Chapter 11
Abstractive Turkish Text Summarization and Cross-Lingual Summarization Using Transformer

Eymen Kagan Taspinar
https://orcid.org/0000-0002-2653-6482
Marmara University, Turkey

Yusuf Burak Yetis
https://orcid.org/0000-0003-0056-7309
Marmara University, Turkey

Onur Cihan
https://orcid.org/0000-0002-5729-2417
Marmara University, Turkey

ABSTRACT

Abstractive summarization aims to comprehend texts semantically and reconstruct them briefly and concisely where the summary may consist of words that do not exist in the original text. This chapter studies the abstractive Turkish text summarization problem by a transformer attention-based mechanism. Moreover, this study examines the differences between transformer architecture and other architectures as well as the attention block, which is the heart of this architecture, in detail. Three summarization datasets were generated from the available text data on various news websites for training abstractive summarization models. It is shown that the trained model has higher or comparable ROUGE scores than existing studies, and the summaries generated by models have better structural properties. English-to-Turkish translation model has been created and used in a cross-lingual summarization model which has a ROUGE score that is comparable to the existing studies. The summarization structure proposed in this study is the first example of cross-lingual English-to-Turkish text summarization.

DOI: 10.4018/978-1-6684-6001-6.ch011

Copyright © 2023, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

INTRODUCTION

Due to the rapid growth of the web, the amount of text data is increasing exponentially which suggests a need for effective techniques and tools to manage this data. Reducing the length of texts while retaining the core meaning, referred to as summarization, has drawn significant attention from researchers in the recent past. There are two main classes of summarization methods: extractive and abstractive. The primary goal of extractive summarization, which uses identical sentences from the original text as part of the summary, is to identify the text's most essential phrases and clauses. In abstractive summarization, however, the aim is to create novel sentences by generating new words or rephrasing the existing ones. To achieve this, the text's semantic content should be examined using deep analysis and reasoning (Rachabathuni, 2017). Abstractive summarization methods provide concise and coherent summaries that are rich in information, short in length, and different from the original text.

Abstractive summarization of English text became a popular research topic thanks to the recent advances in natural language processing (NLP) algorithms. Performance evaluation of the summarization algorithms is done by comparing their ROUGE scores, which is a measure that compares obtained summaries against a reference summary set or translation (Lin, 2004). For English texts, the summarization algorithms in the current literature have ROUGE scores of around 40 which corresponds to a very successful summarization. As a result of the success of summarizing English texts, the authors of this chapter examine the problem of abstractive Turkish text summarization.

BACKGROUND

For abstractive summarization, a number of models utilizing sequence-to-sequence architecture have been presented recently. The transformer model, which exclusively relies on the attention process, was introduced by Vaswani et al. (2017). The attention mechanism was further utilized by the researchers to provide promising results in summarization (Lewis et al., 2019; Raffel et al., 2020). Lewis et al. (2019) proposed the BART model which contains both a bidirectional encoder and an autoregressive decoder. In the BART model, random noise is added to the text data and the original text is reconstructed using a sequence-to-sequence architecture. Raffel et al. (2020) introduced the T5 model which is a text-to-text framework based on an attention mechanism that can be used for various text processing tasks including translation, classification, and summarization. These models are remarkably successful in making sense of sentences since they consist of both the encoder and decoder structures of the Transformer language model, which makes them preferred for translation and summarization problems.

English text summarization problem has been examined by many authors in the literature (Rush et al., 2015; Chopra et al., 2016; Lin et al., 2018). Rush et al. (2015) used a convolutional and attention-based encoder for summarization. Chopra et al. (2016) utilized RNN cells to create a decoder block. Nallapati et al. (2016) suggested an abstractive summarization system for English texts using RNN cells in both encoder and decoder blocks. However, these attention-based structures lead to grammatical errors, semantic irrelevance, and repetition. Lin et al. (2018) provided a solution to this problem using CNN filters and LSTM cells. The studies containing both encoder and decoder structures of the Transformer architecture show higher performances in perceiving text and produce better texts (Raffel et al., 2020; Lewis et al., 2019). Zhang et al. (2019) performed a pre-trained model for English with the C4 corpus that is proposed by Raffel et al. (2020). The fine-tuning stage is performed with ready-to-use datasets

16 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/abstractive-turkish-text-summarization-and-cross-lingual-summarization-using-transformer/314143

Related Content

Mispronunciation Detection and Diagnosis Through a Chatbot
Marcos E. Martinez, Francisco López-Orozco, Karla Olmos-Sánchezand Julia Patricia Sánchez-Solís (2021). *Handbook of Research on Natural Language Processing and Smart Service Systems (pp. 31-45).*
www.irma-international.org/chapter/mispronunciation-detection-and-diagnosis-through-a-chatbot/263095

An Opinion Mining Approach for Drug Reviews in Spanish
Karina Castro-Pérez, José Luis Sánchez-Cervantes, María del Pilar Salas-Zárate, Maritza Bustos-Lópezand Lisbeth Rodríguez-Mazahua (2021). *Handbook of Research on Natural Language Processing and Smart Service Systems (pp. 445-480).*
www.irma-international.org/chapter/an-opinion-mining-approach-for-drug-reviews-in-spanish/263116

Emotional Intelligence in the Digital Workplace key to Enhancing Retention & Reducing Resignations: A Literature Review
Rinki Mishraand Ankita Tewari (2025). *Intersecting Natural Language Processing and FinTech Innovations in Service Marketing (pp. 1-10).*
www.irma-international.org/chapter/emotional-intelligence-in-the-digital-workplace-key-to-enhancing-retention--reducing-resignations/377497

Enhanced Virtual Reality Experience in Personalised Virtual Museums
Chairi Kiourt, Anestis Koutsoudisand Dimitris Kalles (2020). *Natural Language Processing: Concepts, Methodologies, Tools, and Applications (pp. 1348-1366).*
www.irma-international.org/chapter/enhanced-virtual-reality-experience-in-personalised-virtual-museums/239994

Location Extraction to Inform a Spanish-Speaking Community About Traffic Incidents
Alejandro Requejo Flores, Alejandro Ruiz, Ricardo Marand Raúl Porras (2021). *Handbook of Research on Natural Language Processing and Smart Service Systems (pp. 347-367).*
www.irma-international.org/chapter/location-extraction-to-inform-a-spanish-speaking-community-about-traffic-incidents/263110