EvalIta 2011: the *Frame Labeling over Italian Texts* Task

Roberto Basili*, Diego De Cao*, Alessandro Lenci†,
Alessandro Moschitti‡, and Giulia Venturi∨

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

†University of Pisa, Italy
alessando.lenci@ling.unipi.it

‡University of Trento, Italy
moschitti@disi.unitn.it

∨ILC-CNR, Scuola St. Anna, Pisa, Italy
giulia.venturi@ilc.cnr.it

Abstract. The FLaIT task held within the EvalIta 2011 challenge is here described. Systems were asked to label semantic frames and their arguments as evoked by input predicate words over plain text sentences. Proposed systems are based on a variety of learning techniques and achieve very good results, over 80% of accuracy, in most subtasks.

Keywords: System Evaluation, Shallow Semantic Parsing, Frame Semantics

1 The Frame Labeling Task

In the "Frame Labeling over Italian Texts" (FLaIT) task, the general goal is to come forward with representation models, inductive algorithms and inference methods which address the Semantic Role Labeling (SRL) problem. This is the first time that such a task is proposed in the framework of the EVALITA campaign.

So far, a number of shared tasks (CoNLL–2004, 2005, 2008, 2009 and Senseval/Semeval–2004, 2007, 2010) have been concerned with SRL. Typically, two main English corpora with semantic annotations from which to train SRL systems have been used: PropBank [3][1] and FrameNet [1][2]. These previous experiences have been focused on developing SRL systems based on partial parsing information and/or increasing the amount of syntactic and semantic input information, aiming to boost the performance of machine learning systems on the SRL task.

Since 2009, CoNLL has been accompanied by a shared task dedicated to SRL not restricted to a monolingual setting (i.e. English) [2]. The Evalita 2011 FLaIT challenge is the first evaluation exercise for the Italian language, focusing on different aspects of the SRL task over Italian texts. The interest in organizing

[1] http://verbs.colorado.edu/∼mpalmer/projects/ace.html
[2] https://framenet.icsi.berkeley.edu/fndrupal/

this challenge has been prompted by the recent development of FrameNet–like resources for Italian that are currently under development in the iFrame project[3].

1.1 Task Definition

In the task, the complete annotation of frame information for target predicate words marked in input sentences was requested. As an example in the sentence "*Rilevata la presenza di gas in uno dei tubi, i guardiani hanno fatto scattare il piano d'emergenza*", the two frames PRESENCE, as evoked by the LU *presenza*, and PROCESS_START, given the LU *scattare* should be labeled as in the following two separate lines:

Rilevata la presenza [*di gas*] ENTITY [*in uno dei tubi*] LOCATION,
i guardiani hanno fatto scattare, [*il piano d'emergenza*] EVENT

where arguments are typed in square brackets. FLaIT was organized into three subtasks:
Task 1: Frame Prediction (FP). In the first subtask, the assignment of the correct frame for a sentence, given a marked and possibly ambiguous lexical unit, was due. This aimed at verifying the ability in recognizing the true frame of an occurring predicate word, and to select it even against possibly ambiguous lexical units.
Task 2 Boundary Detection (BD) and Task 3 Argument Classification (AC) Participants have been asked to locate and annotate all the semantic arguments of a frame, which are explicitly realized in a sentence, given the marked lexical unit. This task corresponds to the traditional Semantic Role labeling challenge as defined by the CoNLL 2005 task.

