Syntax Parsing of Morphologically Rich Languages and Its Application

Summary of the PhD thesis

Zsolt Szántó
Supervisor: Richárd Farkas, PhD

Doctoral School of Informatics
Department of Computer Algorithms and Artificial Intelligence
Faculty of Science and Informatics
University of Szeged

Szeged
2024

1 Introduction

The internet is a vast repository of information that is constantly growing, with new pages being created and updated every day. Much of this information is in the form of text, such as news articles, blog posts, and product descriptions.

Natural language processing (NLP) is a field of computer science that deals with the interaction between computers and human language. NLP has a wide range of applications, including machine translation, text summarization, and question answering. One of the key challenges in NLP is understanding the syntactic structure of human language sentences. Syntax determines the order of words, how words of a sentence relate to each other, and how phrases and clauses are built within a sentence.

Syntactic parsing is the process of analyzing the structure of sentences to uncover their underlying grammatical relationships. Because of the complexity and ambiguity of human language, syntactic parsers mostly rely on machine learning-based solutions.

In most cases, syntactic parsing is used as a tool for higher-level text-processing applications like information extraction, machine translation, or sentiment analysis. For example, in aspect-based sentiment analysis we aim to assign different sentiment values to different parts of the text. Take the following social media post as an example:

(1) I love the last jedi, but not a fan of the rise of skywalker

There is positive sentiment about The Last Jedi Star Wars movie, but for The Rise of Skywalker the writer shared a negative opinion. If we know the syntax of the sentence we can easily see the `love` is connected to the `last jedi` and the `not a fan` is related to the `rise of skywalker`.

Features extracted from syntax parses has led to state-of-the-art machine learning application in many text processing fields (Björne et al., 2009; Lapponi et al., 2012; Johansson and Moschitti, 2013) since 2006. Since 2019, a major trend in natural language processing research has been the use of end-to-end approaches based on large, pre-trained neural language models. These models are often fine-tuned for specific applications. Although incorporating syntactic parsing as an intermediate step can further enhance the effectiveness of these methods (Zeng et al., 2019), the added value of externally given syntactic information is much lower than previously.

While deep learning solutions often achieve high accuracy, the demand for interpretable output remains crucial in real-world language processing systems. Industrial applications are frequently fully or partially rule-based solutions, as (sufficient) training data for a pure machine learning solution is not available and each and every real-world application has its own requirements. Moreover, rule-based components provide tight control over the behavior of the systems in contrast to other approaches. Experts in a particular field can design application-specific rules based on the relationship of certain words thanks to syntactic parsing. In general, although their relevancy has been decreased, we believe that syntactic parsers are useful, even in the Large Language Model era.

The most popular approaches to syntactic parsing are constituent parsing and dependency parsing, each offering a unique perspective on sentence structure. Constituency parsing breaks down sentences into nested phrases (e.g., noun phrases, verb phrases), exposing the hierarchical structure of language. This type of analysis facilitates applications that require an understanding of how sentence components function together. Dependency parsing, on the other hand, shows the direct, grammatical connections between words, highlighting their roles within a sentence. This information is vital for tasks where accuracy depends on precise interpretation of linguistic relationships such as information extraction.

			Chapters				
			3	4	5	6	7
EACL	2014	Szántó and Farkas (2014)	•				
ACTA	2015	Szántó and Farkas (2015)	•				
COLING	2014	Simkó et al. (2014)		•			
EACL	2017	Vincze et al. (2017)		•			
TSD	2023	Orosz et al. (2023)		•			
ICCIA	2017	Hangya et al. (2017)			•		
EPE	2017	Szántó and Farkas (2017)				•	
AIAI	2023	Szántó et al. (2023)					•

Table 1: *Connection between the chapters of the thesis and the corresponding publications.*

Besides potential applications, this thesis focuses on Hungarian and other morphologically rich languages, that express syntactic information at the level of the morphology of the words instead of encoding it in the word order. Among its contributions, the thesis presents techniques for constituent and dependency parsing that achieve state-of-the-art accuracy on morphologically rich languages and in some cases at least competitive results. Additionally, it introduces methods for automatic, rule-based conversion between constituent and dependency corpora, as well as between different dependency representations for Hungarian that were used to create the Hungarian Universal Dependencies dataset.

2 Structure of the Dissertation

The document is separated into three parts, each describing a specific field of syntactic parsing: constituent parsing, dependency parsing, and the high level application of these tools.

Part I introduces novel approaches for constituent parsing, especially for morphologically rich languages. To enhance the efficiency of parsing systems, it introduces techniques like a novel preterminal merger procedure and leveraging external corpora within the lexical model. This part also shows improvement on the reranking step of constituent parsers and demonstrates that incorporating features based on morphological details leads to improved outcomes for morphologically rich languages (Szántó and Farkas, 2014; Szántó and Farkas, 2015).

Part II provides an overview of Hungarian dependency parsing. It delves into the relationship between dependency and constituent parsing in Hungarian (Simkó et al., 2014). Additionally, it analyzes the development of the Universal Dependency dataset for Hungarian (Vincze et al., 2017). Finally, the chapter highlights HuSpaCy, a Hungarian language processing framework that provides a cutting-edge dependency parser (Orosz et al., 2023).

