Completed work

UniDive

2nd General Meeting

University of Naples "L'Orientale

Naples, Italy, 8-9 February 2024

https://unidive.lisn.upsaclay.fr/

cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Funded by the European Union

Abstractive text summarization datasets, models, and tokenization approaches for Turkish and Hungarian

Batuhan Baykara, Tunga Güngör
Boğaziçi University, Computer Engineering, Istanbul, Turkey

Relevant UniDive working groups: WG1, WG3

Introduction

- Text summarization
  Automatically generating brief, fluent, and salient text from a document
- Two types of summaries (Hahn and Mani, 2000)
  - Extractive text summarization
    Selecting most important sentences/phrases from the document
  - Abstractive text summarization
    Generating a summary reflecting the content of the document

Motivation

- Text summarization works are mostly limited to English
- Turkish and Hungarian possess rich affixation
  Words carry morphological and syntactic information
- Utilizing morphology was shown to be effective (Güngör et al., 2019; Eşref and Can, 2019; Dobrossy et al., 2019; Üstün et al., 2018, Pan et al., 2020)

Contributions

- Two large-scale publicly available summarization datasets for Turkish and Hungarian
- Strong baselines for both datasets
- Comparing pointer-generator model (commonly-used baseline model for summarization) with BERT-based models
- Two morphological tokenization methods
  - SeparateSuffix
  - CombinedSuffix

Related Work

- Turkish text summarization studies are limited to extractive summarization
  - Latent semantic analysis and singular value decomposition (Özsoy et al., 2010)
  - Similarity and frequency based metrics (Çığır et al., 2009)
  - Non-negative matrix factorization (Güran et al., 2011)
  - Semantic information (Güran et al., 2013)
  - Query-based models (Pembe and Güngör, 2008)
- Datasets are limited in size
  - 50 documents (Özsoy et al., 2010)
  - 120 documents (Çığır et al., 2009)
- Hungarian text summarization studies are even less
  - Traditional scoring methods (Beke and Szaszák, 2016)
  - Analyzing error propagation in speech summarization (Ákos Tündik et al., 2019)

Datasets

- Dataset compilation
  - All publicly available newspapers were obtained from Wikipedia
  - 3 news sites were identified for each language
  - Relevant fields were extracted
    URL Author
    Title Source
    Abstract Topic
    Content Tags
    Date of publish
  - Documents with missing values were eliminated

Datasets

	TR-News	HU-News
Training	277,573	211,860
Validation	14,610	11,151
Test	15,379	11,738

Number of documents in datasets

Methodology

- Two models were used
  - Pointer-generator model (See et al., 2007) – baseline model
  - BERT-Transformer model
- Two tokenization methods were used
  - SeparateSuffix
    Root and each suffix are considered as tokens
  - CombinedSuffix
    Root and combined suffixes are considered as tokens

(Example:
Sentence: şampiyon yüzücünün görüntüleri ortaya çıktı
(*the photos of the champion swimmer have been revealed*)
SeparateSuffix: şampiyon yüz #ücü #nün görüntü #ler #i orta #ya çık #tı
CombinedSuffix: şampiyon yüz #ücünün görüntü #leri orta #ya çık #tı)

Experiments and Results

	TR-News			HU-News			
Model	R1	R2	RL	R1	R2	RL	
LEAD-2	31.37	17.91	26.92	24.34	7.87	17.61	Baseline
LEAD-3	28.64	16.21	24.07	23.70	7.78	16.75	
WhiteSpace	31.61	18.55	29.57	22.92	7.69	19.78	1st experiment
Unigram LM	33.38	19.77	31.15	**24.33**	**8.25**	**20.91**	
SeparateSuffix	**34.94**	**20.89**	**32.56**	23.86	8.10	20.53	
CombinedSuffix	33.93	20.07	31.57	23.57	7.97	20.23	
mBERT-uncased	21.70	8.95	18.41	21.88	4.51	17.62	2nd experiment
mBERT-cased	**30.99**	**18.09**	**26.54**	**26.54**	9.72	**19.51**	
BERTurk-uncased-32K	27.40	15.60	23.36	-	-	-	
BERTurk-uncased-128K	26.92	15.25	22.96	-	-	-	
huBERT-uncased	-	-	-	25.40	**10.03**	18.54	

Rouge-1, Rouge-2, and Rouge-L results of pointer-generator models with different tokenizations and BERT models

- 1st experiment:
  - Effects of tokenization methods
  - Pointer-generator model
  - SeparateSuffix outperforms CombinedSuffix
  - Both outperform WhiteSpace tokenization

- 2nd experiment
  - Compares pointer-generator model and BERT-based models
  - mBERT: Multilingual BERT
    BERTurk: Turkish BERT (Schweter, 2020)
    huBERT: Hungarian BERT (Nemeskey, 2020)
  - Multilingual BERT outperforms for both languages

Contact Information:
batuhan.baykara@boun.edu.tr, gungort@boun.edu.tr

https://github.com/batubayk/datasets
https://github.com/batubayk/MorphologicalTokenizers
https://github.com/batubayk/newscrawler