Enhancing Universal Dependency Treebanks: A Case Study

Joakim Nivre[*] **Paola Marongiu**[†] **Filip Ginter**[‡] **Jenna Kanerva**[‡]
Simonetta Montemagni[◇] **Sebastian Schuster**[⋆] **Maria Simi**[•]

[*]Uppsala University, Department of Linguistics and Philology
[†]University of Pavia, Department of Linguistics
[‡]University of Turku, Department of Future Technologies
[◇]Institute for Computational Linguistics «A. Zampolli» – CNR, Italy
[⋆]Stanford University, Department of Linguistics
[•]University of Pisa, Department of Computer Science

Abstract

We evaluate two cross-lingual techniques for adding enhanced dependencies to existing treebanks in Universal Dependencies. We apply a rule-based system developed for English and a data-driven system trained on Finnish to Swedish and Italian. We find that both systems are accurate enough to bootstrap enhanced dependencies in existing UD treebanks. In the case of Italian, results are even on par with those of a prototype language-specific system.

1 Introduction

Universal Dependencies (UD) is a framework for cross-linguistically consistent treebank annotation (Nivre et al., 2016). Its syntactic annotation layer exists in two versions: a *basic* representation, where words are connected by syntactic relations into a dependency tree, and an *enhanced* representation, which is a richer graph structure that adds external subject relations, shared dependents in coordination, and predicate-argument relations in elliptical constructions, among other things.

Despite the usefulness of enhanced representations (see e.g., Reddy et al. 2017; Schuster et al. 2017), most UD treebanks still contain only basic dependencies[1] and therefore cannot be used to train or evaluate systems that output enhanced UD graphs. In this paper, we explore cross-lingual methods for predicting enhanced dependencies given a basic dependencies treebank. If these predictions are accurate enough, they can be used as a first approximation of enhanced representations for the nearly 100 UD treebanks that lack them, and as input to manual validation. Further, enhanced UD graphs are in many respects very similar to semantic dependency representations that encode predicate-argument structures (e.g., Böhmová et al. 2003; Miyao and Tsujii 2004; Oepen and Lønning 2006). While the latter exist only for a small number of languages and are typically either produced by complex hand-written grammars or by manual annotation, basic UD treebanks currently exist for more than 60 languages. Hence, automatic methods capable of predicting enhanced dependencies from UD treebanks, have the potential to drastically increase the availability of semantic dependency treebanks.

In this paper, we evaluate a rule-based system developed for English and a data-driven system trained on de-lexicalized Finnish data, for predicting enhanced dependencies on a sample of 1,000 sentences in two new languages, namely Swedish and Italian. For Italian, we also compare to a rule-based system developed specifically for that language using language-specific information. The results show that both cross-lingual methods give high precision, often on par with the language-specific system, and that recall can be improved by exploiting their complementary strengths.

2 Basic and Enhanced Dependencies

Basic dependencies are strict surface syntax trees that connect content words with argument and modifier relations, and attach function words to the content word they modify (Figure 1). Enhanced dependencies restructure trees and add relations that have been shown useful for semantic downstream applications. Although the enhanced representation is in most cases a monotonic exten-

[1]Out of 102 treebanks in UD release v2.1, only 5 contain enhanced dependencies.

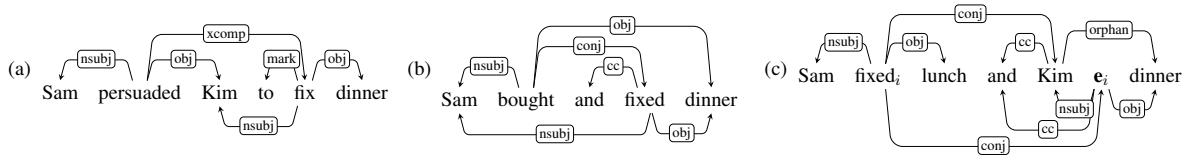

Figure 1: UD basic (top) and enhanced (bottom) dependencies: (a) control, (b) coordination, (c) gapping. For clarity, we show only those enhanced dependencies that are not shared with the basic layer.

sion of the basic one, this does not hold in general (as shown by the treatment of ellipsis below). The current UD guidelines define five enhancements:

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

The last two enhancements can in most cases be predicted deterministically from the basic representation and are mainly a practical convenience. We therefore limit our attention to the first three types, illustrated in Figure 1 (a–c).

