INTISARI

Berbagai karakteristik linguistik yang muncul dari proses interaksi antara bahasa dan struktur sosial seiring berjalannya waktu menciptakan variasi dalam langgam bahasa yang berbeda-beda. Ciri kebahasaan ini mencakup penggunaan ragam bahasa yang formal, pemilihan kata yang sopan, pengekspresian sentimen, keefektifan dalam keringkasan, penyesuaian nada dan suasana, serta penggunaan fitur-fitur unik dalam penyampaian gagasan dan informasi melalui bahasa. Bahasa merupakan entitas yang fleksibel dan dapat menyesuaikan diri dengan latar waktu, tempat, situasi, dan kondisi yang berbeda. Variasi dalam bahasa dapat terjadi sesuai dengan konteks komunikasi, mulai dari percakapan sehari-hari yang cenderung takformal hingga penulisan formal dalam lingkungan akademik atau profesional.

Penelitian ini berfokus pada pengembangan dan peningkatan model alih ragam formalitas teks dalam bahasa Indonesia menggunakan arsitektur Transformer dan metode prapelatihan dengan data sintesis. Data sintesis dihasilkan dari sumber seperti Wikipedia, korpus berita Leipzig, dan artikel ilmiah dari jurnal-jurnal UGM, dengan fokus pada pembangkitan himpunan kerancuan yang mencakup leksikon ungkapan informal-formal, token-lema, token-kelas kata, sinonim WordNet, dan pemeriksa ejaan. Hasil prapelatihan menunjukkan skor BLEU yang tinggi dan perpleksitas yang rendah, menandakan model telah mempelajari ciri-ciri bahasa Indonesia secara efektif. Melalui proses pemelajaran transfer, model ini kemudian berhasil mencapai skor BLEU tertinggi pada himpunan data uji STIF, yaitu **56.93**, mengungguli metode sebelumnya dan menegaskan pentingnya data prapelatihan yang relevan.

Penelitian ini menggarisbawahi pentingnya memilih data prapelatihan yang relevan dan spesifik untuk tugas dalam pemrosesan bahasa alami. Meski berhasil mencapai skor BLEU yang tinggi, model ini masih memiliki ruang untuk peningkatan, khususnya dalam menangani campur kode dan istilah asing. Penelitian lanjutan dapat fokus pada pengembangan metode pembangkitan data sintesis yang lebih canggih, integrasi campur kode dan istilah asing, eksplorasi arsitektur model lain, dan penerapan model pada konteks penggunaan bahasa Indonesia yang lebih luas.

Kata kunci : alih ragam teks, cipta bahasa alami, pemrosesan bahasa alami, data sintesis, Transformer

ABSTRACT

The various linguistic characteristics that emerge from the interaction between language and social structure over time create variations in language styles. These linguistic features include the use of formal language, polite word choices, expression of sentiments, effectiveness in brevity, adjustment of tone and mood, and the use of unique features in conveying ideas and information through language. Language is a flexible entity that can adapt to different times, places, situations, and conditions. Language variation can occur according to the context of communication, ranging from informal daily conversations to formal writing in academic or professional environments.

This research focuses on the development and improvement of a model for text formality style transfer in Indonesian using the Transformer architecture and pretraining methods with synthetic data. The synthetic data is generated from sources such as Wikipedia, the Leipzig news corpus, and UGM scientific articles, focusing on generating a confusion set that includes lexicon of informal-formal expressions, token-lemma, token-Part of Speech (PoS), WordNet synonyms, and spell checker. The pretraining results show a high BLEU score and low perplexity, indicating that the model has effectively learned the characteristics of the Indonesian language. Leveraging transfer learning process, the model further achieved the highest BLEU score on the STIF test data set, a score of ***56.93****, surpassing previous methods and emphasizing the importance of relevant pretraining data.*

This study highlights the importance of choosing relevant and task-specific pretraining data for tasks in natural language processing. Although successful in achieving high BLEU scores, the model still has room for improvement, particularly in handling code-switching and foreign terms. Future research can focus on developing more sophisticated methods for generating synthetic data, integrating code-switching and foreign terms, exploring other model architectures, and applying the model to a broader context of Indonesian language use.

Keywords : text style transfer, natural language generation, natural language processing, synthetic data, Transformer