Part III introduces three application of syntactic parsing, exploring its potential in three distinct areas. Firstly, it examines how utilizing syntactic structures can enhance sentiment analysis (SA) on both individual words and entire sentences (Hangya et al., 2017). Secondly, it delves into Team Szeged's participation in the First Shared Task on Extrinsic Parser Evaluation (EPE 2017), presenting three methods that leverage the challenge's shared generalized dependency graph representation (Szántó and Farkas, 2017) for various downstream applications. Finally, the chapter explores strategies for medication event extraction and classification, demonstrating how syntax-based information extraction can lead to efficiency gains even on the top of pre-trained large language

models (Szántó et al., 2023).

Table 1 outlines the connections between the thesis chapters and the key publications they reference.

3 List of Publications

- Szántó, Z., Farkas, R.: Special techniques for constituent parsing of morphologically rich languages. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. pp. 135–144. Gothenburg, Sweden (April 2014)
- Szántó, Z., Farkas, R.: Constituency parse reranking for morphologically rich languages. Acta Polytechnica Hungarica 12(8) (2015)
- Simkó, K.I., Vincze, V., Szántó, Zs., Farkas, R.: An Empirical Evaluation of Automatic Conversion from Constituency to Dependency in Hungarian. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. pp. 1392–1401 (2014)
- Vincze, V., Simkó, K., Szántó, Z., Farkas, R.: Universal Dependencies and morphology for Hungarian - and on the price of universality. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. pp. 356–365. Association for Computational Linguistics, Valencia, Spain (Apr 2017)
- Orosz, G., Szabó, G., Berkecz, P., Szántó, Z., Farkas, R.: Advancing hungarian text processing with huspacy: Efficient and accurate nlp pipelines. In: International Conference on Text, Speech, and Dialogue. pp. 58–69. Springer (2023)
- Hangya, V., Szántó, Z., Farkas, R.: Latent syntactic structure-based sentiment analysis. In: 2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA). pp. 248–254. IEEE (2017)
- Szántó, Z., Farkas, R.: Szeged at epe 2017: First experiments in a generalized syntactic parsing framework. In: Proceedings of the 2017 Shared Task on Extrinsic Parser Evaluation at the Fourth International Conference on Dependency Linguistics and the 15th International Conference on Parsing Technologies. Pisa, Italy. pp. 75–79 (2017)
- Szántó, Z., Bánáti, B., Zombori, T.: Enhancing medication event classification with syntax parsing and adversarial learning. In: Maglogiannis, I., Iliadis, L., MacIntyre, J., Dominguez, M. (eds.) Artificial Intelligence Applications and Innovations. pp. 114–124. Springer Nature Switzerland, Cham (2023)

4 Constituent Parsing

Chapter 3 presents different techniques to improve constituent parsing; especially for handling the challenges of morphologically rich languages. Section 3.3 introduces my proposals that utilize the information from the constituent trees, while Section 3.4 demonstrates my approaches that use external information like large amounts of unlabeled data and dependency trees.

One of the chief contributions of Section 3.3 is to propose a novel automatic procedure to find the optimal set of preterminals by merging morphological feature values. The size of the preterminal set in the standard context free grammar environment is crucial. If we use only the main POS tags as preterminals, we lose a lot of information encoded in the morphological description of the tokens. On the other hand, using the full morphological description as preterminal yields a set of over a thousand preterminals, which results in data sparsity and performance problems as well. The main novelties of our approach over previous work are that it is very fast – it operates inside a probabilistic context free grammar (PCFG) instead of using a parser as a black box with re-training for every evaluation of a feature combination – and it can investigate particular morphological feature values instead of removing a feature with all of its values. I also experimented with exploiting external corpora in the lexical model. A new scientific result is that automatic tagging of an off-the-shelf supervised morphological tagger can also contribute to the results. My last experiment was carried out with the feature set of an n-best reranker. We showed that incorporating feature templates built on morphological information improves the results. The results were published in the article "Special Techniques for Constituent Parsing of Morphologically Rich Languages" at the *EACL'14 conference* as a long paper (Szántó and Farkas, 2014).

Section 3.4 focused on approaches that use external information like large amounts of unlabeled data and dependency trees to improve the accuracy of constituent parsers. In the discriminative reranking step I introduced a new feature template that employs dependency-based information. By using the unlabeled data, I applied Brown clustering which I also applied as features in the final system. These methods were published in the "Constituency Parse Reranking for Morphologically Rich Languages" paper in the *Acta Polytechnica Hungarica* journal (Szántó and Farkas, 2015).

5 Dependency Parsing

Chapter 4 delves into the current landscape of Hungarian dependency parsing. It introduces the available datasets, explores the relationship between dependency parsing and constituent parsing in Hungarian (Simkó et al., 2014), and examines the development of the Universal Dependency dataset for Hungarian (Vincze et al., 2017). Additionally, it highlights HuSpaCy, which provides a cutting-edge dependency parser for Hungarian (Orosz et al., 2023).

In Section 4.2, I introduce a Hungarian constituency to a dependency converter. Based on that system I demonstrate that although the results obtained by training on the constituency treebank and converting the output to dependency format and those obtained by training on the automatically converted dependency treebank are similar in terms of accuracy scores, the typical errors made by different systems differ from each other. These results were published in the article *An Empirical Evaluation of Automatic Conversion from Constituency to Dependency in Hungarian* at the COLING 2014 conference (Simkó et al., 2014).

In Section 4.3, I presented how the principles of Universal Dependencies and Morphology have been adapted to Hungarian. Experiments were introduced on the new, manually annotated corpus for evaluating automatic conversion and the added value of language-specific, i.e. non-universal, annotations. This work was published at the EACL 2017 conference with the title *Universal Dependencies and Morphology for Hungarian - and on the Price of Universality* (Vincze et al., 2017).

