Dependency Parsing beyond Simple Trees

Thesis proposal

Kira Droganova
Charles University
Faculty of Mathematics and Physics
Prague, Czechia
droganova@ufal.mff.cuni.cz

Abstract

Quite a few linguistic theories and annotation frameworks employ a layer that goes deeper than surface syntax, be it a deep- syntactic or semantic layer. This representation is universally considered more powerful for expressing the meaning. At the same time, it is usually much more difficult to obtain; therefore not surprisingly it is available only for a very limited set of languages. In this thesis proposal I outline phenomena which could be considered useful for deeper representation. My proposal is built upon and extends Enhanced Universal Dependencies (UD). I describe my future plans and ideas concerning parsing experiments and extrinsic evaluation.

1 Introduction

It has been a long journey since Tesnière (1959) introduced the concept of dependency tree, in which words are connected to each other with head-dependent relations to represent syntactic structure of a sentence, through the linguistic theories that distinguish between morphological, syntactic and semantic dependency relations — Meaning-Text Theory (Žolkovskij and Mel’čuk, 1965), Functional Generative Description (Sgall, 1967) — to the frameworks that contain a deep-syntactic, tectogrammatical, or semantic dependency layer — the Prague Dependency Treebank (Böhmová et al., 2003), ETAP-3 (Apresian et al., 2003), the Proposition Bank (Kingsbury and Palmer, 2002), Sequoia (Candito and Seddah, 2012), or Abstract Meaning Representation (Banarescu et al., 2013).

Despite different names and variations of annotated phenomena, all the above frameworks share a common idea of getting closer to representing the meaning by going deeper — a representation deeper than surface syntax is more useful for natural language understanding (but also more difficult to obtain).

Many of the deep frameworks have been applied to more than one language, sometimes just to demonstrate that it is possible; however, interest in multilingual research has grown dramatically since the CoNLL-X shared task on Multilingual Dependency Parsing (Buchholz and Marsi, 2006), which was run for 13 languages. Nowadays, with existing large multilingual treebank collections, multilingual experiments are practically a necessary component of any research on parsing.

Universal Dependencies (UD) (Nivre et al., 2016) annotation guidelines have become a de facto standard for cross-linguistically comparable morphological and syntactic annotation. A significant factor in the popularity of UD is a steadily growing and heavily multilingual collection of corpora: release 2.5 (Zeman et al., 2019) contains 157 treebanks of 90 languages. The UD guidelines have been designed as surface-syntactic, although their emphasis on cross-linguistic parallelism sometimes leads to decisions that are normally associated with deeper, semantics-oriented frameworks (the primacy of content words and lower importance of function words may serve as an example).

UD itself proposes an attempt to provide deeper annotations, dubbed Enhanced Universal Dependencies (Schuster and Manning, 2016). Enhanced UD is an optional extension, which is only available in a handful of treebanks (Droganova and Zeman, 2019). Enhanced UD faces the same threat as the other deep frameworks mentioned above: more complex annotation requires more annotation effort, and semantic annotations are often coupled with huge lexical resources such as verb frame dictionaries. Therefore, it is less likely that sufficient manpower will be available to annotate data in a new language. All this does not mean that the attempts
to enrich annotation should be abandoned; on the contrary, it might inspire development of alternative, less demanding methods.

There are two dimensions along which annotation of a resource can be improved. It can provide the same type of annotation as the light, semi-automatic version, but verified by human annotators. But it may also provide additional types of annotations that cannot be obtained automatically. In my research I am going to concentrate on both of these dimensions. I identify and propose a deep (enhanced) representation for a selection of phenomena that represent information potentially desirable for parsing — a selection of phenomena that are annotated in popular semantic dependency frameworks.

Some of the proposed phenomena can be derived semi-automatically from surface UD trees, in acceptable quality. These annotations will not be as precise as they would if carefully checked by humans, but they will be available for (almost) all UD languages. Moreover, it will be possible to generate them for new UD languages and the deep extension will thus keep up with the growth of UD. This dimension continues the direction that originated in the Deep UD project (Drogoanova and Zeman, 2019).