Added subject relations Basic dependencies do not specify whether the implicit subject of *fix* in (a) is controlled by *Sam* (subject) or *Kim* (object), but enhanced dependencies do. Similar relations are added also in raising constructions.

Shared heads and dependents in coordination In coordinated structures, incoming and outgoing relations are connected only to the first conjunct in basic dependencies. Enhanced dependencies add explicit links from all conjuncts to shared dependents, like the subject *Sam* and the object *dinner* in (b), as well as from the shared head (not shown).

Null nodes for elided predicates Basic dependencies cannot represent predicate-argument relations in gapping constructions like (c), because of the missing verb, and therefore connect arguments and modifiers using a special *orphan* relation. By adding a null node with lexical information copied from the verb in the first clause, enhanced dependencies can assign the real argument relations to the subject *Kim* and the object *dinner*.

3 Adding Enhanced Dependencies

We describe three systems for predicting enhanced dependencies from basic dependencies. The first two systems have been adapted for cross-lingual use, while the third one uses language-specific information and will be used only for comparison in the evaluation in the next section. Other language-specific systems have been developed, such as the one by Candito et al. (2017) for French, but this is the first attempt to predict enhanced dependencies in a language-independent way.

3.1 The Rule-Based English System

The system is an adaptation of the work by Schuster and Manning (2016), developed for English. It relies on Semgrex (Chambers et al., 2007) patterns to find dependency structures that should be enhanced and applies heuristics-based processing steps corresponding to the five types of enhancement described in Section 2. We briefly discuss the three steps that are relevant to our study.

Added subject relations For any node attached to a higher predicate with an `xcomp` relation, the system adds a subject relation to the object of the higher predicate if an object is present (object control) or to the subject of the higher predicate if no object is present (subject control or raising). This heuristic gives the right result in Figure 1 (a).

Shared heads and dependents in coordination For conjoined clauses and verb phrases, the system adds explicit dependencies to shared core arguments (i.e., `(i)obj`, `n/csubj`, `x/ccomp`). Thus, in Figure 1(b), the system adds the `nsubj` and `obj` relations from *fixed* to *Sam* and *dinner*, respectively. For other types of coordination, it only adds dependencies from the shared head.

Null nodes Following Schuster et al. (2018), the system aligns arguments and modifiers in the gapped clause to the full clause. This alignment determines main clause predicates for which an empty node should be inserted. Finally, the gapped clause arguments and modifiers are reattached to the empty node, obtaining a structure such as the one in Figure 1 (c). This method uses word embeddings for the alignment of arguments; here we use the embeddings from the 2017 CoNLL Shared Task (Zeman et al., 2017).

	Subjects					Coordination					Null	
	Swe		Ita			Swe		Ita			Swe	Ita
	RBE	DDF	RBE	DDF	LSI	RBE	DDF	RBE	DDF	LSI	RBE	RBE
Count	127	36	115	43	88	559	981	421	653	660	112	162
Precision	0.87	0.83	0.80	0.95	0.91	0.94	0.91	0.89	0.82	0.85	0.85	0.76
Recall (pooled)	0.98	0.27	0.79	0.35	0.69	0.55	0.97	0.64	0.91	0.78		
Basic errors	12	1	14	0	2	25	28	12	32	19	15	0
Enhanced errors	4	5	9	2	6	9	69	34	86	65	2	35

Table 1: Evaluation of predicted enhanced dependencies for Italian and Swedish (RBE = rule-based English system, DDF = data-driven Finnish system, LSI = language-specific Italian system).

