Pemodelan Topik Kontekstual N-Gram Lintas Bahasa untuk Bahasa Indonesia dengan Zero Shot Learning Menggunakan Knowledge Graph = Cross-Lingual Contextualized N-Gram Topic Modeling for Indonesian with Zero Shot Learning Using Knowledge Graph

Jessica Naraiswari Arwidarasti, author

Abstrak

Seringkali kita mengelompokkan dokumen berdasarkan hasil identifikasi topik. Identifikasi topik terhadap sejumlah dokumen tidak terstruktur, contohnya abstrak, dapat dibantu dengan algoritma pemodelan topik. Namun, pelatihan model topik membutuhkan dokumen dengan jumlah yang memadai. Dengan pembelajaran zero shot, kita dapat melakukan prediksi topik terhadap dokumen dengan jumlah yang kurang memadai dengan mentransfer hasil pembelajaran dari dokumen dalam bahasa lain, contohnya Bahasa Inggris, walaupun tidak ada contoh dari bahasa yang diuji (Bahasa Indonesia). Pemanfaatan zero-shot learning sudah dilakukan oleh Bianchi et al. (2021) dengan Contextual Topic Model (CTM). Koherensi topik yang diprediksi CTM dapat ditingkatkan contohnya jika dokumen terkait dengan knowledge graph (KG). Dengan penambahan informasi dari KG, frekuensi kemunculan kata penting menjadi lebih tinggi. Adapun kualitas topik juga dapat ditingkatkan dengan memodifikasi bag-of-word (BoW) kata tunggal menjadi n-gram. Namun, CTM terbatas pada 1-gram. Penelitian ini bertujuan untuk memperkaya topik serta meningkatkan koherensi prediksi topik untuk dokumen unseen dengan memanfaatkan KG dan kualitas topik dengan memodifikasi BoW pada CTM menjadi n-gram. Hasil eksperimen menunjukkan koherensi topik (dalam ukuran NPMI) tertinggi terhadap dokumen Bahasa Inggris yaitu dengan abstrak singkat dan BoW n-gram sebesar 0,24 dengan margin 0.1019 terhadap Bianchi et al.. Namun, prediksi topik terhadap dokumen Bahasa Indonesia memiliki tingkat similaritas yang lebih baik dengan penambahan KG dilihat dari peningkatan nilai Match sebesar 6% untuk 1-gram dan 4.34% untuk n-gram, centroid similarity sebesar 0.02 untuk 1-gram, dan Kullback-Leibler Divergence 0.1 untuk 1-gram dan 0.04 untuk n-gram. Peningkatan kualitas topik juga terjadi dengan modifikasi BoW menjadi n-gram yang ditunjukkan oleh kemunculan topik yang tidak didapatkan sebelum modifikasi BoW. Adapun, model juga dapat memprediksi dokumen dari sumber lain, contohnya berita. Namun, jika topik dokumen tidak tampak pada pelatihan, topik yang diprediksi kurang koheren terhadap dokumen.

.....Often we group documents based on the results of topic identification. Topic identification against a number of unstructured documents, for example abstracts, can be assisted by topic modeling algorithms. However, topic model training requires a sufficient number of documents. With zero shot learning, we can predict the topic of an inadequate number of documents by transferring learning outcomes from documents in other languages, for example English, even though there are no examples from the tested language (Indonesian). The use of zero-shot learning has been carried out by Bianchi et al. (2021) with the Contextual Topic Model (CTM). The coherence of topics predicted by CTM can be improved, for example if the document is related to a knowledge graph (KG). With the addition of information from KG, the frequency of occurrence of important words becomes higher. The topic quality can also be improved by modifying the single word bag-of-word (BoW) into n-grams. However, CTM is limited to 1-gram. This study aims to enrich the topic and improve the coherence of topic prediction for unseen documents by utilizing KG and

topic quality by modifying BoW on CTM to n-grams. The experimental results show the highest topic coherence (in terms of NPMI) to English documents with a short abstract and a BoW n-gram of 0.24 with a margin of 0.1019 to Bianchi et al.. However, topic predictions for Indonesian documents have a better level of similarity with the addition of KG seen from the increase in the Match value by 6% for 1-gram and 4.34% for n-gram, centroid similarity of 0.02 for 1-gram, and Kullback-Leibler Divergence 0.1 for 1-gram and 0.04 for n-gram. An increase in the quality of the topic also occurs with the modification of BoW to n-grams which is indicated by the appearance of topics that are not obtained before the BoW modification. Meanwhile, the model can also predict documents from other sources, for example news. However, if the topic of the document does not appear in the training, the predicted topic is less coherent with the document.