Other phenomena are available in their native frameworks only for a limited set of languages, but still provide valuable information and can be converted into deep representation. This direction may not look very attractive because the effort required to convert one phenomenon for one language may be comparable to effort required to convert another phenomenon for many languages. However, the availability of different phenomena for different languages makes it possible at least to try to expand these phenomena to other languages.

My proposal is built upon and extends Enhanced Universal Dependencies (see detail in sections 2.1 and 4). I intend to develop a parser that would allow anyone concerned to produce annotations with proposed enhancements (see details in section 5). It might be advantageous to extrinsically evaluate the proposed enhancements. However, this task involves a lot of external factors and restrictions (see section 6).

2 Related Work

2.1 Enhanced Universal Dependencies

The Enhanced UD (Schuster and Manning, 2016) serves as a basis of this research. UD v2 guidelines define five types of enhancements that can appear in treebanks released as part of UD. All the enhancements are optional and it is possible for a treebank to annotate one enhancement while ignoring the others. The enhanced representation is a directed graph but not necessarily a tree. It may contain ‘null’ nodes, multiple incoming edges and even cycles. The following enhancements are defined:

Null nodes for elided predicates. In certain types of ellipsis (gapping and stripping), multiple copies of a predicate are understood, each with its own set of arguments and adjuncts, but only one copy is present on the surface. Example: Mary flies to Berlin and Jeremy [flies] to Paris. The enhanced graph contains an extra node for each copy of the predicate that is missing on the surface. Note that the current UD guidelines do not license null nodes for other instances of ellipsis, such as dropped subject pronouns in pro-drop languages.

Propagation of conjuncts. Coordination groups several constituents that together play one role in the superordinate structure. They are all equal, despite the fact that the first conjunct is formally treated as the head in the basic UD tree. For example, several coordinate nominals may act as subjects of a verb, but only the first nominal is actually connected with the verb via an nsubj relation. In the enhanced graph, this relation is propagated to the other conjuncts, i.e., each coordinate nominal is directly connected to the verb (in addition to the conj relation that connects it to the first conjunct). Likewise, there may be shared dependents that are attached to the first conjunct in the basic tree, but in fact they modify the entire coordination. Their attachment will be propagated to the other conjuncts, too. (Note that not all dependents of the first conjunct must be shared. Some of them may modify only the first conjunct, especially if the other conjuncts have similar dependents of their own.)

While Schuster and Manning (2016) remains the most suitable reference for Enhanced UD to date, its publication pre-dates the v2 UD guidelines and the proposals it contains are only partially compliant with the guidelines. See https://universaldependencies.org/u-overview/enhanced-syntax.html for the current version.
External subjects. Certain types of non-finite, ‘open’ clausal complements inherit their subject from the subject or the object of the matrix clause. Example: Susan wants to buy a book. In the basic tree, Susan will be attached as the nsubj of wants, while there will be no subject dependent of buy. In contrast, the enhanced graph will have an additional nsubj relation between buy and Susan.

Relative clauses. The noun modified by a relative clause plays a semantic role in the frame of the subordinate predicate. In the basic UD tree, it is represented by a relative pronoun; however, in the enhanced graph it is linked from the subordinate predicate instead of the pronoun. (The pronoun is detached from the predicate and attached to the noun it represents, via a special relation ref.) This is the reason why enhanced graphs may contain cycles: in The boy who lived, there is an acl:relcl relation from boy to lived, and an nsubj relation from lived to boy.

Case information. The labels of certain dependency relations are augmented with case information, which may be an adposition, a morphological feature, or both. For example, the German prepositional phrase auf dem Boden (on the ground) may be attached as an oblique dependent (obl) of a verb in the basic tree. The enhanced label will be ob1:auf:dat, reflecting that the phrase is in the dative case with the preposition auf. This information is potentially useful for semantic role disambiguation, and putting it to the label is supposed to make it more visible; nevertheless, its acquisition from the basic tree is completely deterministic, and there is no attempt to translate the labels to a language-independent description of meaning.

Several extensions of the enhanced representation have been proposed. The enhanced++ graphs proposed by Schuster and Manning (2016) extend the set of ellipsis-in-coordination types where null nodes are added; they also suppress quantifying expressions in sentences like a bunch of people are coming.