1.2 Dataset Definition

The dataset used for **training** derives from the merging of two independently annotated resources. The first set, hereafter denoted as FBK has been developed at the Fondazione Bruno Kessler [4]. It includes the annotation of 605 sentences (605 predicates and 1074 roles) at the syntactic and semantic level under the XML Tiger format also used by the Salsa project. The reference syntactic formalism of the is a constituency-based formalism obtained as output from the constituency-based parser by [5]. The second set, hereafter ILC set, has been developed at the ILC in Pisa by Alessandro Lenci and his colleagues [6]. It consists of the ISST–TANL Corpus, a dependency-annotated corpus originating as a revision of a subset of the Italian Syntactic-Semantic Treebank or ISST [9], enriched with Semantic Frames under the XML Tiger format also used by the Salsa project[4]. These amount to 650 sentences with 1763 roles. The resulting

[3] http://sag.art.uniroma2.it/iframe/doku.php
[4] The ISST–TANL Corpus has been used for the Dependency Parsing Track of Evalita 2009 [7] and for the Domain Adaptation for Dependency Parsing Tarck of Evalita 2011 [8].

training set thus includes 1255 sentences for about 38 frames. The total amount of roles completely annotated correspond to 2837 arguments.

The **test set** has been obtained through the exploitation of the aligned English-Italian Europarl section [10]. The English Framenet lexicon has been first used to locate candidate sentences of each of the 38 frames already covered by the training dataset. Annotators completed the annotation of all boundaries and their corresponding Frame Elements, by removing possibly wrong or useless (e.g. too short) sentences. At the end of the labeling we gathered 318 sentences, again focusing on 36 of the training set frames, for a total of 318 targets and 560 other arguments. Notice that the above process was frame driven and not lexical unit driven so that in a not negligible set of cases (27 out of 318 sentences), the lexical unit of the test sentences was never observed in the training set. This had the beneficial effect to measure also the generalization power of the machine learning methods applied during training towards poorly (or never) observed phenomena.

Test and Runs. Given the structured nature of the FLaIT task, test data have been submitted in an incremental fashion, with a growing number of marked details. In the **first run**, sentences were only marked with the targeted lexical unit, but no frame information was provided in order to test the quality of the frame detection process. In the **second run**, the correct frame of the lexical unit was provided but no boundary information was made available in order to test the quality of the boundary detection task also in presence of gold information about the frame. Finally, in the **third run**, the systems were requested to annotate argument roles (i.e. Frame Elements) but exact boundary information was provided. Notice that, in every run, systems have been asked to perform all the three above tasks, i.e. FP, BD and AC. This allowed to evaluate the impact of early labeling errors on the quality of the later annotation steps.

1.3 Evaluation Measures

The traditional evaluation metrics of precision and recall have been used for the three tasks:

Frame Detection. Any (sentence,predicate) pair for which the correct frame is provided by the system is counted as a true positive.

Boundary Detection (BD). True positives here are those semantic arguments whose boundaries are precisely determined (i.e., all and only those tokens belonging to the argument are correctly detected). The average across the overall number of sentences is computed as the microaverage across all arguments. The token based version of this measure accounts for the number of individual tokens correctly classified instead of the number of exact arguments.

Argument Classification (AC). Arguments whose semantic role (i.e. Frame Element label) is correctly assigned are the true positives tp while false positives fp are arguments whose assignment does not correspond to the label in the oracle. Unlabeled arguments correspond to false negatives fn. As usual, AC precision is given by $tp/(tp + fp)$, while AC recall is $tp/(tp + fn)$. The average across the overall number of sentences is computed as the microaverage across

all arguments. The token based version of this measure accounts for the number of tokens correctly classified instead of the number of arguments. The **AC F1-measure** is the weighted harmonic mean of AC precision and AC recall

Table 1. Results of the Frame Detection task

Systems	CELI_NT	CELI_WT	TV_SVM-SPTK	TV_SVM-HMM
Gold Frame Total	318	318	318	318
Frame Correct	207	207	257	250
Frame Untagged	38	38	0	0
Frame Precision	73.93%	73.93%	**80.82%**	78.62%
Frame Recall	65.09%	65.09%	**80.82%**	78.62%
Frame F1	69.23%	69.23%	**80.82%**	78.62%

