role α is a true positive iff it is part of the segment for α as defined in the oracle. Similarly, tokens are considered false negatives and positives if they belong only to the oracle or only to the system output. In Table 6 the overall system results are reported. The percentages are referred to the targets and to the semantic role (or FEs in the FrameNet jargon). Baselines and previous work can be described according to the example shown in Figure 5.

The upper part of Fig. 5 represents the word-level alignment as proposed by the Giza tool. The baselines are reported in the bottom part (B) of the figure. The first alignment derives from the Moses alignment ([14]): it select among the partial segments suggested by the Moses phrase-translation tables the max-

Model	Perfect Matching (FE only)	Partial Matching (FE only)	Token Precision (FE only)	Token Recall (FE only)	Token F1 (FE only)
baseline	66.88% (28,37%)	71.78% (41,13%)	.7 (.59)	.31 (.14)	.4 (.23)
Cicling09	59% (45.3%)	**80,6**% (**81**%)	.80 (.80)	**.86** (**.87**)	.83 (.84)
HMM system	**60.3**% (**56.7**%)	78,8% (80.2%)	**.86** (**.87**)	.85 (.86)	**.86** (**.87**)

Table 6. Accuracy of the role alignment task over the Gold Standard

imal segment (i.e. the longest translation of tokens internal to the targeted role $es(\alpha)$). The row named Cicling09 reports the results obtained by the system discussed in [10] over the test set adopted in this work. That system takes as input the Moses phrase-translation tables. It then performs a boundary detection phase, where possibly useful translation subsequences are merged: all the collected Italian segments are here processed and the best boundary is selected. Pruning of some unsuited solutions is then obtained through a post-processing phase. Here the computed boundaries are refined by applying heuristics based on the entire sentence, i.e. according to candidate solutions for all the different semantic roles in a sentence. In Figure 5 the comparison of the three semantic

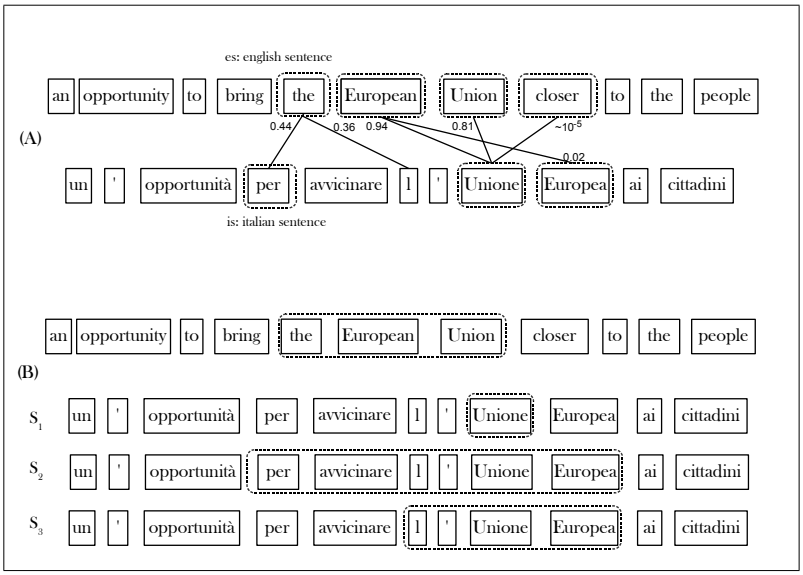

Fig. 5. Comparison among three semantic transfer methods based on the Moses alignments shown in A). The results over the argument "*the European Union*" are shown in the last row of the B) part, as compared with the Moses baseline and the Cicling09 system (first and second row in B), respectively).

transfer methods is presented in the last three rows of the B) part. In the example a role α (e.g. THEME) is characterized by $es(\alpha)$=["*the European Union*"].

The third one is the one proposed in this study. As we can see the Moses baseline method, i.e. the second one, suggests the maximum boundary among those proposed by the alignment shown in A). The Cicling09 system, i.e. the first one, refines this result by applying some heuristics. Although more precise, it does not retrieve all the tokens. The HMM system retrieves all tokens without relying on post processing. The Viterbi algorithm in fact is applied for each role to the entire sentence. The resulting best path through the trellis states receive all the syntagmatic constraints from the available translations and from the transitions that characterize the closed labeling as shown in Fig. 4. This method is thus more robust achieving higher precision scores without post processing.