Candido et al. (2017) define the enhanced-alt graphs, which neutralize syntactic alternations, that is, passives, medio-passives, impersonal constructions and causatives. They also suggest annotating external arguments of non-finite verb forms other than just open infinitival complements and relative clauses: most notably, for participles, even if they are used attributively. Hence in ceux embauchés en 2007 (those hired in 2007), embauchés heads a non-relative adnominal clause (acl) that modifies the nominal ceux, but at the same time ceux is attached as a passive subject (nsubj:pass) of embauchés.

2.2 Other approaches to deep syntactic annotation

Manual semantic annotation is a complex and highly time-consuming process, therefore the data is available only for a limited set of languages. To deal with this issue, a number of researchers have experimented with (semi-)automatic approaches to semantic annotation. Padó (2007) proposes a method that uses parallel corpora to project annotation to transfer semantic roles from English to resource-poorer languages. The experiment was conducted on an English-German corpus. Van der Plas et al. (2011) experimented with joint syntactic-semantic learning aiming at improving the quality of semantic annotations from automatic cross-lingual transfer. An alternative approach is proposed by Exner et al. (2016). Instead of utilizing parallel corpora, they use loosely parallel corpora where sentences are not required to be exact translations of each other. Semantic annotations are transferred from one language to another using sentences aligned by entities. The experiment was conducted using the English, Swedish, and French editions of Wikipedia. Akbik et al. (2015) describe a two-stage approach to cross-lingual semantic role labeling (SRL) that was used to generate Proposition Banks for 7 languages. First, they applied a filtered annotation projection to parallel corpora, which was intended to achieve higher precision for a target corpus, even if containing fewer labels. Then they bootstrapped and retrained the SRL to iteratively improve recall without reducing precision. This approach was also applied to 7 treebanks from UD release 1.4.² However, the project seems to be stalled. Mille et al. (2018) propose the deep datasets that were used in the Shallow and Deep Tracks of the Multilingual Surface Realisation Shared Task (SR’18, SR’19). The Shallow Track datasets consist of unordered syntactic trees with all the word forms replaced with their lemmas; part-of-speech tags and morphological information are preserved (available for 10 languages). The Deep Track

²https://github.com/System-T/UniversalPropositions
datasets consist of trees that contain only content words linked by predicate-argument edges in the PropBank fashion (available for English, French and Spanish). The datasets were automatically derived from UD trees v.2.0. Gotham and Haug (2018) propose an approach to deriving semantic representations from UD structures that is based on techniques developed for Glue semantics for LFG. The important feature of this approach is that it relies on language-specific resources as little as possible.

3 Previous Experiments

I started my research with a preliminary study of the data. I chose elliptic constructions as the phenomenon of my primary interest. Ellipsis, i.e. omission of linguistic content that can be reconstructed from the context formed by the remaining elements, is present in various forms in many languages and obviously makes natural language understanding harder. My experiments were limited to certain types of ellipsis — gapping and stripping (Johnson, 2009; Coppock, 2001; Merchant, 2016): the types that are specified in the UD guidelines. Although the phenomenon is naturally rare, I considered it the most suitable for exploring the data because not only is it annotated in the Enhanced UD, but also it is visible in the basic representation of UD — elliptic constructions can be traced through the orphan relation that is used to attach unpromoted dependents of a predicate to the promoted dependent (Figure 1). It should be noted that when I started experimenting with the data, enhanced representation was available only for 3 languages, thus the initial experiments were conducted on the basic representation.

I conducted a survey (Droganova and Zeman, 2017) on annotation of ellipsis in UD treebanks (version 2.0). The main motivation here was to investigate the types and frequencies of elliptical constructions that are present in the treebanks of different languages. The findings were not encouraging:

- around 40% of the treebanks did not contain sentences with gapping;
- the number of sentences with gapping within a treebank was not sufficient for further experiments — neither parser learning nor linguistic or cross-linguistic studies; only 12 treebanks had more than 100 sentences with orphans;
- after manual analysis, it turned out that the number of annotation errors is rather high, which definitely reflects the complexity of this linguistic phenomenon.