Table 2. Results of the Boundary Detection (BD) task

First Run					
Systems	CELI_NT	CELI_WT	TV_SVM-SPTK	TV_SVM-HMM	
Gold Arg. Size	560	560	560	560	
Gold Arg. Token Size	3492	3492	3492	3492	
Sys. Arg. Size	255	332	609	568	
Sys. Arg. Token Size	1165	1477	3592	3962	
Correct Bound.	117	135	406	288	
Correct Tk. Bound.	945	1162	2945	2695	
BD Prec.	45.88%	40.66%	**66.67%**	50.70%	
BD Rec.	20.89%	24.11%	**72.50%**	51.43%	
BD F1	28.71%	30.27%	**69.46%**	51.06%	
BD Token Prec.	81.12%	78.67%	**81.99%**	68.02%	
BD Token Rec	27.06%	33.28%	**84.34%**	77.18%	
BD Token F1	40.58%	46.77%	**83.15%**	72.31%	
Second Run					
Systems	CELI_NT	CELI_WT	TV_SVM-SPTK	TV_SVM-HMM	RTV_SVM_Geom
Sys. Arg. Size	263	349	609	565	494
Sys. Arg. Token Size	1150	1487	3592	3930	3569
Correct Bound.	124	148	406	282	357
Correct Token Bound.	949	1193	2945	2678	2969
BD Prec.	47.15%	42.41%	66.67%	49.91%	**72.27%**
BD Rec.	22.14%	26.43%	**72.50%**	50.36%	63.75%
BD F1	30.13%	32.56%	**69.46%**	50.13%	67.74%
BD Token Prec.	82.52%	80.23%	81.99%	68.14%	**83.19%**
BD Token Rec.	27.18%	34.16%	84.34%	76.69%	**85.02%**
BD Token F1	40.89%	47.92%	83.15%	72.16%	**84.10%**

2 Results

The participating teams refer to two different institutions: CELI and the University of Roma, Tor Vergata. Their systems are described elsewhere in these proceedings, and will be hereafter shortly outlined.

The FLaIT CELI System. This system applied a legacy parser ([11]) to the input sentences and relied upon a combination of dependency based rules (such as subcategorization patterns) and machine learning techniques, based on Markov Logic Networks (MLN). The authors developed an early version of their Frame Labeling and Boundary detection subsystems just for the FLaIT challenge. Two systems are presented. The first (i.e CELI_WT) makes use of hand coded rules for Semantic Role Labeling, while the second (CELI_NT) only relies on learned rules.

Structured Learning SRL system by the University of Roma, Tor Vergata. These two systems are strongly based on the notion of structured learning as realized by SVM learning. In both cases a discriminative approach is applied but structures are accounted for in the first system, TV_SVM_SPTK, through the adoption of syntagmatic (i.e. tree) kernels. $SPTK$ is a model that extends the standard tree kernels formulation by embedding a corpus-driven lexical similarity metrics between terminal nodes (i.e. words in the leaves) [12]. The second system, named TV_SVM_HMM is a combination of discriminative and generative model often referred to as SVM_HMM. It is also interesting as it maps the BD and AC task into a labeling task, without resorting to any information about grammatical dependencies and the parse tree.

The Semi-Supervised SRL system by the University of Roma, Tor Vergata. The second team in Roma Tor Vergata, made use of an hybrid architecture for just the BD and AC tasks. The first BD component makes use of an SVM-based learning model based on manually engineered features derived from the sentence dependency tree. In the second AC step, a simple generative model is extended with probability estimators based on a distributional semantic, i.e. geometrical, method, that optimizes against small training sets. The RTV_SVM_Geom system is based on the work discussed in [13].

2.1 Discussion

Results for the Frame Detection task are reported in Table 1. The top scores are fairly high (F1>80%), because of the relatively small number of frames to be identified and of the "closed world" assumption of this task. Since the target was overtly marked in the test corpus and systems had to choose the correct frame to be assigned among those attested for that lexical unit in the training corpus the overall task was relatively easy. This is also confirmed by the rather high baseline score (68.39%) that can be simply achieved by randomly assigning one of the possible (according to training data) candidate frame to the target. While the four systems achieve rather close precision values, significant differences exist in recall. This was expected, given the approach of the CELI team to maximize precision over recall.