In Section 4.4, I introduced the HuSpaCy a new industrial-grade text processing pipeline for Hungarian and presented a thorough evaluation showing their (close to) state-of-the-art performance. This work was published in the paper *Advancing Hungarian Text Processing with HuS-*

paCy: Efficient and Accurate NLP Pipelines at the TSD 2023 conference (Orosz et al., 2023).

Contribution

In the Simkó et al. (2014) and Vincze et al. (2017) I made the following contributions:

- I implemented the constituency to dependency and the Universal Dependencies converters.
- I provided statistical support to linguists to develop the rules.
- I did the Machine Learning experiments in both works to compare various representations.

In the Orosz et al. (2023), my contribution was the proposal and experiments on different dependency parsers.

6 Application Of Syntax Parsing

6.1 Latent Syntactic Structure-Based Sentiment Analysis

People publicly share their opinions using social media on a variety of topics, like products and political issues. The task of sentiment analysis (SA) is to automatically extract opinions from textual content. Most of the SA systems assign polarity labels (e.g. positive, negative, and neutral) to textual elements like documents and sentences. The basic solution for SA is to represent the texts in a bag-of-word model and train supervised classifiers or/and employ polarity lexicons for polarity classification (Ravi and Ravi, 2015).

Previous studies have been investigating the utilization opportunities of the syntactic structure of sentences for enhancing sentiment analyzers. Most of these proposals use hand-crafted rules based on the syntactic parse of the sentence (Vilares et al., 2013). These rules are engineered to address certain restricted sets of in-sentence SA's challenges, like negation and intensification.

In the Stanford Sentiment Treebank (Socher et al., 2013) a polarity label was manually assigned to each constituent of the sentence's phrase structure parse. This treebank can be utilized as a training dataset for statistical structure prediction methods and it introduces the opportunity of exploiting the syntactic structure of sentences without restricting the models to a closed set of language phenomena (like negation and intensifiers), neither demands the direct modeling of those phenomena. It enables the application of supervised machine learning techniques to model how morphosyntactic and lexical structures alter the polarity of a constituent. On the other hand, the supervised approach has the disadvantage of requiring a manually annotated treebank. This treebank is domain-dependent, i.e. sentiment analyzers trained on it work fine only on movie reviews and the annotation of new treebanks for other domains is expensive.

In Chapter 5, I focus on the exploitation strategies of syntactic structures for in-sentence and sentence-level SA (Hangya et al., 2017). Usually, sentence-level polarity labels can be easily obtained in a huge amount for various domains, take for instance pro/con or bottomline summaries of the product review sites. Hence this chapter proposes a machine learning framework for sentence-level and in-sentence polarity classifiers by using exclusively sentence-level polarity annotation for training. This approach can predict the sentiment labels assigned to the constituents of a phrase structure parse tree without an annotated sentiment treebank by handling the polarity labels of internal nodes in parse trees as latent variables. Figure 1 exemplifies the difference between a fully annotated sentiment tree and the proposed latent representation.

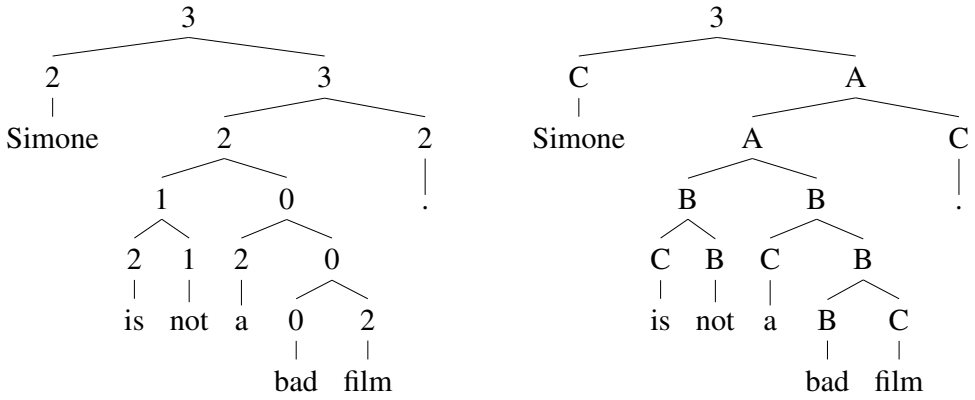

Figure 1: *Representation of sentiment trees in the Stanford Sentiment Treebank (Socher et al., 2013) (left) contains 5-level polarity annotation {`0=very negative`, `4=very positive`} for each node of the binary syntactic tree. On the other, we assume that we have access only sentence-level polarity annotation, i.e. only the label of the root is given (right). Here, the states of the inner nodes are described by latent discrete variables {`A`, `B`, `C`}.*

I introduce two experimental setups for the investigation of the proposed approach. The objective of the experiments' first batch (in Section 5.3) is to investigate whether the sentence-internal latent structure helps the prediction of sentence-level polarity. The second batch of experiments (in Section 5.4) shows that the sentence-internal latent structures themselves are also meaningful when we extract features from them for a target-oriented sentiment analysis task.

The chief added value of this chapter is to propose a latent syntactic structure-based approach that requires only sentence-level polarity labels for training. The experiments on three domains (movies, IT products, restaurants) support that sentiment analyzers are domain-dependent.

This work was published at the *2nd IEEE International Conference on Computational Intelligence and Applications* conference with the *Latent syntactic structure-based sentiment analysis* title (Hangya et al., 2017).

Contribution

In the Hangya et al. (2017) my contribution is:

- Proposal of the sentiment tree representation.
- The latent structured decoder and the training algorithm.