Additionally, we proposed a method of identifying sentences where an orphan is missing. We have shown that our automatic tests can at least partially help to detect erroneously annotated sentences with gapping and to improve future heuristics for identifying ellipsis in UD.

The annotation errors discovered and our experiments on automatic identification of erroneously annotated sentences with gapping inspired a series of further experiments (Droganova et al., 2018b), in which I examine the then latest parsers in order to learn about parsing accuracy and typical errors that they yield on elliptic constructions. For the purpose of these experiments I adapted and extended the evaluation script which had been created to evaluate system output files for the 2017 CoNLL Shared Task (Zeman et al., 2017). The main idea of such adaptation was to preserve the original evaluation techniques that were used within the shared task; following the same line, especially regarding word alignments and sentence segmentation, allows for more precise results. It turned out that parsers make mistakes in similar conditions: the error types and their frequencies are almost the same from parser to parser.

The number of orphan labels is just a tiny fraction of all labels and the contribution of their low Recall and F-measure to the final figures calculated for the whole amount of data goes virtually unseen. The important question is whether the parsers perform really poorly on elliptical constructions or whether it is simply the lack of data. To address this issue we created a collection of artificial treebanks for parsing experiments on elliptical constructions. Re-applying the idea of typical patterns that can be used for detection of elliptical constructions, I implemented general conversion rules, which transform a full sentence of a certain structure into a sentence with gapping by deleting certain linguistic material. Then we tuned the rules and tested them for languages that we included in the study — Czech, English and Finnish.
We further developed this line of research by conducting experiments in enrichment of training data for this specific construction, evaluated for five languages: Czech, English, Finnish, Russian and Slovak (Droganova et al., 2018a). We proposed data enrichment methods that draw upon self-training and tri-training, combined with a stratified sampling method mimicking the structural complexity of the original treebank. We started with experiments on enriching data in general, without a specific focus on gapping constructions. Then we focused on elliptical sentences, comparing general enrichment of training data with enrichment using artificially constructed elliptical sentences. Although we were able to demonstrate small improvements over the CoNLL-17 parsing shared task winning system for four of the five languages, not only restricted to the elliptical constructions, our enrichment experiments focused on gapping led to mixed results. For several languages we did not obtain a significant improvement in the parsing accuracy of ellipsis. At the same time, enrichment experiments on artificial treebanks demonstrated promising results. Experiments with English seemed the most promising — the best F-score of the predicted orphan relation was more than ten times higher compared to using the original treebank; this also tests the method’s applicability when a treebank contains almost no elliptical constructions and training on it results in parsers that only generate the orphan relation very rarely. In general, it seems that the techniques work better for smaller treebanks that do not contain sufficient numbers of sentences with gapping.

The work described in (Droganova and Zeman, 2019) is probably the most important step in the right direction. With the knowledge that the UD data mostly lacks the enhanced annotation layer, we designed a prototype of Deep Universal Dependencies, a concept where minimal deep annotation can be derived automatically from surface UD trees. First, we generated enhanced graphs with the Stanford Enhancer\(^4\) for corpora that lack them,\(^5\) even though it does not guarantee that even if all five types of enhancements are present in the data, all of them will be correctly identified and annotated in the resulting annotation. Then we selected phenomena and prepared extraction procedures for them:

- We started with verbal predicates and identification of their arguments, if present in the same sentence.
- We made sure that the argument with a particular semantic role would always get the same label/number by neutralizing valency-changing operations such as passivization.\(^6\)
- We added a heuristic that connects infinitives to their subjects that should be inherited from the matrix clause in specific cases of adverbial and adnominal clauses.
- Similar to a relative clause, in which the enhanced graph would identify the modified noun, we added a heuristic for participles that are attached as amod — they take the modified noun as their argument; in order to determine whether the noun is argument 1 or 2, we distinguish active and passive participles.

What is important is that it is possible to generate deep annotation for new UD languages and thus the deep extension will keep up with the growth of UD.

