**DeftEval**

**Abstract**

Definition extraction has been a popular topic in NLP research for well more than a decade, but has been historically limited to well defined, structured, and narrow conditions. In reality, natural language is complicated, and complicated data requires both complex solutions and data that reflects that reality. The DEFT corpus expands on these cases to include term-definition pairs that cross sentence boundaries, lack explicit definitors, or definition-like verb phrases (e.g. is, means, is defined as, etc.), or appear in non-hypernym structures.

**Introduction**

As the computational linguistics community moves further towards comprehensive natural language understanding, it has become increasingly clear that our methods need to consider scenarios that match a complex linguistic reality. In the case of term-definition pairs, that means exploring how explicit in-text definitions and glosses work in free and semi-structured text, especially those whose term-definition pair span crosses a sentence boundary and those lacking explicit definition phrases. In this paper we present a new corpus of natural language term-definition pairs, as well as a novel schema that can be generally applied for a wide range of domains

Our project must recognize and extract all definitions from a given text. To do that, we adnotate each definition with B.I.O. tags. That means the application will know with those tags that B is for begin, I for inside and O for the end of the definition. Those tags help us to identify special cases when a definition can contain one or more definition, and this incrase the ambiguity.

**Related Work**

Most related work on definition extraction has relied on the idea that definitions can be captured by common “definitor” verb phrases like “means”, “refers to”, and “is”. Early work in the field incorporated rule-based methods that extracted sentences that met this narrow standard (JL Clavens, 2001; Cui and Chua, 2004, 2005; Fahmi and Bouma, 2006; Zhang and Jiang, 2009). While predictable and easily applied, these models subsequently failed to extract sentences that lack these explicit markers. In an effort to expand on the type of phrases used to extract definitions, Cui et al. (2007) used soft pattern matching in a modified HMM (PHMM). More recent work from Espinosa Anke and Schockaert (2018) makes use of a neural approach, which reached state-of-theart performance on the word class lattices (WCL) datasets (Navigli et al., 2010). Even so, these methods require both term and definition to appear in the same sentence and for terms to appear before definitions. Hypernym detection, a related field, has also garnered interest for quite some time (see e.g., Hearst (1992); Snow et al. (2005); Ritter et al. (2009); Shwartz et al. (2017)). Because many hypernym glosses follow the pattern X, such as Y or X is a (type of) Y, this work contains a subset of cases considered for definition extraction. Navigli and Velardi (2010) demonstrated the use of word class lattices for both hypernym detection and definition extraction, and Yin and Roth (2018) proved the effectiveness of including definitions in the training of hypernym detection models. Most work on definition extraction has been applied solely to English datasets, including the WCL dataset mentioned above (Navigli et al., 2010), the ukWaC dataset (Ferraresi et al., 2008), a large crawled dataset of the .uk domain name, and the W00 dataset, a small, expertly annotated corpus introduced by Jin et al. (2013). There does exist a smaller effort for multilingual explorations, including German (Storrer and Wellinghoff, 2006), Portuguese (Del Gaudio and Branco, 2007), and Slavic (Przepiorkowski et al. ´ , 2007), as well as some language-independent approaches (Del Gaudio and Branco, 2009). The vast majority of these approaches are for unstructured text, typically scraped from online sources, as in the ukWaC dataset, though some interest has been given specifically for semi-structured text in legal contracts (see e.g. Curtotti and McCreath (2010) and Winkels and Hoekstra (2012)).

**Subtasks**

DeftEval is split into three subtasks

**Subtask 1: Sentence Classification**

Given a sentence, classify whether or not it contains a definition. This is the traditional definition extraction task.

**Subtask 2: Sequence Labeling**

Label each token with BIO tags according to the corpus' tag specification (see Data page).

**Subtask 3: Relation Classification**

Given the tag sequence labels, label the relations between each tag according to the corpus' relation specification (see Data page).

**References**

Hang Cui, Min-Yen Kan, and Tat-Seng Chua. 2007.Soft pattern matching models for definitional question answering. *ACM Trans. Inf. Syst*., 25(2).

Kan M.Y. Cui, H. and T.S. Chua. 2004. Unsupervised learning of soft patterns for generating definitions from online news. In *Proceedings of WWW*, 90-99.

Kan M.Y. Cui, H. and T.S. Chua. 2005. Generic soft pattern models for definitional question answering. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Micha Marcinczuk Degorski, ukasz and Adam Przepiorkowski. 2008. Definition extraciton using a sequential combination of baseline grammars and machine learning classifiers. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation* (LREC 2008).

Rosa Del Gaudio and Antonio Branco. 2007. Automatic extraction of definitions in portuguese: A rulebased approach. In *Proceedings of the TeMa Workshop*.

Rosa Del Gaudio and Antonio Branco. 2009. ´ Language independent system for definition extraction:First results using learning algorithms. In Proceedings of the 1st Workshop on Definition Extraction,WDE ’09, pages 33–39, Stroudsburg, PA, USA. Association for Computational Linguistics.

Luis Espinosa Anke and Steven Schockaert. 2018. Syntactically aware neural architectures for definition extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 378–385. Association for Computational Linguistics.