Bootstrapping Enhanced Universal Dependencies for Italian

Maria Simi
Dipartimento di Informatica
Università di Pisa
Largo B. Pontecorvo 3, Pisa
simi@di.unipi.it

Simonetta Montemagni
Istituto di Linguistica Computazionale
"A. Zampolli" - CNR
Via Moruzzi 1, Pisa
simonetta.montemagni@ilc.cnr.it

Abstract

English. The paper presents an extension of the Italian Universal Dependencies Treebank with an "enhanced" representation level (e-IUDT), aimed at simplifying the information extraction process. The modules developed to semi-automatically build e-IUDT were delexicalized to perform cross-language enhancements: preliminary experiments in this direction led to promising results.

Italiano. *L'articolo presenta l'estensione della Universal Dependencies Treebank italiana (e-IUDT) con un livello di rappresentazione arricchito ("enhanced"), finalizzato a rendere più efficiente ed efficace il processo di estrazione dell'informazione. I moduli sviluppati per la costruzione semi-automatica della risorsa sono stati delessicalizzati e utilizzati per il trattamento di diverse lingue: esperimenti preliminari in questa direzione mostrano risultati promettenti.*

1 Introduction

The Universal Dependencies (UD) project, launched in 2015, aims at developing cross-linguistically consistent treebank annotation for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning, and parsing research from a language typology perspective (Nivre et al., 2016). UD represents an open community effort with over 200 contributors producing more than 100 treebanks in over 60 languages.
Starting from the Stanford Dependencies project, from which Universal Dependencies (UD) originate, two syntactic representation options are made available, suited to different use cases (De Marneffe and Manning, 2008): the so-called "basic" representation where a close parallelism to the source text is maintained (i.e. where each word of the original sentence is present as a node), and the so-called "collapsed and propagated" representation which was conceived with a specific view to information extraction tasks.
Within the current version of UD, the "collapsed and propagated" representation has evolved into the graph-based *enhanced* representation proposed by Schuster and Manning (2016).
Since UD version 2.2 (officially released on July 2018), "enhanced treebanks" started to appear for a limited number of languages, i.e. English, Finnish, Russian, Polish, Dutch, Latvian. In order to foster the development of enhanced treebanks for other languages, transfer experiments exploiting existing treebanks are reported in the literature, following both rule-based (Schuster and Manning 2016) and data-driven (Nyblom *et al.*, 2013) approaches.
This paper describes the approach we used for developing and validating the enhanced version of the Italian UD Treebank and reports the first results of transfer experiments to English.

2 Enhanced dependencies

Enhanced dependencies were proposed as a way to simplify the process of information extraction. Enhancements, for the most part, result in additional links added to the dependency tree, motivated by inferences, which remain however anchored at the surface representation level. The result of enhancing a dependency tree is a graph, possibly with cycles, but not necessarily a super graph (since some of the original arcs may be discarded).
The current UD guidelines are quite conservative, i.e. they suggest practically feasible enhancements only. Despite this, enhancements cannot always be achieved automatically, and the task is challenging enough to be interesting. Ac-

cording to the guidelines *enhanced graphs* may contain some or all of the following enhancements, described with particular emphasis on Italian:

1. Added subject relations in control and raising constructions;
2. Shared heads and dependents in coordination;
3. Co-reference in relative clause constructions;
4. Modifier specialization by means of case markers;
5. Null nodes for elided predicates.

2.1 Added subject relations

In the case of control and raising constructions, the subject of the subordinated non-finite clause is added. Consider the following examples, with controlled and raised subjects marked in bold:

1) Subject control: *La **mamma** ha promesso a Maria di comprare il pane* 'The **mother** promised Maria to buy the bread'
2) Object control: *La mamma ha convinto **Maria** a comprare il pane* 'The mother convinced **Maria** to buy the bread'
3) Oblique control: *La mamma ha chiesto a **Maria** di comprare il pane* 'The mother asked **Maria** to buy the bread'
4) Subject raising: *La **mamma** sembra apprezzare il pane integrale* 'The **mother** seems to like whole bread'

Figure 1 shows the UD representation of sentence 3), where the added subject relation (marked as `nsubj:xsubj`) is represented as an "enhanced arc" (in blue).

