Abstrak

Ketersediaan korpus paralel pada pasangan bahasa Sunda-Indonesia masih sangat sedikit. Korpus paralel tersebut penting dan bisa dimanfaatkan sebagai sumber data latih dalam sistem *machine translation* atau sistem *natural language processing*. Penelitian ini mencoba untuk mengumpulkan kalimat paralel yang didapatkan dari pasangan artikel Wikipedia berbahasa Sunda dan berbahasa Indonesia menggunakan fasilitas *interlanguage links*. Sebuah *bilingual lexicon* dan beberapa filter yang berdasarkan pada kemunculan kata, panjang kalimat dan *word overlap* antar kalimat digunakan untuk mendapatkan kalimat paralel. Metode *bootstrapping* kemudian digunakan untuk meningkatkan kualitas kalimat paralel dengan cara memperbarui *bilingual lexicon* memanfaatkan IBM Model 4 *expectation maximization* (EM) *learner* di dalam *tool* GIZA++. GIZA++ dijalankan pada kandidat kalimat paralel yang dihasilkan di setiap iterasi sistem sampai kondisi konvergensi tercapai. Hasil evaluasi manual menggunakan penilaian manusia menunjukkan bahwa 79,5% dari korpus paralel hasil bentukan sistem terbukti paralel.

Kata kunci: korpus paralel, Wikipedia, *bootstrapping*, *expectation maximization*