Graph analysis of the dependency-based lexical structures

Benedikt Perak
Faculty of Humanities and Social Studies, University of Rijeka
Sveučilišna avenija 4, Rijeka, Croatia
bperak@uniri.hr

Relevant UniDive working groups: WG2, WG3

1 Introduction

The sequence of lexical items within a linguistic structure is organised by a set of syntactic relations construing a conceptual relation and representing an emergent semantic structure or a meaning. According to the Universal Dependencies (UD)[1] (Nivre et al., 2016, 2020; de Marneffe et al., 2021), a framework for morphosyntactic annotation of human language, which to date has been used to create treebanks for more than 100 languages, the classification of syntactic relations offers a linguistic representation that is useful for morphosyntactic research, semantic interpretation, and for practical natural language processing across different human languages.

A syntactic dependency, in general, is a binary phrasal asymmetric grammar relation with a lexical head and other lexical items as dependents of that head, represented in diagrams by an arrow from the headword to the dependent word. Syntactic dependencies are typically used in natural language processing (NLP) to represent the grammatical structure of sentences. This structure can be described as a graph, where each word is defined as a node, and the relationships between words are represented as edges connecting the nodes. The edges in the graph reflect the grammatical relationships between words, such as subject-verb or modifier-head relationships.

The paper aims to highlight the power of Universal Dependencies (UD) relations in uncovering the complex interplay of semantic and syntactic aspects in language through the use of ConGraCNet methodology and its accompanying web app. The ConGraCNet syntactic-semantic methodology (Perak and Kirigin, 2022) has been developed into a web application[2]. The app utilizes various computational tools and technologies, including tagged corpus data obtained from the Sketch Engine (https://app.sketchengine.eu/), and the Croatian academic VPS-SRCE service for data storage, processing, and visualization. The app also uses Python-based data processing and enrichment programs, including ConGraCNet graph algorithms, WordNet definitions, the Open Multilingual Wordnet, and WordNet domains lexicon. These tools and technologies work together to integrate data structures and provide a comprehensive representation of the syntactic-semantic hierarchy of UD dependencies.

[1] https://universaldependencies.org
[2] http://emocnet.uniri.hr

2 Syntactic dependencies in ConGraCNet

The ConGraCNet methodology is a syntactic-semantic approach for the analysis of the text that represents the cognitive hierarchy of UD dependencies as a graph structure. It uses computational tools and technologies such as tagged corpus data, graph algorithms, as well as WordNet synsets, and Sentiment Dictionaries to extract and integrate data structures that reflect the relationships between words and concepts.

The ConGraCNet presents a novel approach to resolving semantic tasks by leveraging various dependency relations in its fundamental graph structure. The approach contrasts with the conventional employment of word embeddings, which entail representing words as high-dimensional vectors in a continuous vector space. The ConGraCNet's dependency graph approach offers a novel perspective on semantic similarity identification, with potential ramifications for enhancing the precision and resilience of natural language processing tasks. Furthermore, the network's architectural design may serve as a framework for developing more advanced algorithms that integrate a broader spectrum of linguistic attributes and word relationships.

While both sequential distribution and graph distribution approaches aim to represent the meaning of words and their relationships, the ConGraCNet methodology is specifically focused on UD dependencies, utilizing a graph-based representation.

The ConGraCNet methodology takes a systemic approach to exploring the syntactic-semantic constructions in natural language. This approach is

based on the idea of an emergent ontology of syntactic-semantic constructions, which can be used to categorize the world. The underlying theoretical grounding of the ConGraCNet app recognizes the role of natural language syntactic-semantic patterns in identifying ontological relations between concepts and categories.

Each syntactic-semantic pattern expresses a different ontological function, providing insight into the relationships and connections between different concepts. By leveraging these patterns, the ConGraCNet app builds a hierarchical representation of syntactic-semantic constructions, capturing the categorization of the world and the ontological relationships between concepts. This allows for a more complete and nuanced understanding of the relationships and connections between concepts in natural language.

The dependencies can be classified based on their emergent cognitive properties, which are represented through a hierarchical ontological schema of semantic roles and agent-based representations (Perak, 2017). By formalizing these dependencies into multi-layered structures extracted from tagged corpora, it is possible to explore the possibility of using graph representations to capture common knowledge of conceptual entities, attributes, and processes.