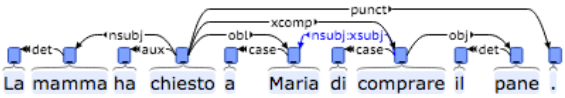
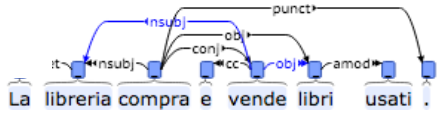


Figure 1. Enhanced representation of oblique control

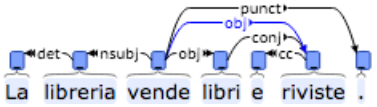
Control and raising predicates are superficially very similar, with a main difference: whereas Raising predicates have a 'non-thematic' argument, all arguments of Control predicates are 'thematic'. Such a distinction is neutralized in the enhanced UD representation. In both cases, however, the selection of the controlled/raised argument is lexically-driven.

2.2 Sharing in coordination

Coordination is another major source of potential enhancements, as information shared among conjuncts is typically attached only to the first conjunct and could be propagated to the other conjuncts, where this is applicable. In propagating information, it is useful to distinguish two cases, according to whether *dependents* of the first conjunct are propagated or the *head* of the first conjunct is propagated instead. Figure 2 shows Italian examples for each case.



The book store buys and sells used books.

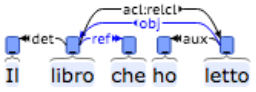


The book store sells books and magazines

Figure 2. a) Dependents propagation b) Head propagation

2.3 Co-reference in relative clauses

In basic UD, relative pronouns are normally attached to the main predicate of the relative clause, typically as nominal subjects (`nsubj`) or direct objects (`obj`). In the corresponding *enhanced* graph, the relative pronoun is linked to its antecedent with the `ref` relation and its dependency to the head of the relative clause is transferred to the antecedent itself, as exemplified in Figure 3 where it can be observed that the resulting enhanced representation contains a cycle.



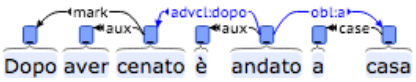
The book that I read

Figure 3. Relative clauses

2.4 Specialization of relations

Adding case information to the relation name of non-core dependents serves the purpose of disambiguating their semantic role. This information is expressed in terms of the preposition or the subordinating conjunction introducing non-core dependents. In particular: `nmod` and `obl` relation labels, respectively marking nominal and oblique modifiers introduced by prepositions, are augmented with language specific case information; `acl` and `advcl` labels, corresponding respectively to noun modifying clauses and adverbial clauses, are augmented with markers introducing them. A similar type of specialization also applies to the `conj` dependency label linking conjuncts in coordinated structures, which is specialized with respect to the conjunction type (*e*, *o*, *oppure* …), as identified by the lemma of the `cc` dependency (i.e. the relation between a

conjunct and a preceding coordinating conjunction).

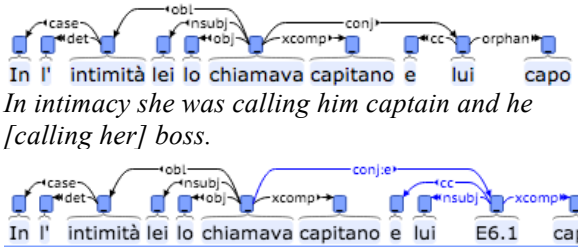


After having dinner he went home

Figure 4. Adding case and mark information to labels

2.5 Null nodes for elided predicates

Special null nodes are added in clauses to stand for a predicate which is elided; other cases of ellipsis are not being dealt with in the current UD guidelines due to major difficulties in their reconstruction. This type of enhancement occurs when the basic (i.e. pre-enhancement) tree contains an orphan relation which in the enhanced graph is removed and replaced by the reconstructed explicit syntactic structure. A new null node is added in place of the missing predicate and dependencies are redirected. Figure 5 shows an example of predicate elision, along with the enhanced version which introduces a new node (labeled as E6.1) obtained as a copy of the token 'chiamava'.



In intimacy she was calling him captain and he [calling her] boss.

