Peringkasan Lintas Bahasa Berbasis Transformer Menggunakan Multilingual Word Embeddings untuk Domain Bahasa Inggris-Indonesia = Transformer-Based Cross-Lingual Summarization Using Multilingual Word Embeddings for English-Indonesian Domain

Achmad Fatchuttamam Abka, author

Abstrak

Peringkasan lintas bahasa adalah sebuah proses menghasilkan ringkasan dalam bahasa target dari dokumen sumber berbahasa lain. Secara tradisional, peringkasan lintas bahasa dilakukan dalam skema pipeline yang melibatkan dua langkah, yaitu penerjemahan dan peringkasan. Pendekatan ini memiliki masalah, yaitu munculnya error propagation. Untuk mengatasi masalah tersebut, penelitian ini mengusulkan peringkasan lintas bahasa abstraktif end-to-end tanpa secara eksplisit menggunakan mesin penerjemah. Arsitektur peringkasan lintas bahasa yang diusulkan berbasis Transformer yang sudah terbukti memiliki performa baik dalam melakukan text generation. Model peringkasan lintas bahasa dilatih dengan 2-task learning yang merupakan gabungan peringkasan lintas bahasa dan peringkasan satu bahasa. Hal ini dilakukan dengan menambahkan decoder kedua pada Transformer untuk menangani peringkasan satu bahasa, sementara decoder pertama menangani peringkasan lintas bahasa. Pada arsitektur peringkasan lintas bahasa juga ditambahkan komponen multilingual word embeddings. Multilingual word embeddings memetakan kedua bahasa yang berbeda ke dalam ruang vektor yang sama sehingga membantu model dalam memetakan relasi antara input dan output. Hasil eksperimen menunjukkan model usulan mendapatkan kenaikan performa hingga +32,11 ROUGE-1, +24,59 ROUGE-2, +30,97 ROUGE-L untuk peringkasan lintas bahasa dari dokumen sumber berbahasa Inggris ke ringkasan berbahasa Indonesia dan hingga +30,48 ROUGE-1, +27,32 ROUGE-2, +32,99 ROUGE-L untuk peringkasan lintas bahasa dari dokumen sumber berbahasa Indonesia ke ringkasan berbahasa Inggris.

.....Cross-lingual summarization (CLS) is a process of generating summaries in the target language from source documents in other languages. Traditionally, cross-lingual summarization is done in a pipeline scheme that involves two steps, namely translation and summarization. This approach has a problem, it introduces error propagation. To overcome this problem, this study proposes end-to-end abstractive cross-lingual summarization without explicitly using machine translation. The proposed cross-lingual summarization architecture is based on Transformer which has been proven to have good performance in text generation. The cross-lingual summarization model is trained with 2-task learning, which is a combination of cross-lingual summarization and monolingual summarization. This is accomplished by adding a second decoder to handle monolingual summarization, while the first decoder handles cross-lingual summarization. The multilingual word embeddings component is also added to the cross-lingual summarization architecture. Multilingual word embeddings map both different languages into the same vector space so that it helps the model in mapping the relationship between input and output. The experimental results show that the proposed model achieves performance improvements of up to +32.11 ROUGE-1, +24.59 ROUGE-2, +30.97 ROUGE-L for cross-lingual summarization from English source documents to Indonesian summaries and up to +30,48 ROUGE-1, +27.32 ROUGE-2, +32.99 ROUGE-L for cross-lingual summarization from Indonesian source documents to English summaries.