

ABSTRAK

Nama : Bernadia Puspasari

Program Studi : Ilmu Komputer

Judul : Penerjemahan Teks Indonesia – Jepang Otomatis Menggunakan Mesin Penerjemah Statistik Berdasarkan Frase

Mesin Penerjemah digunakan untuk menerjemahkan teks dari suatu bahasa ke bahasa lain secara otomatis. Mesin Penerjemah Statistik adalah Mesin Penerjemah yang menggunakan pendekatan statistik dalam proses menerjemahkan teks. Penelitian dilakukan dengan menggunakan Mesin Penerjemah Statistik berdasarkan frase yang memanfaatkan korpus dwibahasa paralel sebagai data pelatihannya. Korpus dwibahasa Indonesia – Jepang yang digunakan berupa koleksi dokumen Kitab Suci, artikel berita, dan percakapan sehari-hari dengan jumlah keseluruhan kalimat sebanyak 24.365 kalimat. Koleksi dokumen dalam bahasa Jepang tersedia dalam dua macam bentuk tulisan, yaitu Kanji atau Romaji. Penelitian dilakukan pada korpus dwibahasa tanpa faktor tambahan dan korpus yang menggunakan perangkat bahasa tambahan, yaitu lema. Dari hasil penelitian, didapati bahwa kinerja penerjemahan teks Indonesia – Jepang menggunakan Mesin Penerjemah Statistik berdasarkan frase pada penelitian ini, nilai akurasi tertinggi berdasarkan BLEU *score* mencapai 0,2027. Nilai akurasi tertinggi tersebut didapatkan pada jenis dokumen artikel berita tanpa faktor tambahan dengan model bahasa 5-gram. Sedangkan penambahan perangkat bahasa lema pada korpus pelatihan menurunkan kinerja dari Mesin Penerjemah Statistik berdasarkan frase.

Kata kunci : Mesin Penerjemah, statistik, frase, lema

ABSTRACT

Name : Bernadia Puspasari

Study Program : Computer Science

Title : Automatic Indonesian – Japanese Text Translation Using
Phrase-based Statistical Machine Translation

Machine Translation translates text from one language to another automatically. Statistical Machine Translation uses statistical approach to translate text. This research uses phrase-based Statistical Machine Translation system. We use Indonesian – Japanese bilingual corpora as the training data which consist of holy writings, news article, and daily conversation with total of 24.365 sentences. Japanese document collections are written in Kanji and Romaji. This research uses unfactored training corpora and factored training corpora (lemma). The highest accuracy based on evaluation of the translation result is 0.2027 in BLEU score which is the score of news article document written in Romaji using 5-gram language model. Factored training corpora (lemma) decreases the performance of the machine translation system.

Key words : Machine Translation, statistical, phrase-based, lemma.