4 Enhancements

In this section I outline additional phenomena that could be useful for deeper representation. In my research I utilize the same CoNLL-U Plus file format\(^7\) that is described in (Droganova and Zeman, 2018a).\(^4\) The Stanford UD Enhancer was adapted from an older tool that was designed to work with the Stanford Dependencies, a predecessor of UD.

\(^5\)Some corpora were excluded for copyright reasons; we also excluded corpora with incomplete or non-existent lemmatization.

\(^6\)Note that we do not label the actual semantic roles.

\(^7\)https://universaldependencies.org/ext-format.html
two new columns, DEEP:PRED and DEEP:ARGS, contain annotation that was added on top of Enhanced UD; without them, the file is still a valid CoNLL-U file. I begin by extending our ideas from the Deep UD project.

Valency frames. Information concerning valency frames is extremely useful, but at the same time it is hard to obtain. It would be beneficial to link information from valency dictionaries or FrameNets — argument labels and frame ids — to corresponding instances in the UD data. These lexical resources exist for a number of languages, such as English, German, Swedish, French, Spanish, Brazilian Portuguese, Czech, Russian, Chinese, Japanese, Korean. A prominent case concerns Czech and English, which were involved in a project dedicated to cross-linguistic comparison of valency behavior of Czech and English verbs, CzEngVallex (Urešová et al., 2016), thus a Czech-English parallel corpus enhanced with a manual linguistic annotation up to the tectogrammatical (deep syntax) layer is available. Although the resource utilizes the tectogrammatical layer, which is not available in the UD treebanks, the availability of the parallel data allows one to experiment with machine learning techniques and even try to evaluate the results on different lexical resources which are available for English.

Oblique arguments. In some languages, such as Czech (Havránek and Jedlička, 1966) and Russian (Testelets, 2001), according to their grammar, arguments can be expressed by prepositional noun phrases. The obl:arg relation is designed to retain this information and distinguishes oblique arguments (Figure 2) from adjuncts (Figure 3), which use the plain obl relation. The relation obl:arg is a language-specific subtype, therefore it occurs only in 11 languages: Arabic, Czech, German, Latin, Lithuanian, Maltese, Naija, Polish, Sanskrit, Slovak, Tamil. For this set of languages it should be easy enough to work this information into the deep representation level (Figure 2). For treebanks that do not use this label it should be possible to reconstruct it using machine learning techniques and/or external valency dictionaries. An important prerequisite here would be checking how a particular language deals with the argument-adjunct distinction.

Predicates other than verbs. Apart from verbal predicates, annotation of predicative nominals and predicative adjectives such as My brother is an artist and The book was interesting could be beneficial for deep representation. Such cases are not limited to constructions with the copula (see Figure 4 and 5 where an adjective appears as the head of an adverbial clause and a clausal complement respectively), and include more complicated cases (Figure 6) in which a participle is tagged ADJ and appears in the head position of clausal modifier of a noun, but it is not a relative clause. Although it is hard to come up with an example in which more than one (first) argument would be present in such sentences, this approach unifies the representation and makes it more consistent in a sense that predicate-argument structure will be present on the deep representation level for sentences with different surface structure.

Probably the most difficult case in this category is annotation of deverbal ("eventive") nouns — nouns that are derived from verbs, often by adding a derivational affix (for instance, the noun consultation was derived from the verb consult), that behave grammatically as nouns. This issue is especially difficult to approach from a multilingual perspective: to my knowledge, there are just a few resources that provide relevant annotation — the NomBank (Meyers et al., 2004) project covers nominalizations of verbs and adjectives in English, NomLex-PT (Paiva et al., 2014) is a lexicon of Portuguese nominalizations and PDT-Vallex (Hajič et al., 2003) is a resource that contains valency patterns of verbs, nouns, adjectives and adverbs as they occurred in the Prague Dependency Treebank, Prague Czech-English Depen-

---

8UD version 2.4
Even if physically not everything is very comfortable, it is bearable.

Я считаю, что для проведения научной работы пол года вполне достаточно.