The idea of fine-grained analysis with just sentence annotations cannot be distributed between the coauthors.

6.2 Application of Generalized Syntactic Parsing Framework

In Chapter 6, I introduce the work of Team Szeged for the *First Shared Task on Extrinsic Parser Evaluation* (EPE 2017) (Szántó and Farkas, 2017). I present three approaches to exploit the opportunities of the general dependency graph representation of the shared task.

The goal of the EPE 2017 was to estimate "*the relative utility of different types of dependency representations for a variety of downstream applications that depend heavily on the analysis of grammatical structure*". (Oepen et al., 2017).

To enable different types of dependency representations, the organizers of the shared task introduced a very general graph-based representation of '*relational*' structure reflecting syntactic-semantic analysis. The nodes of this graph correspond to lexical units, and its edges represent labeled directed relations between two nodes. Nodes can be defined in any terms of (in principle arbitrary) sub-strings of the surface form of the input sentence. This representation allows overlapping and empty (i.e. zero-span) node sub-strings as well. Moreover, nodes and edges are labeled by attribute–value maps without any restriction on the attribute set.

This very general graph-based representation opens brand new ways for expressing syntactic or semantic information besides the standard dependency tree formalism. We understood the call of the shared task in a generalized way and came up with ideas that aim to leverage the opportunities of the general representation beyond dependency parse trees. We experimented with a couple of such ideas (instead of trying to achieve high scores in the shared task).

In the first set of experiments (in Section 6.2), we start from the classic dependency parsing approach but instead of a single dependency parse, we express the distribution of possible dependency parses given a sentence in the graph-based general representation. In Section 6.3, I introduce a possible solution for enriching the dependency parse by constituent information given by a standard phrase-structure parser. In this way, various syntactic representations can be represented in the graph and information is not lost because the downstream application can only accept a single dependency parse tree. Furthermore, in the EPE 2017 setting, we can send a blended relational structure to the downstream task, like a parse distribution and blended version of different syntactic approaches, and the downstream application is able to machine learn which type of syntactic structure or phenome or even which combination of syntactic information is useful for itself.

Our last batch of experiments (in Section 6.4) is a consequence of this objective, i.e. the relational representation has to be useful for the downstream application. Here, we tried to automatically recognize which dependency parse labels are useful to a downstream task and collapsed the useless ones.

The results of this chapter were published in the *Proceedings of the 2017 Shared Task on Extrinsic Parser Evaluation at the Fourth International Conference on Dependency Linguistics and the 15th International Conference on Parsing Technologies* with the title *Szeged at EPE 2017: First experiments in a generalized syntactic parsing framework* Szántó and Farkas (2017).

6.3 Enhancing Medication Event Classification with Syntax Parsing

In Chapter 7, I introduce strategies for medication event extraction and classification, revealing how syntax-based information extraction unlocks efficiency gains in the presence of pretrained large language models too (Szántó et al., 2023).

Understanding the complete medication history is necessary for having a fuller picture of the patient, but in many cases, medication-related information is documented only as unstructured clinical notes. This can make it challenging for healthcare providers to obtain a comprehensive view of a patient's medication history including information on medication changes, dosages, and adverse reactions. The automatic analysis of these notes could help medical providers have a fuller background on the patient, better understand the reasons behind medication changes, and identify the healthcare provider who ordered a medication change, as well as the reason for the change.

This would allow for more informed medical decisions and improve patient safety.

The Contextualized Medication Event Dataset (CMED) (Mahajan et al., 2021) and the *National NLP Clinical Challenges 2022 Track 1* aimed at the extraction of these medication events from clinical notes. As well as identifying names of medications, this dataset allows for the detailed analysis of the context of medication-related events. It aims to extract more detailed information from the text about the mentioned medications: like whether the use of the medication was started or stopped, or identifying the person requesting the change. This is a context classification problem where the goal is to find the information that relates to the specific expression, eg. being able to correctly identify if one medicine was started, but another was stopped for a patient within the same note.

Nowadays the most generic approach for this type of problem is using a pre-trained language model and fine-tuning it for our tasks. We applied two main additions to this standardized framework. (Szántó et al., 2023) One of our modifications is aiming to handle the noisy, error-ridden nature of the clinical notes, for this problem we applied adversarial attacks throughout the training process. The other was used for the context classification task and motivated by the state-of-the-art algorithms of aspect-based sentiment analysis as they both aim to identify the context relevant to the selected phrase. We used syntactic relations to help find the more closely related parts of the sentences.

Contribution

In the Szántó et al. (2023) my contribution is:

- The idea of the full system except for the adversarial attack-based solution.
- The idea and the implementation of the syntax-based local context mechanism for context classification.

7 Magyar Nyelvű Összefoglaló

A dolgozat elsősorban a magyar és más morfológiailag gazdag nyelvek szintaktikai elemzésére koncentrál, emelett bemutatja a szintaktikai elemzés lehetséges alkalmazásait olyan magasabb szintű feladatokban, mint például a szentiment elemzés vagy az orvosi dokumentumokban található gyógyszerszedési események kinyerése.

7.1 Konstituens Elemzés

A 3. fejezet a morfológiailag gazdag nyelvek konstituens elemzésének a kihívásaira mutatott be megoldásokat. A konstituens fákban található információk jobb kihasználására ismertetett módszereket a 3.3. rész, míg a 3.4. rész olyan külső forrásból származó információk, mint a nagymennyiségű címkézetlen szövegekes korpuszok felhasználására adott megoldási javaslatokat.