For instance, the "conj" dependency in the Universal Dependencies (UD) annotation scheme refers to a type of grammatical relationship that exists between words in a sentence, where two or more words are linked together to form a single syntactic unit by means of their logical conjunction. In the sentence "I like pizza and pasta," the words "pizza" and "pasta" are linked by the conjunction "and," and this relationship is labeled as a "conj" relationship in the UD annotation. Similarly, in the sentence "She is both smart and beautiful," the words "smart" and "beautiful" are linked by the conjunction "and," and this relationship is also labeled as a "conj" relationship. The "conj" label is used to reflect the grammatical relationship between words in a sentence, and it provides important information about the meaning and structure of the concepts in the sentence. In this sense, the "conj" dependency can be seen as a way to structure the underlying network of semantically related concepts in a sentence, by indicating the relationships between words. By extracting the conj collocates is possible to introduce network and graph methods to capture the underlying semantic similarity and create a layer of "conj" related concepts. In NLP, network and graph-based methods are used to model relationships between words in a sentence, and they can be used to capture the underlying semantic similarity between concepts. In our previous work we have shown how "conj" dependency can be represented as the underlying network of semantically related concepts, much similar to the word embeddings.

The "conj" relationship is an edge between two concepts in a graph, where the concepts are represented as nodes. The weight of the edge is be based on the strength of the semantic similarity between the concepts, which is determined using a variety of corpus collocation techniques, such as word frequency or logDice methods.

By applying the ConGraCNet methodology on the corpus-based conj dependency structures it is possible to develop various lexical tasks related to understanding the semantic and syntactic relationships between lexemes.

1. Identifying conceptually similar lexemes using the "conj" relationship.(Perak and Kirigin, 2022)

2. Clustering conceptually similar lexemes into semantic domains by identifying clusters of semantically related concepts.(Perak and Ban Kirigin, 2020)

3. Identifying polysemic structures of a lexeme by analyzing the relationships between different meanings of a lexeme.(Ban Kirigin et al., 2021)

4. Identifying antonymicity by analysing statistical co-occurrence patterns of antonym pairs, defined by WordNet or other lexicons, in clusters of graph lexical structures.

5. Identifying category label for a semantic domain by analyzing the hypernym or categorial relationships of the concepts in the domain.(Ban Kirigin et al., 2021)

6. Propagating sentiment values and properties on a set of semantically similar concepts from a sparse sentiment dictionaries.(Kirigin et al., 2021)

7. Calculating sentiment value for sentiment domains: By analyzing the relationships between concepts in a sentiment domain, the ConGraCNet methodology can be used to calculate the sentiment value for the domain.(Ban Kirigin et al., 2022)

By providing a graph-based representation of the relationships between concepts in text, the ConGraCNet app can facilitate the development of

more advanced NLP systems that are better able to understand and process natural language data.

Future research could build on this work by employing deep graph learning techniques to enhance the performance of the ConGraCNet methodology. By incorporating more complex graph structures and deep neural network architectures, we may be able to further improve the accuracy and efficiency of natural language processing systems. These efforts could have significant implications for a range of practical applications, such as text generation, text classification, sentiment analysis and machine translation. As such, the ConGraCNet methodology represents a promising direction for advancing the field of computational linguistics and paving the way for more sophisticated natural language processing systems.

References

Tajana Ban Kirigin, Sanda Bujačić Babić, and Benedikt Perak. 2021. Lexical sense labeling and sentiment potential analysis using corpus-based dependency graph. *Mathematics*, 9(12):1449.

Tajana Ban Kirigin, Sanda Bujačić Babić, and Benedikt Perak. 2022. Semi-local integration measure of node importance. *Mathematics*, 10(3):405.

Marie de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics*, 47:255–308.

Tajana Ban Kirigin, Sanda Bujacic Babic, and Benedikt Perak. 2021. Building a sentiment dictionary for croatian. *Logic and Applications LAP 2021*, page 67.

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)*, pages 1659–1666.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Dan Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC)*, pages 4034–4043.

Benedikt Perak. 2017. Conceptualisation of the emotion terms: Structuring, categorisation, metonymic and metaphoric processes within multi-layered graph representation of the syntactic and semantic analysis of corpus data. *Cognitive Modelling in Language and Discourse across Cultures; Cambridge Scholars Publishing: Newcastle upon Tyne, UK*, pages 299–319.

Benedikt Perak and Tajana Ban Kirigin. 2020. Corpus-based syntactic-semantic graph analysis: Semantic domains of the concept feeling. *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje*, 46(2):957–996.

Benedikt Perak and Tajana Ban Kirigin. 2022. Construction grammar conceptual network: Coordination-based graph method for semantic association analysis. *Natural Language Engineering*, pages 1–31.