Partitives and light noun constructions. For the analysis of partitive noun phrases such as both of the girls, the authors propose to treat the first part of the phrase as a quantificational determiner by promoting the semantically salient noun phrase girls to be the head of the partitive; the quantificational determiner is analyzed as a flat multi-word expression that is headed by its first word (Figure 7). Quantificational determiners reside in a closed class, thus it should be possible to create a list of quantificational determiners for other languages.

The authors propose a similar analysis for light noun constructions, such as a bunch of people, in which the second noun phrase tends to be the semantically salient one while the first part of the phrase serves as a quantificational determiner. It would be tricky to detect such constructions for different languages — even if a complete list of such constructions existed for English, their adaptation requires translation and at least some fluency in the target language; this does not guarantee that the target language does not have its own specific constructions that should be treated in the same manner.

Another case that was not mentioned in the paper, but still can be considered similar in a sense that the second noun phrase is more semantically salient than the first part of the phrase, concerns larger numbers like thousand, million, billion, which in phrases like thousands of people indicate quantity even if it is a noun or a numeral rather than a determiner. For such cases it also should be possible to create a list of lemmas for other languages.

Conjoined prepositions and prepositional phrases. There are challenging issues connected with the omission of a word or several words that can occur in conjoined prepositions or prepositional phrases, such as I bike to and from work or I flew to Paris or to Moscow. The authors suggest that all the information should be encoded in UD graphs (Figure 8): nmod:to, as well as an nmod:from, should explicitly mark an edge between bike and work; in order to show that bike to work and bike from work are conjoined by and the representation contains copied node bike, which is attached to the original node as a conjunct. It might be possible to identify such cases by specific subtree patterns (Figure 9, 10).

I do not consider cases such as Children drew with red crayons and markers due to structural ambiguity — such cases are difficult for automatic

9This depends on the traditional grammar of a language. For instance, in Russian tradition these lemmas are treated as nouns — they have a paradigm structure and endings similar to the regular classes of substantives.
Indiáni přijeli z Kanady vedeni Josefem Brantem

Both of the girls are singing

I bike to and from work

Figure 6: An example of a sentence with an adjective as a clausal modifier of a noun: *Indiáni přijeli z Kanady vedeni Josefem Brantem.* *Indians arrived from Canada, led by Joseph Brant.*

Figure 7: An example of annotation for the quantificational determiner det:gmod.

Figure 8: An example of annotation for conjoined prepositions: basic UD on top and the proposal from (Schuster and Manning, 2016) re-annotated according to the UD guidelines v 2.0 at the bottom.

5 Parsing

In this work, I do not aim to implement a new state-of-the-art graph parser. However, I think that it is crucial to provide a reasonably good parser that would allow anyone concerned to work with the proposed enhancements as well as with already existing enhanced UD and Deep UD. For this purpose I came up with the following ideas:

- Utilize an existing parser for Enhanced UD and use it as a separate module; create another separate module for parsing the proposed new enhancements.

Utilizing an existing parser for Enhanced UD would allow me to concentrate more on developing a parser for the proposed enhancements. To try this idea, I consider using Stanford Enhancer, which is available as a part of Stanford CoreNLP (Manning et al., 2014). The main disadvantage of this approach is that Stanford Enhancer forms a part of complex software which provides a set of human language technology tools; thus it might be difficult to detach or tune this module separately. Therefore, it is highly likely that the enhancer can be used only as a black box.

The shared task on Cross-Framework Meaning Representation Parsing (MRP 2019) is the most recent effort to advance data-driven parsing into graph-structured representations of sentence meaning (Oepen et al., 2019). Reusing information on the algorithms that were implemented in the best MRP 2019 systems, I have created a shortlist of tools employing methods that are worth exploring for developing a parser for the new enhancements. The main points of interest are the top parsers for DELPH-IN MRS Bi-Lexical Dependencies (DM) and Prague Semantic Dependencies (PSD) frameworks. DM graphs were originally produced by a two-stage conversion process from the underspecified logical forms — of-
ten referred to as English Resource Semantics, or ERS (Bender et al., 2015) — into bi-lexical semantic dependency graphs (Ivanova et al., 2012). PSD graphs were produced by reduction of tectogrammatical trees (or t-trees) from the linguistic school of Functional Generative Description (Sgall et al., 1986; Hajic et al., 2012). Despite the difference in the represented linguistic information — DM and PSD graphs mainly contain information concerning valency frames while the enhanced graphs mainly represent other linguistic phenomena, such as gapping, propagation of conjuncts, etc. — these graphs seem highly relevant for my purpose: in DM and PSD graphs as well as in the proposed enhanced graphs some tokens do not contribute to the graph, and there are multiple incoming edges for some nodes. Importantly, all tools that I chose for further experiments have GitHub pages and documentation.