Table 3. Results of the Argument Classification (AC) task

First Run					
Systems	CELI_NT	CELI_WT	TV_SVM-SPTK	TV_SVM-HMM	
Gold Arg. Size	560	560	560	560	
Gold Arg. Token Size	3492	3492	3492	3492	
Sys. Arg. Size	255	332	609	568	
Sys. Arg. Token Size	1165	1477	3592	3962	
Correct Arg.	83	91	295	188	
Correct Token Arg.	558	731	2248	1853	
AC Prec.	32.55%	27.41%	**48.44%**	33.10%	
AC Rec.	14.82%	16.25%	**52.68%**	33.57%	
AC F1	20.37%	20.40%	**50.47%**	33.33%	
AC Token Prec.	47.90%	49.49%	**62.58%**	46.77%	
AC Token Rec	15.98%	20.93%	**64.38%**	53.06%	
AC Token F1	23.96%	29.42%	**63.47%**	49.72%	
Second Run					
Systems	CELI_NT	CELI_WT	TV_SVM-SPTK	TV_SVM-HMM	RTV_SVM_Geom
Sys. Arg. Size	263	349	609	565	494
Sys. Arg. Token Size	1150	1487	3592	3930	3569
Correct Arg.	95	109	312	212	256
Correct Token Arg.	716	960	2479	2147	2198
AC Prec.	36.12%	31.23%	51.23%	37.52%	**51.82%**
AC Rec.	16.96%	19.46%	**55.71%**	37.86%	45.71%
AC F1	23.09%	23.98%	**53.38%**	37.69%	48.58%
AC Token Prec.	62.26%	64.56%	**69.01%**	54.63%	61.59%
AC Token Rec.	20.50%	27.49%	**70.99%**	61.48%	62.94%
AC Token F1	30.85%	38.56%	**69.99%**	57.86%	62.26%
Third Run					
Systems	CELI_NT	CELI_WT	TV_SVM-SPTK	TV_SVM-HMM	RTV_SVM_Geom
Sys. Arg. Size	247	300	560	549	543
Sys. Arg. Token Size	1657	2160	3492	3481	3475
Correct Arg.	181	225	394	366	363
Correct Token Arg.	1269	1798	2736	2705	2489
AC Prec.	73.28%	**75.00%**	70.36%	66.67%	66.85%
AC Rec.	32.32%	40.18%	**70.36%**	65.36%	64.82%
AC F1	44.86%	52.33%	**70.36%**	66.01%	65.82%
AC Token Prec.	76.58%	**83.24%**	78.35%	77.71%	71.63%
AC Token Rec.	36.34%	51.49%	**78.35%**	77.46%	71.28%
AC Token F1	49.29%	63.62%	**78.35%**	77.59%	71.45%

Moving to the Boundary Detection task, we can see in Table 1 that the differences between the first and the second run results do not appear to be significant. This means that knowing the frame evoked by the target does not help systems in identifying the boundaries of its Frame Element. This is indeed predictable, since the Frame Element spans do not seem to be related to the

particular type of Frame. The RTV_SVM_Geom, which did not participate in the first run, achieves the best precision, but TV_SVM-SPTK shows up again as the best model, given its better tradeoff between precision and recall.

Knowing the frame does not facilitate systems in the AC task either. As can be seen from the results in Table 3, there is just a minor improvement in the second run, with respect to the first one. Conversely, all the systems significantly improve their performance in the third run. The frame type and the Frame Element boundaries are crucial information to boost system ability to assign the proper role. In this run, CELI_WT scores the highest precision, but at the cost of a rather low recall whereas TV_SVM-SPTK again achieves the best tradeoff between the two measures.

3 Conclusions

The first experience with the FLaIT task at EVALITA has been successfull. The participation of two research centers with 5 systems is very good if we consider the complexity of designing an SRL chain and making it operational. A lexicon of 105 different lexical units for 36 frames has been made available by the challenge.