Figure 5. Null nodes for elided predicates

This is the most problematic among the foreseen UD enhancements, due to several reasons such as: correct insertion points are difficult to anticipate; phraseological verbs and verbs with clitics (either in pronominal form or with clitic complements, see example in Figure 5) would require copying a variable number of tokens (the verb and the object with a shift in gender in the case at hand), which is not always easy to be identified; the appropriate syntactic role of the dependents of the added (i.e. recovered) predicate must be inferred by proper alignment with the dependents of the originally explicit predicate. Moreover, the proposed UD treatment requires a major change in the treebank format with the addition of new tokens with special labeling and numbering. Therefore, the introduction of null nodes calls for an *ad hoc* treatment and introduces a complexity in the processing of the treebank which is not fully justified if the aim is only to address the cases of predicate elision, for the fact that this is a rare phenomenon in treebanks. Other cases of elision, such as subject elision, are much more meaningful for Italian.

2.6 Open issues

Besides the standard enhancements foreseen for UD illustrated above, we are currently evaluating cases that could be treated as such for Italian, and could possibly be relevant for other languages as well. These include:

- case information, which could also be added for some core relations such as `ccomp`. Consider as an example the following sentences: *Non so se verrà domani* 'I don't know whether (he) will come tomorrow' vs *Non so quando arriverà* 'I don't know when (he) will arrive'. Without enhancing the `ccomp` relation, the semantics of the subordinated clause (conditional vs temporal) remains underspecified;
- null nodes for elided subjects: Italian is a pro-drop language and the omission of explicit subjects occurs quite frequently in actual language usage; according to Bates (1976), the pro-drop rate by adults is 70%. The addition of null nodes for subject ellipsis could significantly enhance the syntactic representation with a view to information extraction tasks.

The typology of representation enhancements could also be further extended to neutralize diathesis alternations, as proposed by Candito *et al.* (2017) for French. In what follows, we focus on the standard UD enhancements, excluding the treatment of predicate elision for which more careful investigation and detailed guidelines are required.

Table 1. Guessing step: additional annotations

ExtraSubjOf=*id*	token *id* is head of a new arc to be added to current token
RefOf=*id*	
PropagateDepTo=*id*	
PropagateHeadWith=*label*	*label* is the string suggested to propagate or to specialize a relation
CaseSpec=*label*	
MarkSpec=*label*	
CcSpec=*label*	

3 Developing an enhanced UD gold treebank for Italian

UD enhanced representation cannot be generated through a completely automatic process: this is a task that entails a global vision of the tree to be completed and often requires additional linguistic knowledge concerning e.g. raising/control

properties and/or selectional preferences of predicates. To build the enhanced Italian UD Treebank (henceforth, e-IUDT), we followed a three-step approach, articulated as follows:

1. *Guessing*: by making use of heuristics, a script suggests target nodes whose representation might be enhanced, e.g. the best extra subject candidate(s) in raising/control constructions, or the heads/dependents to be propagated in coordinated constructions. During this step, additional annotations are produced in the representation of involved tokens. For example, the annotation ExtraSubjOf = *j* added to token *i* is an indication that *i* is an additional subject headed by *j*. In other cases, the additional annotation indicates a label to be used for specializing a given relation or whether a conjunct should be propagated. Table 1 summarizes the additional annotations used;
2. *Revising*: the human annotator is called to validate the proposed changes, automatically generated during the previous step;
3. *Enhancing*: validated additional annotations are used to automatically generate the enhanced UD representation. Enhancements are not limited to retyping or addition of dependencies; in some cases, they involve the reshaping of the dependency graph, and for this reason an automatic transformation reduces the chances of occasional errors.

The heuristics behind the guessing step make use of lexical resources extracted from the corpus itself: this is the case, for example, of lexical information on raising/control properties of predicates, guiding the identification of extra-subject candidates.

Following the three-step strategy sketched above, we built a gold standard e-IUDT resource on top of the development data set of the Italian UD treebank (Release 2.2), constituted by 11,908 tokens. In Table 2, the first two columns (headed by "IT DEV (GOLD)") summarize the enhancements contained in the developed resource, which involve 21,75% of the words. Most of them are represented by the specialization of modifiers and conjoining relations, immediately followed by head propagation, relative clauses and extra-subjects. Interestingly enough, it can be noticed that the distribution of enhancements remains quite similar across different subsets of the same language (e.g. the development vs test sets for Italian), whether manually revised (dev) or not (test), or for another language, English.