A 3.3. rész egyik fő hozzájárulása, hogy újszerű, automatikus eljárást javasolt a preterminálisok optimális halmazának megtalálására a morfológiai jellemzők értékeinek az összevonásával. A preterminálisok halmazának mérete nagyon fontos a hagyományos környezetfüggetlen nyelvtan alapú megközelítések esetén. Ha csak a főszófajt használjuk preterminálisokként, akkor sok, a szavak morfológiai leírásában kódolt információt veszítünk. Ezzel szemben, ha a teljes morfológiai leírást használjuk a preterminálisok szintjén, akkor több mint ezer különböző elemünk

lesz, aminek a következtében kevés példánk lesz az egyes preterminálisokra és teljesítményproblémákba is ütközünk. Az általam javasolt megközelítés legfőbb újításai a korábbi munkákhoz képest, hogy nagyon gyors – egy valószínűségi kontextusfüggetlen nyelvtanon (PCFG) belül működik, ahelyett, hogy egy feketedobozként használt elemzőt kellene minden jellemző kombináció vizsgálatához újratanítani – ezen felül képes páronként vizsgálni bizonyos morfológiai jellemzők értékeit ahelyett, hogy egy jellemzőt az összes értékével együtt dobna el vagy tartana meg.

Ezen felül a kísérleteimben a lexikai modell javítására külső korpuszok felhasználását is megvizsgáltam. Eredményeim alapján a rendszer teljesítménye tovább javítható szófaji elemzési statisztikák felhasználásával. Az utolsó kísérletem bemutatta, hogy a konstituens elemzésben gyakran használt újrarangsorolási lépés hatékonysága javítható morfológiai információkra építő jellemzők készítésével.

A 3.4. rész olyan megközelítésekre összpontosított, amelyek külső információkat, például nagy mennyiségű címkézetlen adatot és függőségi fákat használnak a konstituens elemző pontosságának javítására. Az újrarangolási lépésben új jellemző készletet vezettem be, ami függőségi elemzés alapú információkat használ. A címkézetlen adatok felett Brown klaszterezést készítettem, amit szintén jellemzőként építettem be az újrarangsoroló rendszerbe.

7.2 Függőségi Elemzés

A 4. fejezet a magyar nyelvű függőségi elemzés jelenlegi helyzetével foglalkozott. Bemutatta a rendelkezésre álló korpuszokat, megvizsgálta a magyar (Simkó et al., 2014) függőségi és konstituens elemzés közötti kapcsolatot, és bemutatta a Universal Dependencies projekt magyar alkorpuszának az elkészültét (Vincze et al., 2017). Emellett ismertette a (közel) state-of-the-art függőségi elemzővel rendelkező, ipari igényekre optimalizált HuSpaCy magyar nyelvű szövegfeldolgozó keretrendszert (Orosz et al., 2023).

A 4.2. részben ismertettem egy rendszert, ami magyar nyelvű konstituens fákat képes függőségi fákká átalakítani. Ezen rendszer alapján demonstráltam, hogy bár a konstituens elemzővel betanított modellből függőségi formátumba konvertált eredmények és az automatikusan konvertált függőségi elemzővel betanított eredmények pontosság szempontjából hasonlóak, a két rendszer által elkövetett tipikus hibák eltérnek egymástól.

A 4.3. részben bemutattam, hogy a Universal Dependencies and Morphology alapelveit hogyan ültettük át a magyar nyelvre. Az új, manuálisan létrehozott korpuszon kísérleteket végeztünk az automatikus átalakítás értékelésére, valamint a nyelvspecifikus, azaz nem univerzális annotációk hozzáadott értékének vizsgálatára.

A 4.4. részben ismertettem a HuSpaCy-t, egy új, ipari célokra kialakított magyar nyelvű szövegfeldolgozó keretrendszert és kísérleteken keresztül annak (közel) state-of-the-art teljesítményét.

7.3 Szintaxis Elemzés Alkalmazásai

7.3.1 Rejtett Szintaktikai Struktúra Alapú Szentiment Elemzés

Az 5. fejezetben szintaktikai struktúrák kiaknázására összpontosítottam a mondatokon belüli és a mondatszintű szentiment elemzésben (Hangya et al., 2017). A mondatszintű szentiment címkék általában könnyen és nagy mennyiségben beszerezhetőek, például a termékismertető oldalakon található értékelések letöltésével. Ezért ebben a fejezetben egy olyan gépi tanulási keretrendszert javasoltam, ami kizárólag mondatszintű szentiment annotációt használ a tanításhoz, de mondaton belüli elemekhez is rendel szentiment címkéket. Ez a megközelítés képes előrejelezni a konstituens

fa csomópontjaihoz rendelt szentiment címkéket annotált szentimenteket tartalmazó fa nélkül, a belső csomópontok szentiment címkéit látens változókként kezelve.

Két kísérleti környezetet alakítottam ki a javasolt megközelítés vizsgálatára. Az első kísérletsorozat (5.3. rész) célja annak vizsgálata, hogy a mondaton belüli látens struktúra segíti-e a mondatszintű szentiment előrejelzését. A 5.4. részben szereplő második kísérletsorozat azt mutatja be, hogy maguk a mondaton belüli látens struktúrák is jelentéssel bírnak, amikor jellemzőket készítünk belőlük egy célorientált szentiment elemzés feladathoz.

7.3.2 EPE: Általános Szintaktikai Keretrendszer Alkalmazása

A 6. fejezetben bemutattam a *First Shared Task on Extrinsic Parser Evaluation* (EPE 2017) versenyen a Team Szeged munkáját (Szántó and Farkas, 2017). Ismertettem három megközelítést a verseny céljául szolgáló általános gráfalapú függőségi reprezentáció kihasználására.