3.2 The Data-Driven Finnish System

This data-driven approach is adapted from the supervised method of Nyblom et al. (2013) originally developed for Finnish. First, patterns identify candidate relations, which are subsequently classified with a linear SVM, trained on gold standard annotation. The original method does not predict null nodes, and therefore we only discuss added subject relations and coordination below.

Added subject relations A binary classifier is used to decide whether an `nsubj` relation should be added from an `xcomp` dependent to the subject of its governor, accounting for subject control or raising (in the positive case). Object control is not handled by the original system, and we chose not to extend it for this initial case study.

Shared heads and dependents in coordination Candidate relations are created for all possible shared heads (incoming relation to the first conjunct) and dependents (outgoing relations from the first conjunct), striving for high recall. A classifier then predicts the relation type, or negative.

Feature representation To enable transfer from models trained on Finnish to other languages, we remove lexical and morphological features except universal POS tags and morphological categories that we expect to generalize well: `Number`, `Mood`, `Tense`, `VerbForm`, `Voice`. Language-specific dependency type features are generalized to universal types (e.g., from `nmod:tmod` to `nmod`).

3.3 The Language-Specific Italian System

The language-specific Italian system builds on the rule-based enhancer developed for the Italian Stanford Dependencies Treebank (Bosco et al., 2013, 2014). It has been adapted to predict enhanced dependencies according to the UD guidelines but does not yet handle null nodes. It provides an interesting point of comparison for the cross-lingual systems but cannot really be evaluated on the same conditions since it has been developed using data from the Italian treebank.

Added subject relations For any infinitive verb attached to a higher predicate with an `xcomp` relation, the system adds a subject relation to a core or (dative) oblique dependent of the governing verb. In contrast to the other systems, this system uses external language-specific resources that specify the control/raising properties of Italian verbs.

Shared heads and dependents in coordination For coordination, the system works similarly to the English rule-based system but includes additional heuristics for different types of coordination (clausal, verbal, nominal, etc.) to prevent the addition of multiple dependents of the same type (e.g., multiple subjects) if this leads to incorrect graphs.

4 Evaluation

Systems were evaluated on the Italian and Swedish UD treebanks. Since these lack enhanced dependencies annotation, the output is manually evaluated by native speakers with extensive experience with the UD guidelines. This allows us to report precision, while recall can only be measured relative to the union of correct predictions. The data-driven system was trained on the training set of the UD Finnish-TDT treebank.

We evaluate added subjects and coordination in a sample of 1,000 sentences from the training set of each treebank; the evaluation of null nodes for elided predicates, which occur more rarely, is based on the entire training sets. The results are shown in Table 1, with errors categorized as *basic errors* caused by errors in the basic dependencies, and *enhanced errors* attributed to the systems.