4 A language-independent rule-based UD enhancer

Different cross-lingual techniques have been developed for adding enhanced dependencies to existing UD treebanks, both rule-based (Schuster and Manning 2016) and data-driven (Nyblom *et al.*, 2013). The modularity of the approach proposed for e-IUDT construction created the prerequisites for reusing some of these components for implementing an UD enhancing module. In what follows, we report preliminary results achieved by transforming the heuristics of the *Guessing* module into language-independent ones. Instead of using language-specific lexical information on raising/control properties of verbs for identifying extra-subject candidates, following the general UD strategy we used the heuristic according to which the controlled / raised subject of the embedded clause follows the obliqueness hierarchy, i.e. it is the object of the next higher clause, if there is one, or else its subject. Such a strategy was extended to foresee also oblique complements as controlled / raised subjects. The output of the *Guessing* module is directly passed to the *Enhancing* component. In order to test effectiveness and generality of the approach we tested the rule-based language-independent enhancer on the Italian and English development sets, both available as gold datasets.

Table 2. Enhanced relations

	IT DEV (GOLD)		**IT TEST (SILVER)**		**EN DEV (GOLD)**		**EN TEST (GOLD)**	
words	11.908		10.417		25.150		17.658	
enhancements	2.590	21,75%	2.275	21,84%	4.255	16,92%	3.595	20,36%
xsubj	69	2,66%	69	3,03%	342	8.04%	251	6,98%
ref	127	4,90%	210	9,23%	111	2,61%	274	7,62%
conj specializations	322	12,4%	266	11,7%	810	19,03%	532	14,80%
dep propagation*	45	1,7%	36	1,6%	165	3,9%	103	2,87%
head propagation*	250	9,7%	230	10,1%	478	11,2%	413	11,49%
other specializations	1.777	68,6%	1.464	64,4%	2.349	55%	2.022	56,24%

For evaluation, we used an adaptation of the evaluation script used in the evaluation campaign EVALITA 2014 (Bosco *et al.*, 2014), which is based on a set of relations extracted from the enhanced graph and for each of them computes *Precision*, *Recall* and *F1*. The evaluation focused on enhanced relations, thus allowing to analyze the complexity of the task. Table 3 reports the results achieved with the following gold data sets: **IT-dev**, the development dataset from UD-ISDT 2.2, enhanced as described above; **EN-dev** and **EN-test**, the development and test English datasets from UD-EWT 2.2.

Table 3. Precision, recall and F1 for enhanced relations

	UAS			LAS		
	P	R	F1	P	R	F1
IT-dev	99,7	99,8	99,8	99,5	99,6	99,6
EN-dev	98,2	99,3	98,8	96,2	97,2	96,7
EN-test	99,2	99,0	99,0	97,8	97,6	97,6

Table 4. Recall and Precision for enhancement type

	IT-dev		EN-dev		EN-test	
	R	P	R	P	R	P
xsubj	92,7	98,4	100,0	99,4	99,6	99,0
ref	100,0	100,0	99,1	86,6	99,3	94,4
conj spec	99,7	100,0	98,2	94,9	97,9	97,6
other specs	99,9	100,0	97,0	96,7	98,2	98,1
propagation	97,8	95,7	97,1	97,3	95,5	98,2

For Italian, despite the de-lexicalization of the Guessing module, UAS and LAS results are quite high. Results are very high also when enhancement is carried out against different sets of the English UD Treebank. A qualitative error analysis was also performed. Table 4 details recall and precision achieved for the different types of enhancements, for both Italian and English.

The main sources of errors turned out to be:

- the identification of extra-subjects, performed on the basis of heuristics rather than lexical information. This is particularly true for Italian, for both P and R;
- the specialization of relations with case markers, which turned out to be particularly problematic for multi-word markers. This can be observed mainly for English, for which a different strategy is followed in their representation;
- dependent propagation in coordinated constructions, which is not always easy for both languages. For Italian, the interference with pro-drop subjects should also be considered;
- other problematic cases include non-homogenous conjuncts for which the propagation of dependents or heads cannot always be easily carried out.

An example follows where, without lexical information, the identification of extra subjects fails. Consider the sentence *I carri armati ... andavano a Budapest ... a spegnere i fuochi* 'The tanks ... went to Budapest ... to extinguish the fires'. In UD, the `obl` relation covers both lexically realized indirect objects and other oblique complements: however, without distinguishing between the two it is impossible to recover the extra subject of the infinitive clause. A suggestion could be to introduce a specialization of the `obl` relation for identifying indirect objects.