Ahhoz, hogy egyszerre alkalmazni lehessen különböző függőségi reprezentációkat a verseny szervezői egy nagyon általános, gráfalapú reprezentációt vezettek be. A gráf csomópontjai lexikai egységeknek felelnek meg, az élei pedig címkézett irányított kapcsolatokat jelentenek két csomópont között. A csomópontokat a bemeneti mondat egy tetszőleges karakterszámú szakaszával lehet meghatározni. Ez a reprezentáció lehetővé teszi az átfedő és az üres (azaz nulla karakter hosszú) csomópontok létrehozását is. Ezenfelül lehetőség van csomópontokat és az éleket tetszőleges attribútum-érték párokkal címkézni.

Ez a nagyon általános gráfalapú reprezentáció teljesen új lehetőségeket nyit a szintaktikai vagy szemantikai információk kifejezésére a hagyományos függőségi leírás felett. Ennek megfelelően a feladatra olyan ötletekkel álltunk elő, amelyek célja, hogy az általános reprezentáció lehetőségeit a függőségi elemzésen túl is kihasználjuk.

Az első kísérletsorozatban (6.2. rész) a klasszikus függőségi elemzési megközelítésből indultunk ki, de egyetlen függőségi elemzés helyett a lehetséges függőségi elemzések eloszlását határoztuk meg egy adott mondatra, aminek a lekódolására lehetőséget adott a verseny gráfalapú általános reprezentációja. A 6.3. részben bemutattam egy lehetséges megoldást a függőségi elemzés kiegészítésére konstituens elemzésből érkező információk segítségével. Az EPE 2017 verseny struktúrájának köszönhetően ezek a kevert információk egyszerre eljutnak a magasabb szintű feladatokhoz, ahol a gépi tanuló rendszer el tudja dönteni, hogy mely adatok, illetve milyen kombinációik a leghasznosabbak számára.

Az utolsó kísérletsorozatunkban (6.4. rész) megpróbáltuk automatikusan felismerni, hogy mely függőségi elemzési címkék hasznosak a magasabb szintú alkalmazásokhoz, és csak azokat megtartani. A haszontalannak bizonyult címkéket pedig összevontuk egymással.

7.3.3 Gyógyszerszedési Események Osztályozásának Javítása Szintaktikai Elemzéssel

A 7. fejezetben különböző stratégiákat mutattam be orvosi dokumentumokból történő gyógyszerszedési események kinyerésére és osztályozására, demonstrálva, hogy a szintaxis alapú információk felhasználása még nagy nyelvi modellek jelenlétében is javítani tud a rendszer hatékonyságán (Szántó et al., 2023).

A kísérleteink során a Contextualized Medication Event Dataset-et használtuk, amely három feladatot tartalmaz. A gyógyszerkinyerési és eseményosztályozási feladatokhoz egy CRF-BERT alapú megoldást alkalmaztunk, amelyet ellenséges példákkal gazdagítottunk.

Az adatbázisban szereplő harmadik feladat gyógyszerszedési változások osztályozása kontextus alapján. Ehhez szintén egy BERT alapú megoldást valósítottunk meg, amelyet a dokumen-

tumok releváns részének szintaxis alapú súlyozásával bővítettünk. Ennek a megközelítésnek a fő motivációja a szentimentelemzés területéről származik. Az ott már jól bevált *local context focus mechanism* és a szintaktikai elemzésen alapuló súlyozás használata sikeresen javította a gyógyszerszedési változások osztályozásának az eredményeit is.

References

Björne, J., Heimonen, J., Ginter, F., Airola, A., Pahikkala, T., Salakoski, T.: Extracting complex biological events with rich graph-based feature sets. In: Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task. pp. 10–18 (2009)

Hangya, V., Szántó, Z., Farkas, R.: Latent syntactic structure-based sentiment analysis. In: 2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA). pp. 248–254. IEEE (2017)

Johansson, R., Moschitti, A.: Relational features in fine-grained opinion analysis. Computational Linguistics 39(3), 473–509 (2013)

Lapponi, E., Velldal, E., Øvrelid, L., Read, J.: Uio 2: sequence-labeling negation using dependency features. In: * SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012). pp. 319–327 (2012)

Mahajan, D., Liang, J.J., Tsou, C.H.: Toward understanding clinical context of medication change events in clinical narratives. In: AMIA Annual Symposium Proceedings. vol. 2021, p. 833. American Medical Informatics Association (2021)

Oepen, S., Øvrelid, L., Björne, J., Johansson, R., Lapponi, E., Ginter, F., Velldal, E.: The 2017 shared task on extrinsic parser evaluation. towards a reusable community infrastructure. In: Proceedings of the 2017 Shared Task on Extrinsic Parser Evaluation at the Fourth International Conference on Dependency Linguistics and the 15th International Conference on Parsing Technologies. Pisa, Italy. pp. 1–16 (2017)

Orosz, G., Szabó, G., Berkecz, P., Szántó, Z., Farkas, R.: Advancing hungarian text processing with huspacy: Efficient and accurate nlp pipelines. In: International Conference on Text, Speech, and Dialogue. pp. 58–69. Springer (2023)

Ravi, K., Ravi, V.: A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. Knowledge-Based Systems 89, 14–46 (2015)

Simkó, K.I., Vincze, V., Szántó, Zs., Farkas, R.: An Empirical Evaluation of Automatic Conversion from Constituency to Dependency in Hungarian. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. pp. 1392–1401 (2014)