Results in Table 6 show that the HHM defined in this paper produces an improvement in precision and recall over the previously proposed methods. Although the percentage of partially matched roles is in line with the Cicling09 system (i.e. 78.8% vs.80.2%), the perfectly matched phenomena are many more. At the Frame Element level (i.e. over the set of most complex phenomena) a striking performance increase is obtained, that raise accuracy from 45% to 56% (i.e. about 25% increment). Results are also very good for what concerns token-based measures, this suggesting that the HMM-based model approximate in a much more accurate way the perfect labeling. It is interesting to note that the best Cicling09 result reaches a good level of token recall (i.e. about 86%) at the expense of a lower precision (80%). As a matter of fact the precision reachable by the HMM-based system is higher (about 87%) with a corresponding 5% increase also in the token F1 measure. In Table 6, results in brackets are achieved on the set of frame elements that are more complex as for their length (target predicates are usually expressed by one-token segments, e.g. simple verbs or nouns) and for their grammatical structure. On this phenomena the HMM-based system is almost always better with more stable and precise labeling. In general the HMM-based system is more robust and independent from complex heuristics that characterize instead previous works.

4 Conclusion

Unsupervised models for semantic role transfer in bilingual corpora have been recently presented in [10]. Improving these models means making the boundary detection algorithms more robust and introducing new grammatical and syntactic rules. However, this research line may also lead to weaker models that may be not fully applicable to real cases. In this work, an Hidden Markov Model is introduced in order to increase robustness and generalize the semantic transfer system to a larger set of phenomena. First of all, the model should not depend too much on the language pair, in order for it to be adopted in a larger set of cases (i.e. generic semantic role transfer tasks between any language pair and aligned corpus). The model strictly relies on the Giza statistical alignment capabilities and from robust emission and transition probability estimates over a small corpus. Each sentence is mapped into his own model where semantic roles in the target language are states and the source roles are the observations.

Models are just solved at a statistical level (i.e. multiple applications of Viterbi decoding, one for each role to be detected): no rule-based boundary detection or post processing is applied.

The proposed supervised model has been shown to be trainable using a small corpus. In this paper, on the set of 984 sentences taken from European Parliament corpus, an 80% was used for the training phase. Results obtained on the remaining 20% of the sentences allowed to compare the proposed HMM-based approach with the unsupervised system described in [10]. The increase in token-based precision confirms the superiority of the new method. Although they are relative to a subset of the European Parliament used for the evaluation in [10], they are representative of a large set of lexical and grammatical phenomena. Wider experimentation is already in progress to confirm these results over the entire European Parliament corpus: the performance of a Semantic Role Labeling system will be used as an indirect measure of the quality reachable by the approach here proposed.

References

1. Fillmore, C.J.: Frames and the semantics of understanding. Quaderni di Semantica **4** (1985) 222–254
2. Baker, C.F., Fillmore, C.J., Lowe, J.B.: The Berkeley FrameNet project. In: Proc. of COLING-ACL '98. (1998) 86–90
3. Pado, S.: Cross-lingual annotation projection models for role-semantic information. In: PhD Thesis Dissertation, University of Saarlandes, Saarbrucken, Germany (2007)
4. Yarowsky, D., Ngai, G., Wicentowski, R.: Inducing multilingual text analysis tools via robust projection across aligned corpora. In: HLT '01: Proceedings of the first international conference on Human language technology research, Morristown, NJ, USA, Association for Computational Linguistics (2001) 1–8
5. Hwa, R., Resnik, P., Weinberg, A., Kolak, O.: Evaluating translational correspondence using annotation projection. In: In Proceedings of the 40th Annual Meeting of the ACL. (2002) 392–399
6. Smith, D.A., Smith, N.A.: Bilingual parsing with factored estimation: Using english to parse korean". In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. (2004) 49–54
7. Fung, P., Chen, B.: Biframenet: Bilingual frame semantics resource construction by cross-lingual induction. (2004) 931–937
8. Pado, S., Pitel, G.: Annotation precise du francais en semantique de roles par projection cross-linguistique. In: Proc. of TALN 2007, Toulouse, France (2007)
9. Tonelli, S., Pianta, E.: Frame information transfer from english to italian. In: Proc. of LREC Conference, Marrakech, Marocco (2008)
10. Basili, R., Cao, D.D., Croce, D., Coppola, B., Moschitti, A.: Cross-language frame semantics transfer in bilingual corpora. In: Proc. of 10th Int. Conf. on Intelligent Text Processing and Computational Linguistics (CICLing 2009), Mexico City, Mexico (2009)
11. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Computational Linguistics **29** (2003) 19–51

12. Tonelli, S., Pianta, E.: Three issues in cross-language frame information transfer. In: Proceedings of the RANLP 2009. (2009)
13. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: Proc. of the MT Summit, Phuket, Thailand (2005)
14. 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)