Dependency specialization turned out to be a challenging conversion case when applied to the English UD treebank: problems encountered were somehow unexpected, being mostly due to a different strategy for annotating multi-word case markers, not always compliant with the general UD annotation guidelines. This explains the lower results reported in Table 3 for English with respect to Italian.

5 Conclusions

We extended the Italian UD Treebank with an enhanced representation level: Italian is now among the few languages within UD with a gold enhanced Treebank which will be part of Release v2.3. The modules used to semi-automatically build e-IUDT were delexicalized to carry out cross-language enhancements: preliminary results for both Italian and English are promising. The contribution also includes better and more detailed specifications to the constantly in-progress guidelines. Current developments include: from a mono-lingual perspective, extension of the typology of enhancements; from the multi-lingual perspective, testing and extending the enhancement component successfully used with English for other languages.

References

Bates Elisabeth. 1976. *Language and context: The acquisition of pragmatics*. New York, NY: Academic Press.

Cristina Bosco, Vincenzo Lombardo, Leonardo Lesmo, Daniela Vassallo. 2000. Building a treebank for Italian: a data-driven annotation schema. In Proceedings of LREC 2000, Athens, Greece.

Cristina Bosco, Simonetta Montemagni, Maria Simi. 2012. Harmonization and Merging of two Italian

Dependency Treebanks, Workshop on Merging of Language Resources, in Proceedings of LREC 2012, Workshop on Language Resource Merging, Instanbul, May 2012, ELRA, pp. 23–30.

Cristina Bosco, Simonetta Montemagni, Maria Simi. 2013. Converting Italian Treebanks: Towards an Italian Stanford Dependency Treebank. In: *ACL Linguistic Annotation Workshop & Interoperability with Discourse*, Sofia, Bulgaria.

Cristina Bosco, Felice Dell'Orletta, Simonetta Montemagni, Manuela Sanguinetti, Maria Simi. 2014. The Evalita 2014 Dependency Parsing task, CLiC-it 2014 and EVALITA 2014 Proceedings, Pisa University Press, ISBN/EAN: 978-886741-472-7, 1–8.

Marie Candito, Bruno Guillaume, Guy Perrier, Djamé Seddah. 2017. Enhanced UD Dependencies with Neutralized Diathesis Alternation, *Depling 2017 - Fourth International Conference on Dependency Linguistics*, Sep 2017, Pisa, Italy. 2017

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The Stanford typed dependencies representation. In COLING Workshop on Cross-framework and Cross-domain Parser Evaluation.

Marie-Catherine de Marneffe, Miriam Connor, Natalia Silveira, Bowman S. R., Timothy Dozat, Christopher D. Manning. 2013. More constructions, more genres: Extending Stanford Dependencies, Proc. of the Second International Conference on Dependency Linguistics (DepLing 2013), Prague, August 27–30, Charles University in Prague, Matfyzpress, Prague, pp. 187–196.

Marie-Catherine de Marneffe and Christopher D. Manning. 2013. Stanford typed dependencies manual, September 2008, Revised for the Stanford Parser v. 3.3 in December 2013.

Marie-Catherine De Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, Christopher D. Manning. 2014. Universal Stanford Dependencies: a Cross-Linguistic Typology. In: *Proc. LREC 2014*, Reykjavik, Iceland, ELRA.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In Proceedings of LREC.

Simonetta Montemagni, Maria Simi. 2007. The Italian dependency annotated corpus developed for the CoNLL–2007 shared task. Technical report, ILC–CNR.

Jenna Nyblom, Samuel Kohonen, Katri Haverinen, Tapio Salakoski and Filip Ginter. 2013. Predicting conjunct propagation and other extended Stanford Dependencies. Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013), pp 252–261, Prague, August 27–30.

Maria Simi, Cristina Bosco, Simonetta Montemagni. 2008. Less is More? Towards a Reduced Inventory of Categories for Training a Parser for the Italian Stanford Dependencies. In: *Proc. LREC 2014*, 26–31, May, Reykjavik, Iceland, ELRA.

Schuster, Sebastian and Christopher D. Manning. Enhanced English Universal Dependencies: An Improved Representation for Natural Language Understanding Tasks." LREC (2016).