1. Pendahuluan

Parafrasa bisa diartikan sebagai pengungkapan kembali suatu teks dengan ekspresi atau diksi yang berbeda, tapi merujuk pada makna yang sama. Pembangkit parafrasa merupakan task yang sangat penting dan memiliki implementasi luas pada Natural Language Processing (NLP), misalnya Question Answering [5, 31, 28, 7], Machine Translation [20, 24], Question Generation [12, 25], dan Data Augmentation [16, 8]. Bahkan, pembangkit parafrasa juga berpotensi untuk diterapkan dalam penulisan ulang teks yang bersifat teknis menjadi mudah dipahami, misalnya pada teks-teks medis [4].

Sampai saat penelitian ini ditulis, dataset parafrasa bahasa Indonesia dengan jumlah yang cukup banyak dan bisa diakses oleh umum hanya ParaCotta [1]. Pembangunan dataset parafrasa bahasa Indonesia secara manual tentu memerlukan biaya yang mahal. Berdasarkan penelitian [30], terdapat rekomendasi untuk memanfaatkan *task* serta dataset NLP lain untuk membangun sistem pembangkit parafrasa otomatis yang lebih baik. Seiring dengan perkembangan riset NLP bahasa Indonesia yang mengalami kemajuan signifikan pada beberapa tahun terakhir [2], terdapat riset yang berisi temuan menarik yaitu penelitian mengenai pembangunan dataset Abstractive Summarization [14]. Riset ini menyatakan bahwa sebagian besar pasangan dataset yang dibangun cenderung parafrasa. Hal ini didasari oleh hasil evaluasi manual terhadap data uji yang digunakan. Pada penelitian [14], pasangan yang dimaksud adalah hasil ringkasan manual dan otomatis (menggunakan sistem peringkas teks otomatis).

Tabel 1. **Contoh pasangan teks parafrasa pada dataset ParaCotta [1]**

Teks I	Teks II
Burton tidak pernah bersabar kecuali ketika itu benar-benar diperlukan dan sering kali tidak.	Burton tidak pernah sabar kecuali bila itu benar-benar diperlukan dan sering tidak kemudian.
Hanya saja, dia benar-benar meminta kita ... tidak begitu banyak bertanya.	Hanya saja, ia benar-benar meminta kami ... tidak begitu banyak bertanya.

Pada Tabel 1 ditunjukkan bahwa pasangan teks parafrasa yang dihasilkan dataset ParaCotta[1] cenderung memiliki diksi yang kurang beranekaragam. Hal ini dapat disebabkan oleh korpus yang digunakan untuk melatih sistem pembangkit parafrasa otomatis cenderung memiliki korelasi negatif antara kesamaan semantik dan keragaman leksikal pada pasangan teks yang dihasilkan. Akibatnya, diperlukan suatu metode untuk mengekstrak parafrasa dari Abstractive Summarization. Lalu, hasil ekstraksi tersebut dijadikan dataset parafrasa bahasa Indonesia untuk membangun sistem pembangkit parafrasa otomatis.

Kontribusi utama pada penelitian ini adalah membangun dataset parafrasa bahasa Indonesia dengan memanfaatkan *task* Abstractive Summarization. Dalam pembangunan dataset tersebut, dibangun pula suatu metode ekstraksi dan filtrasi untuk sintesis dataset parafrasa.Lalu, dataset parafrasa yang dihasilkan dievaluasi melalui sistem pembangkit parafrasa otomatis dan *human evaluation*.

Penelitian ini bertujuan untuk membangun dataset parafrasa bahasa Indonesia yang akan digunakan untuk melatih sistem pembangkit parafrasa otomatis. Dataset yang dihasilkan akan dievaluasi baik secara otomatis maupun manual. Lalu, akan dilakukan analisis berdasarkan pengujian pada sistem yang dibangun terhadap data uji.

Pada sub selanjutnya dijelaskan studi yang berkaitan dengan penelitian ini. Setelah itu dijelaskan bagaimana teknik dalam pembangunan dataset serta alat evaluasi yang digunakan. Hasil dan analisis penelitian dipaparkan setelahnya diakhiri dengan kesimpulan.