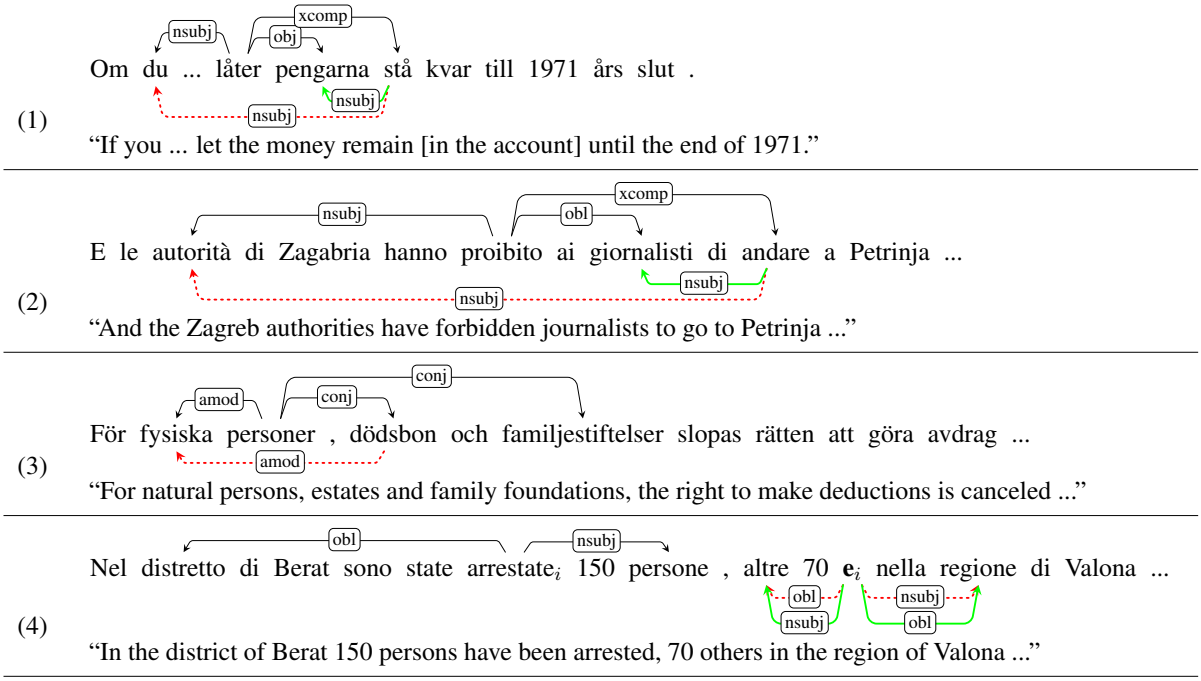

Figure 2: Error examples: added subjects (1–2), coordination (3), null nodes (4). Basic dependencies above, enhanced dependencies below; dotted red = incorrect; solid green = correct.

Added subject relations For Swedish, the rule-based English system (RBE) performs better than the data-driven Finnish system (DDF), especially on recall. The advantage in precision comes from object control, as illustrated in (1) in Figure 2 where DDF predicts subject control despite the presence of a direct object. The lower recall for DDF comes from only considering added subjects of infinitives (as opposed to all `xcomp` predicates). For Italian, the precision results are reversed, which is in part due to non-core arguments occurring more frequently as controllers in Italian. In this case, RBE will always predict a core argument (subject or object) as controller while DDF can abstain from predicting a dependency. The language-specific Italian system (LSI) correctly predicts most of the non-core controllers, thanks to lexical information, leading to higher precision than RBE. This is exemplified in (2) in Figure 2, where RBE predicts subject control while LSI finds the oblique controller and DDF makes no prediction at all. The lower recall for LSI is again caused by its restriction to infinitives.[2]

Shared heads and dependents in coordination The results for coordination are indicative of the different adopted strategies. RBE achieves high precision (0.94 for Swedish, 0.89 for Italian) by limiting shared dependent predictions to core arguments. DDF instead opts for high recall (0.97 for Swedish, 0.91 for Italian) by considering all dependents of the first conjunct as potential shared dependents. As a result, both systems outperform the language-specific system on one metric, but lose out on the other. The most common type of error, especially for the high-recall systems, is to treat a left-dependent of the first conjunct as shared by all conjuncts. This is exemplified by (3) in Figure 2, where DDF incorrectly predicts that the adjectival modifier in *fysiska personer* (natural persons) also applies to *dödsbon* (estates).

Null nodes for elided predicates The method developed to resolve gapping in English seems to generalize very well to Swedish, where almost all the observed errors are in fact due to errors in the basic annotation (mostly incorrect uses of the `orphan` relation). The results are somewhat lower for Italian, which allows word order variations that cannot be captured by the algorithm of Schuster et al. (2018). A case in point is (4) in Figure 2, where the order of the remnants in the gapped clause (`nsubj-obl`) is inverted compared to the complete clause (`obl-nsubj`).

[2]It is worth noting that the recall of both DDF and LSI could easily be improved by lifting the restriction to infinitives, since the non-infinitive cases are rarely ambiguous.

