

LITHUANIAN SUMMARIES CORPORA FOR AUTOMATIC TEXT SUMMARIZATION



VYTAUTAS
MAGNUS
UNIVERSITY
Faculty of
Informatics

AUTHORS: Danguolė Kalinauskaitė danguole.kalinauskaite@vdu.lt Milita Songailaitė milita.songailaite@vdu.lt Justinas Juozas Dainauskas justinas.dainauskas@vdu.lt Justina Mandravickaitė justina.mandravickaite@vdu.lt Virginijus Marcinkevičius virginijus.marcinkevicius@mif.vu.lt Tomas Krilavičius tomas.krilavicius@vdu.lt

1 INTRODUCTION

We are creating a dataset where each text has two summaries: one extractive (using sentences from the text) and one abstractive (rewritten in new words). Our corpus IN PROGRESS can be used to train and compare summarization systems, helping improve tools for fast reading, search, recommendation, and decision support.

3 EXTRACTIVE VS. ABSTRACTIVE

Extractive summary: Built by selecting and copying the most important sentences or phrases directly from the original text.
Abstractive summary: Built by paraphrasing and rewriting the main ideas in new words, like a human would explain them.

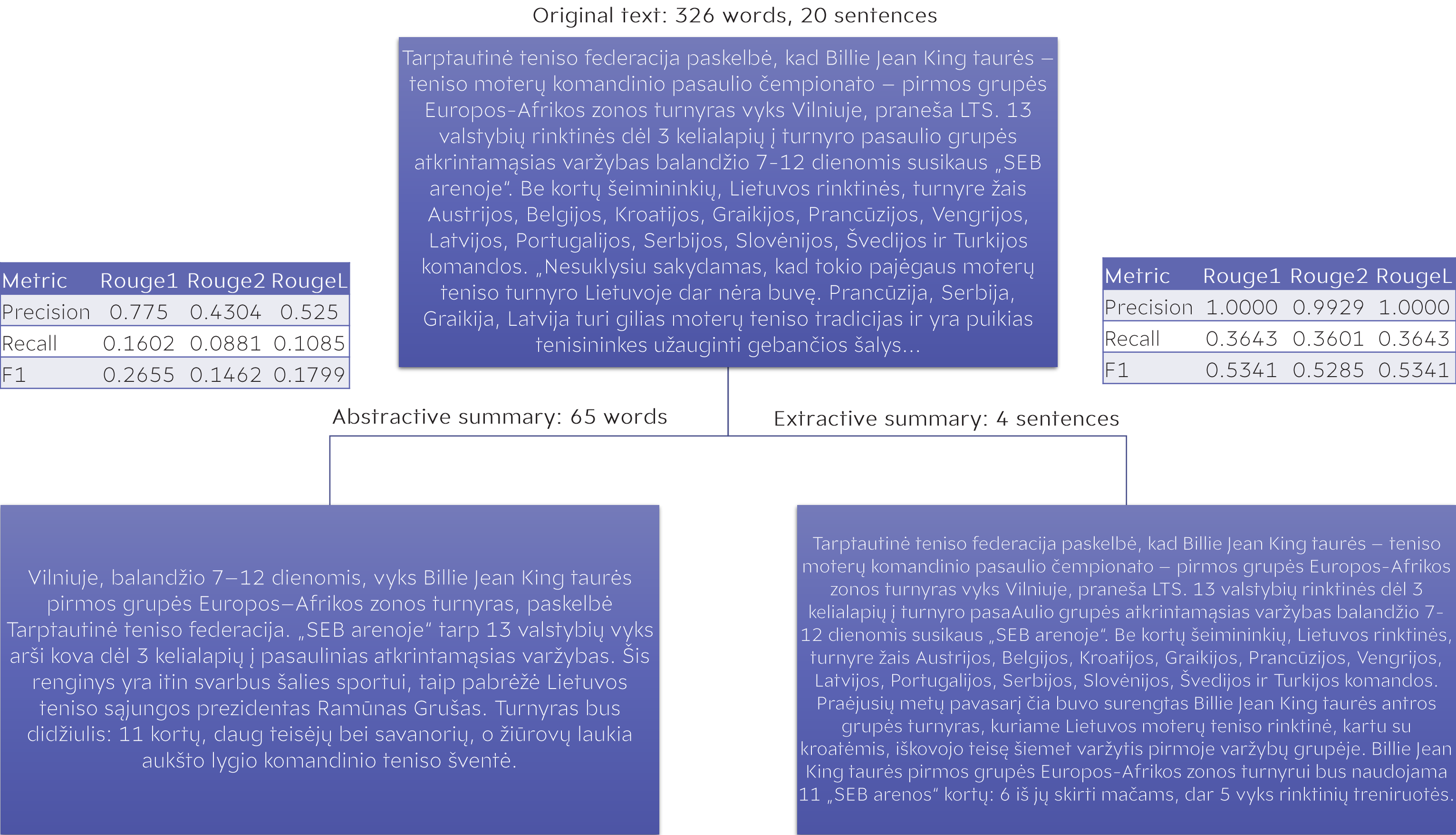
4 ANNOTATION

Extractive summary: Identify key sentences → select the most important ones → keep original wording → ensure logical flow.
Abstract summary: Understand the text → mark main ideas → rewrite in your own words → omit details → maintain coherence.

5 STATISTICS

Type of source	Number of documents	Number of words	Average number of words per document
IT	1 800	809 493	449.72
Health care	700	613 904	877.01
Media	1 500	1 069 725	713.15
Law	785	829 000	1 056.05

6 RESULTS



2 SOURCES

- 1. IT sources:**
- a) IT-related blogs (e.g., <http://technologijos.lt/>)
 - b) Students' bachelor's and master's theses (e.g., <https://www.vdu.lt/cris/home>)
 - c) IT-related VU scientific journals (e.g., <https://www.zurnalai.vu.lt/open-series/en/issue/view/2425>)