It is worth mentioning that there is a good opportunity to test one of the earliest versions of the parser by participating in a shared task on parsing enhanced UD graphs, which will be a part of IWPT 2020.

6 Extrinsic evaluation

An important step of this research would be to assess a contribution of the new information (the proposed enhancements) to the quality of downstream applications. In this section I consider potentially beneficial research directions rather than propose a specific course of actions. Although there has been considerable work on extrinsic evaluation of syntactic parsers from different angles — experiments concerning parsers, parsing models, conversion schemes, representations, etc. — extrinsic evaluation is still a laborious and complex task. To my knowledge, no common procedures have been developed to perform extrinsic evaluation on custom representations. This task is further complicated by the following:

- Downstream systems are rarely freely available and if available, they do not generally provide architecture that would allow for easy modification of the evaluated module.
- Such systems require adaptation for every custom module in question.
- Even if a common extrinsic evaluation mechanism were to be created, applications change fast, thus the evaluation results may not generalize well to newer systems.

- It is not immediately clear which downstream applications are better suited to the evaluation. This poses a research question in itself.

Miyao et al. (2008) propose a comparative evaluation of constituent-based, dependency-based, and deep linguistic parsers on an information extraction system that performs protein-protein interaction (PPI) identification in biomedical papers. The authors evaluate eight parsers using five different parse representations and experiment with several combinations of parser and parse representation. The experiments show that the results are similar for all parsers in question, but utilizing domain-specific data improves accuracy; improvements vary from parser to parser.

Johansson and Nugues (2008) compare constituent-based and dependency-based representations for the semantic role labeling task for English. The authors demonstrate that dependency-based systems perform slightly better on the argument classification task, and the results are slightly lower on the argument identification task. In addition, the results show that dependency-based semantic role classifiers rely less on lexicalized features, which makes them more robust to domain changes.

Buyko and Hahn (2010) compare the 2007 and 2008 CoNLL schemes and Stanford Basic Dependencies for event extraction from biomedical text. The results show that the content-oriented Stanford scheme is less suitable for the task than the CoNLL representations.

Popel et al. (2011) study the influence of different dependency-parsing techniques on the quality of an English-Czech dependency-based machine translation system (TectoMT) (Žabokrtský et al., 2008). The authors experiment with graph-based, transition-based, and phrase-structure parsers and utilize the same syntactic representation. The results show that UAS does not correlate well with the effect on translation quality for parsers that are based on different dependency-parsing techniques.

Yuret et al. (2010) present an overview of the Parser Evaluation using Textual Entailments (PETE) shared task in the SemEval-2010 Evaluation Exercises on Semantic Evaluation. The task involves recognizing textual entailments based on
Elming et al. (2013) focus on comparison of different types of dependency representations and their contributions over several different downstream tasks where syntactic features are known to be effective: negation resolution, semantic role labeling, statistical machine translation, sentence compression, and perspective classification. The results show that not only does the choice of dependency representation have clear effects on the downstream results, but also that these effects vary depending on the task.

Gómez-Rodríguez et al. (2019) evaluate the influence of 4 different dependency parsers on the performance of a rule-based sentiment analysis system that determines the polarity of sentences from their parse trees. The parsers show equally good results in the sentiment analysis task; experiments do not show any relevant influence of the parser accuracy on the results.

A series of shared tasks on Extrinsic Parser Evaluation (Oepen et al., 2017; Fares et al., 2018) is probably the most prominent recent effort to explore the contribution of different types of dependency representations to a variety of downstream tasks. Quantitative intrinsic scores of standard metrics do not immediately indicate corresponding advances in natural language understanding tasks.

