Pembangunan Dataset Parafrasa Bahasa Indonesia untuk Sistem Pembangkit Parafrasa Otomatis

Ryan Abdurohman[1], Arie Ardiyanti Suryani[2]

[1,2]Fakultas Informatika, Universitas Telkom, Bandung
[1]ryanabdurohman@student.telkomuniversity.ac.id, [2]ardiyanti@telkomuniversity.ac.id,

Abstrak

Parafrasa dapat diartikan sebagai pengungkapan suatu teks dengan diksi yang berbeda tapi merujuk pada makna yang sama. Sistem yang dapat membangkitkan parafrasa secara otomatis memiliki peran yang sangat penting pada Natural Language Processing (NLP). Pada penelitian sebelumnya, dataset pararfrasa yang dihasilkan diekstrak menggunakan mesin penerjemah dengan asumsi pasangan teks sudah pasti memiliki kesamaan semantik. Sehingga, filter yang digunakan hanya pada perbedaan ragam diksi. Akibatnya, dataset yang dihasilkan cenderung kurang memuaskan dalam hal keragaman leksikal dan kesamaan semantik. Oleh karena itu, penelitian ini bertujuan untuk meng-*generate* dataset parafrasa dengan memanfaatkan *task* selain mesin penerjemah yaitu Abstractive Summarization pada dataset Liputan6. Hasil ringkasan manusia yang ada dalam dataset Liputan6 akan dipasangkan dengan teks hasil ringkasan sistem. Setelah itu, pasangan teks akan difilter berdasarkan rata-rata dari kesamaan semantik menggunakan BERTScore dan keragaman leksikal menggunakan inverseSacreBLEU. Dataset yang dihasilkan kemudian dievaluasi untuk dijadikan data latih pada pembangkit parafrasa serta dievaluasi pula secara manual oleh manusia. Proses filtrasi yang digunakan terbukti berhasil meningkatkan keragaman leksikal dibanding penelitian sebelumnya yang ditunjukkan peningkatan skor inverseSacreBLEU dari 57,42 ke 72,76. Adapun dataset yang dihasilkan dari liputan6 (146.030 data) memiliki jumlah hampir 40 kali lipat lebih kecil dari penelitian sebelumnya (5.753.296 data), tapi memiliki skor kesamaan semantik dan keragaman leksikal yang lebih tinggi dengan peningkatan sebanyak 1-2 poin skor. Hal ini menunjukkan kualitas dataset yang dihasilkan lebih baik dari penelitian sebelumnya.

Kata kunci : pembangkit parafrasa, kesamaan semantik, keragaman leksikal