The competition resulted in a variety of advanced methods ranging from dependency rules to probabilistic and discriminative methods (e.g. semantically smoothed tree kernels). The obtained accuracy is generally good and in line with the state-of-the-art in other languages such as English, for which larger and richer resources are available. The realistic settings adopted (i.e. no gold information was provided for all steps) make the presented results even more valuable, as they have been derived in standard operational conditions, few annotated data and lack of lexical *ad hoc* resources.

Acknowledgement. We would like to thank all the members of the iFrame group, who greatly supported the FLaIT 2011 experience. In particular, Sara Tonelli and Emanuele Pianta for making their annotated data available to all teams. This work has been partially funded by the Italian Project, PRIN 2008: *Portale per l'Accesso alle Risorse Linguistiche per l'Italiano* (PARLI).

References

1. Baker, C.F., Fillmore, C.J., Lowe, J.B.: The Berkeley FrameNet Project. In: Proceedings of the 36th ACL Meeting and 17th ICCL Conference, Morgan Kaufmann (1998)
2. Hajič, J., Ciaramita, M., Johansson, R., Kawahara, D., Martí, M.A., Màrquez, L., Meyers, A., Nivre, J., Padó, S., Štěpánek, J., Straňák, P., Surdeanu, M., Xue, N., Zhang, Y.: The CoNLL-2009 Shared Task: Syntactic and Semantic Dependencies in Multiple Languages. In: Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL): Shared Task, Boulder, Colorado, June, pp. 1-18 (2009)
3. Palmer, M., Gildea, D., Kingsbury, P.: The Proposition Bank: A Corpus Annotated with Semantic Roles. In: Computational Linguistics Journal, 31(1) (2005)

4. Tonelli, S., Pianta, E.: Frame information transfer from english to italian. In: Proc. of LREC Conference, Marrakech, Marocco (2008)
5. Corazza, A., Lavelli, A., Satta, G.: Phrase-based statistical parsing. In: Proc. of EVALITA 2007 Workshop on Evaluation of NLP Tools for Italian, AI*IA (2007)
6. Lenci, A., Montemagni, S., Vecchi, E., Venturi, G.: Enriching the isst-tanl corpus with semantic frames. In: Forthcoming
7. Bosco, C., Montemagni, S., Mazzei, A., Lombardo, V., Dell'Orletta, F., Lenci, A.: Parsing Task: comparing dependency parsers and treebanks. In: Proceedings of Evalita'09, Reggio Emilia (2009)
8. Dell'Orletta, F., Marchi, S., Montemagni, S., Venturi, G., Agnoloni, T., Francesconi, E.: Domain Adaptation for Dependency Parsing at Evalita 2011. In: Proceedings of Evalita'11 (2011)
9. Montemagni, S., Barsotti, F., Battista, M., Calzolari, N., Corazzari, O., Lenci, A., Zampolli, A., Fanciulli, F., Massetani, M., Raffaelli, R., Basili, R., Pazienza, M. T., Saracino, Zanzotto, F., Mana, N., Pianesi, F., Delmonte, R.: Building the *Italian S*yntactic–*S*emantic *T*reebank. In: Anne Abeillé (ed.), "Building and Using syntactically annotated corpora", Kluwer, Dordrecht (2003)
10. Basili, R., De Cao, 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. Testa, M., Bolioli, A., Dini, L., Mazzini, G.: Evaluation of a Semantically Oriented Dependency Grammar for Italian. In: Proc. of the EVALITA 2009 (2009)
12. Croce, D., Moschitti, A., Basili, R.: Structured Lexical Similarity via Convolution Kernels on Dependency Trees. In: Proc. of the 2011 Conf. on Empirical Methods in Natural Language Processing, Edinburgh, UK (2011)
13. Croce, D., Giannone, C., Annesi P., Basili, R.: Towards open-domain semantic role labeling. In: Proc. of the 48th Annual Meeting of the ACL, Uppsala, Sweden, (2010)