Contribution ID: **53** Tür: **Oral Presentation**

YOKUTM: Systematic Compilation and Analysis of Turkish Theses for Dataset Development

20 Aralık 2024 Cuma 15:00 (30 dakika)

The field of Natural Language Processing (NLP) has witnessed a substantial body of research. Historically, linguistic research and language modeling have predominantly relied on strict, rule-based frameworks, such as the ITU Turkish Natural Language Processing Pipeline (Eryiğit, 2014). However, the advent of artificial intelligence has catalyzed a paradigm shift, introducing sophisticated models including BERT (Devlin et al., 2018). The utilization of these statistical models necessitates extensive datasets characterized by clarity, conciseness, and grammatical precision to attain optimal efficacy.

Despite the proliferation of open-source datasets, many are predominantly available only in English and suffer from grammatical flaws and a lack of textual cleanliness. This limitation underscores an urgent need for high-quality, multilingual datasets to bolster the development and training of robust and diverse NLP models.

In response to this necessity, we propose a novel methodology that leverages the freely accessible and comprehensive thesis database of the National Thesis Center (Council of Higher Education of Turkey, 2007). This methodology aims to establish and sustain a large linguistic dataset derived from Turkish academic theses by using fast and robust open-source software solutions, making it accessible for everyone.

This curated dataset is poised to serve as an invaluable resource for researchers and developers within the NLP community, thereby facilitating significant advancements in language models, retrieval-augmented generation frameworks, text summarization tasks, and other AI-driven linguistic applications. We hope that our work not only meets the critical need for high-quality Turkish linguistic data but also sets a benchmark for analogous efforts in other languages, fostering inclusivity and diversity within NLP research.

Keywords: Natural Language Processing, Large Language Models, Retrieval-Augmented Generation, Dataset Creation, Academic Text Processing

Presentation language / Sunum Dili

TR (Türkçe)

Disciplines / Disiplinler

Linguistics / Dilbilim

E-mail / E-posta

efeozyay@ogr.iu.edu.tr

ORCID ID

0009-0005-5894-6098

Institution / Affiliation / Kurum

Istanbul University, Faculty of Letters, Linguistics

Country / Ülke

Turkey

Başlıca yazarlar:: Efe Özyay (Istanbul University)

Sunu yapanlar: Efe Özyay (Istanbul University)

Session Classification: Session 3.1 (Day 3)

Track Classification: Congist'24: Digital Tools and Techniques