- 2. Health care sources: Data prepared by the State Data Agency**
- a) Pharmacy documents
 - b) Doctors' diagnoses
- 3. Media sources:**
- a) News articles from <https://www.lrt.lt/>
- 4. Law sources:**
- a) Lithuanian Courts Information System
 - b) Lithuanian Register of Legal Acts
 - c) Case Law of the Supreme Court of Lithuania
 - d) Legal newsletters (e.g., <https://www.teise.pro/>)
 - e) Scientific papers (e.g., <https://www.elaba.lt/elaba-portal/talpykla>)

Requirements

- **≥ 4000** documents
- The document must be a full-text file of **≥ 300** words.
- The summary must be no less than **10%** of original text.
- Documents must not be older than **2015**.
- **Two types** of abstracts for 1 text: abstractive & extractive

Composition

- media **≥ 1000** documents
- law **≥ 1,000** documents
- medicine **≥ 1,000** documents
- information technology **≥ 1,000** documents

7 FUTURE PLANS

- The corpus will be used:
- For training a model for automatic summarization
 - For analysis of various aspects of summarization in Lithuanian
 - As a resource for comparing summarization tools for Lithuanian

8 ASK US

- What's the difference between a "good" summary and a "bad" one?
- How do we make sure the summaries are correct and not missing important information?
- How long does it take to write an abstractive summary?

ACKNOWLEDGEMENTS

These results are part of the project "Summaries Corpora for the Needs of Artificial Intelligence" implemented from the funds of the Economic Recovery and Resilience Plan "New Generation Lithuania", project No. 02-101-K-0001. We thank Eglė Rimkienė for helping to prepare the presentation of the results.

SustAIN
Liv
Work



Vilnius
University



Funded by
the European Union
NextGenerationEU

CARD
CENTRE
FOR APPLIED
RESEARCH
AND
DEVELOPMENT