The 2017 Shared Task on Extrinsic Parser Evaluation (EPE 2017) was intended to explore the downstream utility of various representations at available levels of accuracy for different parsers to a selection of state-of-the-art downstream applications — biomedical event extraction, negation resolution, and fine-grained opinion analysis systems — which utilize different types of text. EPE 2017 was limited to parsing only English text. The range of representation varies from syntactic in nature to so-called semantic dependency representations, which necessarily take the form of unrestricted directed graphs. Although the shared task organizers think that it is difficult to compare results due to multiple variables — the parser (and its output quality), the representation, input preprocessing, and the amount and domain of training data, I have noticed that the winning system utilizes enhanced graphs\textsuperscript{10}, and this really helped in the Negation Resolution and Opinion Analysis tasks.

The 2018 Shared Task on Extrinsic Parser Evaluation (EPE 2018) was organized as an optional track of the 2018 Shared Task on Multilingual Parsing from Raw Text to Universal Dependencies (Zeman et al., 2018). The shared task utilizes the same set of downstream applications and basic UD version 2.x as the main shared task representation; training data was limited to the English UD treebanks provided for the core task.

Fortunately, the negation resolution system, Sherlock (Lapponi et al., 2017), is available\textsuperscript{11} for further experiments and it is possible to re-run the experiment on the same dataset as at EPE 2018.

\begin{table}[h]
\centering
\begin{tabular}{|l|l|c|c|l|l|}
\hline
Parser & Reference & DM & PSD & Architecture & Embeddings \\
\hline
HIT-SCIR & Che et al. (2019) & .951 & .905 & transition-based with stack LSTM & enhanced BERT \\
AM-parser & Donatelli et al. (2019) & .947 & .913 & composition-based; BiLSTM & BERT \\
ShanghaiTech & Wang et al. (2019) & .949 & .895 & graph-based & BERT, GloVE \\
\hline
\end{tabular}
\caption{A shortlist of tools. DM and PSD columns show the maximal F$_1$ scores that a system was able to achieve.}
\end{table}
In general, to implement the idea of extrinsic evaluation I need to identify downstream applications that rely heavily on grammatical structure, i.e., are able to recognize complex and interacting relations where the component pieces are often syntactic-to-semantic constituents whose interactions are mediated by grammar. I see some potential in the fact extraction, relation extraction, and question answering disciplines. For instance, the organizers of the shared task on semantic relation extraction and classification in scientific paper abstracts at SemEval-2018 (Gábor et al., 2018) underline the relevance of dependency trees for the task; participants of the community question answering shared task at SemEval-2015 (Nakov et al., 2015) utilize dependency trees with varying degrees of success.

7 Outlook for the future

My thesis is built upon Enhanced Universal Dependencies (UD) and extends the Deep UD project. I propose a set of linguistic phenomena that could be beneficial for deeper representation.

I am going to begin by experimenting with valency frames. The complete success would be to learn to predict full frames for languages with lexical resources available. I will start my experiments with the Czech-English parallel corpus enhanced with a manual linguistic annotation up to the tectogrammatical (deep syntax) layer, gradually extending the evaluation to other lexical resources available for English.

Then I am going to add information concerning predicative nominals and predicative adjectives including the cases where an adjective appears as the head of an adverbial clause or a clausal complement. I am going to experiment with a set of universal rules which potentially can be applied to all UD treebanks.

After that I am going to experiment with conjoined prepositions and prepositional phrases and enhance the annotation of such constructions.

Next I am going to get a closer look at participative and estimate to what extent and for which languages it would be possible to enhance the deep layer.

I seek to develop a reasonably good parser that would allow anyone concerned to work with the proposed enhancements as well as with already existing Deep UD enhancements. For that reason I utilize and (partially) re-implement one of the methods that have already proved well suited for non-tree parsing.

I intend to design experiments and conduct extrinsic evaluation for proposed deep annotation. However, this task is highly dependent on external software, therefore it is hard to foresee the extent to which this task could be accomplished.

References


Lucien Tesnière. 1959. Eléments de syntaxe structurale.


