mCTRLSum: Penggunaan Pretrained Language Generation Model Berbasis Multilingual Pada Abstractive Summarization Terkontrol Menggunakan Keyphrase = mCTRLSum: Utilizing Multilingual Pretrained Language Model For Controllable Abstractive Summarization Using Keyphrase

Sugiri, author

Abstrak

Sebagian besar studi terbaru dalam abstractive summarization melakukan pendekatan dengan melakukan fine-tuning pretrained language generation model (PLGM). PLGM yang digunakan biasanya merupakan versi monolingual, yang hanya memiliki informasi bahasa yang sesuai dengan dataset yang digunakan. Penelitian ini menggunakan PLGM berbasis multilingual, yang menghasilkan kinerja yang cukup kompetitif jika dibandingkan dengan solusi state-of-the-art yang ada. Dengan menggunakan PLGM berbasis multilingual manfaat yang dihasilkan akan berdampak lebih luas sebanyak informasi bahasa yang dimiliki oleh PLGM terkait. Teknik CTRLSum, yaitu penambahan keyphrase di awal source document, terbukti dapat membuat PLGM menghasilkan summary sesuai dengan keyphrase yang disertakan. Penelitian ini menggunakan teknik mCTRLSum, yaitu teknik CTRLSum dengan menggunakan multilingual PLGM. Untuk mendapatkan keyphrase, selain dengan menggunakan teknik keyphrase extraction (KPE) yang memilih kata yang ada di source document, juga digunakan teknik keyphrase generation (KPG) yaitu teknik pembangkitan suatu set kata/frasa berdasarkan suatu source document dataset berbahasa Inggris, tidak hanya dilatih menggunakan oracle keyphrase sebagai pseudo-target dari dataset summarization, model KPG juga dilatih menggunakan dataset khusus permasalahan KPG dengan domain dan bahasa yang sama. Dengan teknik mCTRLSum yang memanfaatkan oracle keyphrase, penelitian ini mendeklarasikan batas atas solusi permasalahan abstractive summarization pada dataset Liputan6, dan XLSum berbahasa Inggris, Indonesia, Spanyol, dan Perancis dengan peningkatan terbesar pada dataset Liputan6 sebanyak 22.54 skor ROUGE-1, 18.36 skor ROUGE-2, 15.81 skor ROUGE-L, dan 7.16 skor BERTScore, dan rata-rata 9.36 skor ROUGE-1, 6.47 skor ROUGE-2, 6.68 skor ROUGE-L dan 3.14 BERTScore pada dataset XLSum yang digunakan pada penelitian ini.

.....Most of the recent studies in abstractive summarization approach by fine-tuning the pre-trained language generation model (PLGM). PLGM used is usually a monolingual version, which only has language information that corresponds to the dataset used. This study uses amultilingual-basedd PLGM, which results in quite competitive performance, compared to existing state-of-the-art solutions. Using a PLGM based on the multilingual benefits generated, it will have a wider impact as much as the language information base owned by the related PLGM. The CTRLSum technique, which is the addition of a keyphrase at the beginning of the source document, is proven to be able to make PLGM produce a summary according to the included keyphrase. This study uses the mCTRLsum technique, namely the CTRLsum technique using multilingual PLGM. To get thekey phrasee, in addition to using the keyphrase extraction (KPE) technique, the words in the source document, keyphrase generation (KPG) techniques are also used, namely the technique of generating a set of words/phrases based on a source document. On the English dataset, not only using the oracle keyphrase as the pseudo-target of the dataset summariza buttion, the KPG model also uses

the dataset specifically for KPG problems with the same domain and language. With the mCTRLsum technique that utilizes the oracle keyphrase, this study declares the upper bound of the solution to the abstractive summarization problem in the Liputan6 and XLSum in English, Indonesian, Spanish, and French datasets with the highest increase in Liputan6 dataset of 22.54 ROUGE-1 score, 18.36 ROUGE-2 score, 15.81 ROUGE-L score, and 7.16 BERTScore, and in average of 9.36 ROUGE-1 score, 6.47 ROUGE-2 score, 6.68 ROUGE-L score, and 3.14 BERTScore on XLSum dataset used in this research.