Sveučilište u Zagrebu

Fakultet organizacije i informatike

Martina Ptiček

SUVREMENI PRISTUPI OBRADI PRIRODNOGA JEZIKA - IDENTIFIKACIJA METAFORA U HRVATSKOM JEZIKU

DOKTORSKI RAD

Varaždin, 2024.

Sveučilište u Zagrebu

Fakultet organizacije i informatike

MARTINA PTIČEK

SUVREMENI PRISTUPI OBRADI PRIRODNOGA JEZIKA - IDENTIFIKACIJA METAFORA U HRVATSKOM JEZIKU

DOKTORSKI RAD

Mentorica:

prof.dr.sc. Jasminka Dobša

Varaždin, 2024.

University of Zagreb

Faculty of Organization and Informatics

Martina Ptiček

CONTEMPORARY APPROACHES TO NATURAL LANGUAGE PROCESSING - METAPHOR IDENTIFICATION IN CROATIAN LANGUAGE

DOCTORAL THESIS

Varaždin, 2024.

SAŽETAK

Predmet istraživanja prikazanog u ovom doktorskom radu pripada području obrade prirodnog jezika i strojnog učenja. Istraživanje se bavi identifikacijom metafora, s obzirom na njihovu rasprostranjenost u ljudskom jeziku te značajnu ulogu u konceptualnom poimanju svijeta. Unutar ovog šireg okvira istražuje se uspješnost kojom metode strojnog učenja, korištenjem neuronskih mreža i jezičnih modela, vrše identifikaciju metafore u korpusu tekstova na hrvatskom jeziku. Za treniranje modela, njihovo testiranje te kasnije i procjenu uspješnosti, izrađen je skup podataka na hrvatskom jeziku s označenim metaforama, koji se sastoji od ukupno 74 teksta iz područja vijesti, kulture te znanosti o književnosti. Skup se sastoji od 3.794 rečenice, odnosno 87.109 riječi, a anotiranje je provedeno temeljem procedure MIPVU (engl. *Metaphor Identification Procedure Vrije Universiteit)* prilagođene hrvatskom jeziku i proširene bilježenjem stupnja konvencionalnosti i primarnosti metafora te je ovo prvi takav skup izrađen za hrvatski jezik. Pripremljeni skup podataka potom je korišten u testiranju modela klasifikacije, u kojima su korišteni jezični modeli trenirani i na hrvatskom jeziku. Rezultati provedenog istraživanja pokazuju da model klasifikacije tokena poduprt lingvističkim teorijama metafore (MelBERT) zajedno s jezičnim modelom CroSloEngual daje dobre rezultate u identifikaciji metafore u hrvatskom jeziku, odnosno s njima je ostvarena vrijednost F1 od 0,5801.

Ključne riječi: anotiranje metafore, MIPVU, identifikacija metafore, hrvatski jezik, neuronske mreže, transformatorska neuronska mreža, jezični modeli

SUMMARY

The subject matter of the survey presented in this paper lies in the field of natural language processing and machine learning. The survey focuses on metaphor identification, acknowledging that metaphors are omnipresent in human language and that they have an important role in conceptual apprehension of the world. Within this broader framework, the success of metaphor identification on a corpus of texts in Croatian language is analysed with the use of neural networks and language models. A dataset consisting of 74 texts in Croatian language encompassing news, culture, and literary science, with annotated metaphors was prepared for the purpose of model training, testing and subsequently their performance evaluation. The scope of the dataset is 3,794 sentences, respectively 87,109 words, and it was annotated with the use of MIPVU procedure (*Metaphor Identification Procedure Vrije Universiteit)* which was adjusted for Croatian language and further upgraded with noting metaphor conventionality and whether a metaphor is primary or not. It is the first dataset of this kind in Croatian language. The prepared dataset was then used for testing classification models which use the language models trained also on Croatian language. The results of this survey show that token classification model based on linguistic theory of metaphor (MelBERT) together with CroSloEngual language model give good results in identifying metaphors in Croatian language, since they accomplish F1 value of 0.5801.

Key words: metaphor annotation, MIPVU, metaphor identification, Croatian language, neural networks, Transformers, language models