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C.D., Ng, A., Potts, C.: Recursive deep models for semantic compositionality over a sentiment treebank. In: Proc. EMNLP. pp. 1631–1642 (2013)

Szántó, Z., Bánáti, B., Zombori, T.: Enhancing medication event classification with syntax parsing and adversarial learning. In: Maglogiannis, I., Iliadis, L., MacIntyre, J., Dominguez, M. (eds.) Artificial Intelligence Applications and Innovations. pp. 114–124. Springer Nature Switzerland, Cham (2023)

Szántó, Z., Farkas, R.: Special techniques for constituent parsing of morphologically rich languages. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. pp. 135–144. Gothenburg, Sweden (April 2014)

Szántó, Z., Farkas, R.: Constituency parse reranking for morphologically rich languages. Acta Polytechnica Hungarica 12(8) (2015)

Szántó, Z., Farkas, R.: Szeged at epe 2017: First experiments in a generalized syntactic parsing framework. In: Proceedings of the 2017 Shared Task on Extrinsic Parser Evaluation at the Fourth International Conference on Dependency Linguistics and the 15th International Conference on Parsing Technologies. Pisa, Italy. pp. 75–79 (2017)

Vilares, D., Alonso, M.A., Gómez-Rodriguez, C.: A syntactic approach for opinion mining on Spanish reviews. Natural Language Engineering 21(01), 139–163 (2013)

Vincze, V., Simkó, K., Szántó, Z., Farkas, R.: Universal Dependencies and morphology for Hungarian - and on the price of universality. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. pp. 356–365. Association for Computational Linguistics, Valencia, Spain (Apr 2017)

Zeng, B., Yang, H., Xu, R., Zhou, W., Han, X.: Lcf: A local context focus mechanism for aspect-based sentiment classification. Applied Sciences 9(16), 3389 (2019)

Társszerzői nyilatkozat

Kijelentem, hogy ismerem Szántó Zsolt PhD fokozatra pályázó Syntax parsing of morphologically rich languages and its application című disszertációját.

A közösen publikált

- Szántó, Zsolt; Bánáti, Balázs; Zombori, Tamás
  Enhancing Medication Event Classification with Syntax Parsing and Adversarial Learning
  In: Artificial Intelligence Applications and Innovations : 19th IFIP WG 12.5 International

cikkben és a disszertációban is szereplő eredményekről az alábbi nyilatkozatot teszem.
A következő eredményekben a pályázó hozzájárulása volt a meghatározó:

- A teljes rendszer ötlete, kivétel az adversarial learning alapú megoldás
- A szintaxis-alapú súlyozásra építő kontextus osztályozó ötlete és megvalósítása

és ezeket az eredményeket nem használtam fel és a jövőben sem használom fel tudományos fokozat megszerzéséhez.

Dátum: 2024. 05. 22.

.....

Bánáti Balázs

Társszerzői nyilatkozat

Kijelentem, hogy ismerem Szántó Zsolt PhD fokozatra pályázó Syntax parsing of morphologically rich languages and its application című disszertációját.

A közösen publikált

- Orosz, György; Szabó, Gergő; Berkecz, Péter; Szántó, Zsolt; Farkas, Richárd
  Advancing Hungarian Text Processing with HuSpaCy: Efficient and Accurate NLP Pipelines
  LECTURE NOTES IN COMPUTER SCIENCE 14102 pp. 58-69. Paper: Chapter 6 , 12 p. (2023)

cikkben és a disszertációban is szereplő eredményekről az alábbi nyilatkozatot teszem.
A következő eredményekben a pályázó hozzájárulása volt a meghatározó:

- A különböző szintaktikai elemzők javaslata és kísérleti validálása

és ezeket az eredményeket nem használtam fel és a jövőben sem használom fel tudományos fokozat megszerzéséhez.

Dátum: 2024. 05. 22.

.....

Berkecz Péter

Társszerzői nyilatkozat

Kijelentem, hogy ismerem Szántó Zsolt PhD fokozatra pályázó Syntax parsing of morphologically rich languages and its application című disszertációját.

A közösen publikált

- Hangya, Viktor; Szántó, Zsolt; Farkas, Richárd
  Latent Syntactic Structure-Based Sentiment Analysis
  In: 2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA)

cikkben és a disszertációban is szereplő eredményekről az alábbi nyilatkozatot teszem.
A következő eredményekben a pályázó hozzájárulása volt a meghatározó:

- Szentiment fa reprezentációjának kidolgozása
- A szentiment fát előállító gépi tanulás alapú algoritmus kidolgozása

és ezeket az eredményeket nem használtam fel és a jövőben sem használom fel tudományos fokozat megszerzéséhez.

Az alábbiakhoz a hozzájárulásunk oszthatatlan:

- A kutatás alap motivációjának és módszertanának kidolgozása
- Rejtett szintaktikai struktúra alapú aprólékos szentiment elemző alapjainak kidolgozása

Dátum: 2024. 05. 22.

.....

Hangya Viktor

Társszerzői nyilatkozat

Kijelentem, hogy ismerem Szántó Zsolt PhD fokozatra pályázó Syntax parsing of morphologically rich languages and its application című disszertációját.

