Penerjemahan teks Indonesia-Inggris otomatis menggunakan mesin penerjemahan statistik berdasarkan frase dengan koleksi dokumen menggunakan part-of-speech tagging dan lema

Metti Zakaria Wanagiri, author

Abstrak

Mesin Penerjemah (MP) adalah sebuah sub-bagian dari computational linguistics yang menggunakan komputer untuk menerjemahkan teks dari sebuah bahasa ke bahasa yang lain. Sementara Mesin Penerjemah Statistik (MPS) adalah sebuah pendekatan MP dimana hasil terjemahan dihasilkan atas dasar model statistik yang parameter-parameternya diambil dari hasil analisis korpus teks dwibahasa (yang paralel). Pada tugas akhir ini, penerjemahan teks Indonesia-Inggris dilakukan dengan menggunakan MPS berdasarkan frase dimana penerjemahan dilakukan dengan menggunakan prinsip penerjemahan berdasarkan frase. Korpus dwibahasa Indonesia-Inggris yang digunakan terdiri dari kategori berita, kitab suci, novel dan percakapan. Jumlah korpus pelatihan yang digunakan adalah 40779 kalimat, yaitu 704 berita, 4025 percakapan, 16050 novel dan 20000 kitab suci. Sementara korpus pengujian yang digunakan adalah 20300 kalimat, yaitu 300 berita, 2000 percakapan, 8000 novel dan 10000 kitab suci. Percobaan penerjemahan ini dilakukan, dievaluasi dan dianalisis dari dua aspek yaitu penggunaan perangkat bahasa tambahan (yang meliputi Part-of-Speech Tagging dan lema) dan n-gram yang digunakan dalam membentuk model bahasa. Hasil percobaan yang didapat adalah nilai akurasi tertinggi dicapai oleh penerjemahan korpus dwibahasa biasa (tidak menggunakan Part-of-Speech Tagging maupun lema) pada kategori novel dengan menggunakan model bahasa 5-gram, yaitu 0,2696.

<hr>Machine Translation (MT) is a sub-field of computational linguistics that uses a computer to translate text or speech from one natural language to another. Meanwhile Statistical Machine Translation (SMT) is a paradigm of MT where translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora (parallel). The Indonesian-English text translation is done using a phrase-based SMT in which the translation is carried out using phrase-based Translation. We use Indonesian and English bilingual corpora which consists of news, holy writings, fiction and daily conversation categories. We use training corpus of 40779 sentences which are 704 for news, 4025 for conversation, 16050 for fiction and 20000 for holy writings. Meanwhile the testing corpus consists of 20300 sentences which are 300 for news, 2000 for conversation, 8000 for fiction and 10000 for holy writings. Experiments have been done, evaluated and analyzed regarding two aspects, namely the use of factored-models (Part-of-Speech Tagging and lemma) and number of n-gram for generating the language model. In this thesis, we found that the translations of default bilingual corpora (without Part-of-Speech Tagging and lemma) for fiction category using 5-gram language model yield the highest accuracy of 0.2696.