

Streszczenie

Przedmiotem niniejszej rozprawy jest analiza składniowa języka naturalnego, nazywana również parsowaniem składniowym. Problem ten polega na przypisaniu zadanemu fragmentowi tekstu formalnej reprezentacji opisującej jego budowę gramatyczną. Jednostkami tekstu poddawanymi analizie najczęściej są zdania lub — ogólniej — wypowiedzenia, a struktura stanowiąca jej wynik to rozbiór składniowy. Analiza składniowa stanowi jedno z podstawowych zagadnień w dziedzinie przetwarzania języka naturalnego, a jej przeprowadzenie jest niezbędne lub bardzo przydatne w wielu zastosowaniach.

Dwoma powszechnie wykorzystywanymi formalizmami reprezentowania rozbioru są drzewa składnikowe oraz zależnościowe. Metoda automatycznej analizy składniowej stanowiąca przedmiot niniejszej rozprawy w systematycznie spójny sposób łączy dwa wspomniane sposoby przedstawiania rozbioru w jedno, hybrydowe drzewo składnikowo-zależnościowe. Podejście takie pozwala na elastyczne wykorzystanie zalet obu reprezentacji, a jednocześnie nie jest szeroko zbadane i stosowane, szczególnie w przetwarzaniu tekstów w języku polskim.

Dla opracowanej przeze mnie nowej techniki parsowania do postaci hybrydowej kluczowe są elementy struktury drzewiastej, które nazywam kręgosłupami składniowymi. Kręgosłupy to swego rodzaju lokalne osie, wokół których zbudowane są fragmenty wynikowego drzewa. Stosując podejście typu „parsowanie przez tagowanie” (*parsing as tagging*), traktuję je jako etykiety (znaczniki), które należy przypisać poszczególnym słowom (segmentom) analizowanego zdania. Aby wyznaczyć sposób, w jaki na bazie kręgosłupów należy zbudować kompletną strukturę hybrydową, wykorzystuję z kolei grafową metodę parsowania zależnościowego.

W celu przeprowadzenia eksperymentalnej weryfikacji skuteczności proponowanej przeze mnie metody zaprojektowałam oraz zaimplementowałam model statystyczny w postaci głębokiej sieci neuronowej, w której skład wchodzi pretrenowany model językowy typu BERT. Model ten wykonuje wszystkie obliczenia konieczne do skonstruowania drzewa hybrydowego zgodnie z moim algorytmem. W ramach dodatkowych badań rozszerzam architekturę parsera o komponenty wykonujące

inne zadania z zakresu przetwarzania tekstu: lematyzację, analizę fleksyjną oraz rozpoznawanie jednostek nazewniczych.

Wyniki wykonanej przeze mnie ewaluacji wykazują, że parser działający zgodnie z przedstawioną metodą ma skuteczność na poziomie najlepszych osiągniętych aktualnie wyników (*state-of-the-art*). W moich eksperymentach skupiam się przede wszystkim na tekstach polskich (zarówno współczesnych, jak i dawnych), ale uwzględniam również dane niemieckojęzyczne. Uzyskana jakość automatycznej analizy składniowej współczesnego języka polskiego, wyrażona powszechnie wykorzystywanymi miarami, to 96% UAS i 90,7% LAS dla wykrywania struktury zależnościowej oraz 97,6% F_1 dla wykrywania składników. Wysokie są również wyniki dla lematyzacji oraz analizy fleksyjnej – F_1 odpowiednio na poziomie około 98,5% oraz 97,5% – a także dla rozpoznawania jednostek nazewniczych – F_1 wynosi w tym przypadku 88%.

Moja implementacja zaprezentowanej metody – parser Hydra – w momencie kończenia tej rozprawy znalazła już praktyczne zastosowania, między innymi w automatycznej anotacji dużego korpusu języka polskiego, wsparciu rozbudowy ręcznie znakowanego banku drzew oraz wykrywaniu neologizmów w tekstach pochodzących z Internetu.

Abstract

The subject of this dissertation is the syntactic analysis of natural language, also known as syntactic parsing. This problem involves assigning a formal representation to a given textual passage, describing its grammatical structure. The text units subjected to analysis are most often sentences or — more generally — utterances, and the resulting structure is called a syntactic parse. Syntactic parsing is one of the fundamental issues in the field of natural language processing, and its results are essential or very useful in many applications.

Two formalisms commonly used for representing a parse are constituency trees and dependency trees. The method of automatic parsing that is the subject of this dissertation systematically and consistently combines the two aforementioned structure types into a single, hybrid constituency-dependency tree. This approach allows for flexible use of the advantages of both representations, but at the same time, it has not been widely researched and applied, particularly in processing of Polish texts.

Central to my new hybrid parsing technique are tree structure elements which I call syntactic spines. These components are a kind of local axes around which fragments of the resulting tree are constructed. Using a parsing-as-tagging approach, I treat them as labels (tags) that should be assigned to individual words (segments) of the analysed sentence. To determine the way a complete hybrid structure should be constructed starting from the spines, I use a graph-based dependency parsing method.

In order to experimentally verify the effectiveness of my proposed method, I designed and implemented a statistical model in the form of a deep neural network, which includes a pre-trained BERT-type language model. This model performs all the calculations necessary to construct the hybrid tree according to my algorithm. As part of additional research, I augmented the parser architecture with components that perform other text processing tasks: lemmatisation, inflectional analysis, and named entity recognition.

The results of my evaluation show that the parser operating according to the presented method achieves state-of-the-art performance. In my experiments, I focus primarily on Polish texts (both contemporary and historical), but I also include German-language data. The achieved quality of automatic syntactic analysis of contemporary Polish, measured by commonly used metrics, is 96% UAS and 90.7% LAS for dependency structure detection, and 97.6% F_1 score for constituent detection. The results for lemmatisation and inflectional analysis are also high – the results being approximately 98.5% and 97.5% F_1 score respectively – as well as for named entity recognition – the F_1 score value in this case is 88%.

My implementation of the presented method – the Hydra parser – had already found practical applications by the time this dissertation was completed, including automatic annotation of a large Polish corpus, support for the expansion of a manually tagged treebank, and detection of neologisms in texts sourced from the Internet.