A közösen publikált

- Orosz, György; Szabó, Gergő; Berkecz, Péter; Szántó, Zsolt; Farkas, Richárd
  Advancing Hungarian Text Processing with HuSpaCy: Efficient and Accurate NLP Pipelines
  LECTURE NOTES IN COMPUTER SCIENCE 14102 pp. 58-69. Paper: Chapter 6 , 12 p. (2023)

cikkben és a disszertációban is szereplő eredményekről az alábbi nyilatkozatot teszem.
A következő eredményekben a pályázó hozzájárulása volt a meghatározó:

- A különböző szintaktikai elemzők javaslata és kísérleti validálása

és ezeket az eredményeket nem használtam fel és a jövőben sem használom fel tudományos fokozat megszerzéséhez.

Dátum: 2024. 05. 22.

.....

Dr. Orosz György

Társszerzői nyilatkozat

Kijelentem, hogy ismerem Szántó Zsolt PhD fokozatra pályázó Syntax parsing of morphologically rich languages and its application című disszertációját.

A közösen publikált

- Simkó, Katalin Ilona; Vincze, Veronika; Szántó, Zsolt; Farkas, Richárd
  An Empirical Evaluation of Automatic Conversion from Constituency to Dependency in Hungarian
  In: Proceedings of COLING 2014 : 25th International Conference on Computational Linguistics

- Vincze, Veronika; Simkó, Katalin; Szántó, Zsolt; Farkas, Richárd
  Universal Dependencies and Morphology for Hungarian - and on the Price of Universality
  In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017) : Volume 1. Long papers

cikkekben és a disszertációban is szereplő eredményekről az alábbi nyilatkozatot teszem.
A következő eredményekben a pályázó hozzájárulása volt a meghatározó:

- A konstituens és dependencia nyelvtanok közötti konvertáló rendszer implementációja
- A Szeged Dependency Treebank és a Universal Dependencies formátuma közötti konvertáló rendszer implementációja
- Mindkét publikáció esetén statisztikák készítése a konvertálási szabályok kidolgozásának a támogatására
- Mindkét publikáció esetén a gépi tanulási kísérletek

és ezeket az eredményeket nem használtam fel és a jövőben sem használom fel tudományos fokozat megszerzéséhez.

Dátum: 2024. 05. 22.

Simkó Katalin Ilona

.....

Simkó Katalin Ilona

Társszerzői nyilatkozat

Kijelentem, hogy ismerem Szántó Zsolt PhD fokozatra pályázó Syntax parsing of morphologically rich languages and its application című disszertációját.

A közösen publikált

- Orosz, György; Szabó, Gergő; Berkecz, Péter; Szántó, Zsolt; Farkas, Richárd
  Advancing Hungarian Text Processing with HuSpaCy: Efficient and Accurate NLP Pipelines
  LECTURE NOTES IN COMPUTER SCIENCE 14102 pp. 58-69. Paper: Chapter 6 , 12 p. (2023)

cikkben és a disszertációban is szereplő eredményekről az alábbi nyilatkozatot teszem.
A következő eredményekben a pályázó hozzájárulása volt a meghatározó:

- A különböző szintaktikai elemzők javaslata és kísérleti validálása

és ezeket az eredményeket nem használtam fel és a jövőben sem használom fel tudományos fokozat megszerzéséhez.

Dátum: 2024. 05. 22.

Szabó Gergő

Szabó Gergő

Társszerzői nyilatkozat

Kijelentem, hogy ismerem Szántó Zsolt PhD fokozatra pályázó Syntax parsing of morphologically rich languages and its application című disszertációját.

A közösen publikált

- Simkó, Katalin Ilona; Vincze, Veronika; Szántó, Zsolt; Farkas, Richárd
  An Empirical Evaluation of Automatic Conversion from Constituency to Dependency in Hungarian
  In: Proceedings of COLING 2014 : 25th International Conference on Computational Linguistics

- Vincze, Veronika; Simkó, Katalin; Szántó, Zsolt; Farkas, Richárd
  Universal Dependencies and Morphology for Hungarian - and on the Price of Universality
  In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017) : Volume 1. Long papers

cikkekben és a disszertációban is szereplő eredményekről az alábbi nyilatkozatot teszem.
A következő eredményekben a pályázó hozzájárulása volt a meghatározó:

- A konstituens és dependencia nyelvtanok közötti konvertáló rendszer implementációja
- A Szeged Dependency Treebank és a Universal Dependencies formátuma közötti konvertáló rendszer implementációja
- Mindkét publikáció esetén statisztikák készítése a konvertálási szabályok kidolgozásának a támogatására
- Mindkét publikáció esetén a gépi tanulási kísérletek

és ezeket az eredményeket nem használtam fel és a jövőben sem használom fel tudományos fokozat megszerzéséhez.

Dátum: 2024. 05. 22.

.....

Dr. Vincze Veronika

Társszerzői nyilatkozat

Kijelentem, hogy ismerem Szántó Zsolt PhD fokozatra pályázó Syntax parsing of morphologically rich languages and its application című disszertációját.

A közösen publikált

- Szántó, Zsolt; Bánáti, Balázs; Zombori, Tamás
  Enhancing Medication Event Classification with Syntax Parsing and Adversarial Learning
  In: Artificial Intelligence Applications and Innovations : 19th IFIP WG 12.5 International

cikkben és a disszertációban is szereplő eredményekről az alábbi nyilatkozatot teszem.
A következő eredményekben a pályázó hozzájárulása volt a meghatározó:

- A teljes rendszer ötlete, kivétel az adversarial learning alapú megoldás
- A szintaxis-alapú súlyozásra építő kontextus osztályozó ötlete és megvalósítása

és ezeket az eredményeket nem használtam fel és a jövőben sem használom fel tudományos fokozat megszerzéséhez.

Dátum: 2024. 05. 22.

Zombori Tamás

.....

Zombori Tamás