5 Conclusion

Our main conclusion is that both the rule-based English and the data-driven Finnish systems are accurate enough to be useful for enhancing treebanks in other languages. Precision is often above 0.9 (and never below 0.8) and recall is complementary, with the English system giving better coverage on added subjects and the Finnish one on coordination. The error analysis furthermore shows how both systems can be further improved. The results are especially encouraging given that one of the "source languages", Finnish, is typologically quite different from the others, which indicates that UD does generalize across languages.

For future research, it would be interesting to investigate how much language-specific training data would be needed for the data-driven system to exceed the cross-lingual results reported here. In addition, the same techniques can of course be used not only for treebank enhancement but also to post-process basic dependencies in parsing, which would potentially be useful for many downstream applications. An interesting question there is how much results would deteriorate because of parsing errors in the basic dependencies.

Acknowledgments

We are grateful to two reviewers for constructive comments on the first version of the paper. This work was supported in part by a gift from Google, Inc.

References

Alena Böhmová, Jan Hajič, Eva Hajičová, and Barbora Hladká. 2003. The Prague dependency treebank. In *Treebanks*, Springer, pages 103–127.

Cristina Bosco, Felice DellâĂŹOrletta, Simonetta Montemagni, Manuela Sanguinetti, and Maria Simi. 2014. The evalita 2014 dependency parsing task. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 and of the Fourth International Workshop EVALITA 2014*. Pisa University Press, volume II, pages 1–8.

Cristina Bosco, Simonetta Montemagni, and Maria Simi. 2013. Converting italian treebanks: Towards an italian stanford dependency treebank. In *Proceedings of the 7th ACL Linguistic Annotation Workshop and Interoperability with Discourse*. pages 61–69.

Marie Candito, Bruno Guillaume, Guy Perrie, and Djamé Seddah. 2017. Enhanced UD dependencies with neutralized diathesis alternation. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*. pages 42–53.

Nathanael Chambers, Daniel Cer, Trond Grenager, David Hall, Chloe Kiddon, Bill MacCartney, Marie-Catherine de Marneffe, Daniel Ramage, Eric Yeh, and Christopher D. Manning. 2007. Learning alignments and leveraging natural logic. In *Proceedings of the Workshop on Textual Entailment and Paraphrasing*. pages 165–170.

Yusuke Miyao and Jun'ichi Tsujii. 2004. Deep linguistic analysis for the accurate identification of predicate-argument relations. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*. pages 1392–1397.

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, and Dan Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*.

Jenna Nyblom, Samuel Kohonen, Katri Haverinen, Tapio Salakoski, and Filip Ginter. 2013. Predicting conjunct propagation and other extended stanford dependencies. In *Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013)*. pages 252–261.

Stephan Oepen and Jan Tore Lønning. 2006. Discriminant-based MRS banking. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*.

Siva Reddy, Oscar Täckström, Slav Petrov, Mark Steedman, and Mirella Lapata. 2017. Universal semantic parsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*. pages 89–101.

Sebastian Schuster and Christopher D. Manning. 2016. Enhanced english universal dependencies: An improved representation for natural language understanding tasks. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*.

Sebastian Schuster, Joakim Nivre, and Christopher D. Manning. 2018. Sentences with gapping: Parsing and reconstructing elided predicates. In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2018)*.

Sebastian Schuster, Éric Villemonte de la Clergerie, Marie Candito, Benoît Sagot, Christopher D. Manning, and Djamé Seddah. 2017. Paris and Stanford at EPE 2017: Downstream evaluation of graph-based dependency representations. In *Proceedings*

of the 2017 Shared Task on Extrinsic Parser Evaluation (EPE 2017).

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajic, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinkova, Jan Hajic jr., Jaroslava Hlavacova, Václava Kettnerová, Zdenka Uresova, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria dePaiva, Kira Droganova, Héctor Martínez Alonso, ÇağrÄś Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonca, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. Conll 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. pages 1–19.