Pengembangan Metode Ekstraksi Sumber Daya NLP dari Kamus Dwibahasa Indonesia dan Bahasa Daerah = Extracting NLP Resources from Bilingual Dictionaries for Regional Languages in Indonesia

Julian Fernando, author

Abstrak

Perkembangan NLP bahasa daerah di Indonesia masih tergolong lambat. Banyak faktor yang melatarbelakangi hal tersebut, seperti dokumentasi bahasa yang buruk, penutur bahasa yang sedikit, dan kurangnya sumber daya untuk mempelajari NLP bahasa daerah. Penelitian ini bertujuan untuk mengembangkan metode ekstraksi kamus dwibahasa Indonesia dan bahasa daerah yang umum untuk menghasilkan sumber daya NLP. Sistem yang dihasilkan mampu mengolah banyak kamus dwibahasa sekaligus menjadi sumber daya NLP. Kamus terlebih dahulu dikonversi ke dalam bentuk machine readable dan diolah ke bentuk korpus entri sebelum dilakukan ekstraksi. Korpus entri adalah korpus yang mengandung informasi lengkap setiap entri di dalam kamus beserta jenis font, ukuran, dan posisi setiap kata pada entri di dalam kamus dwibahasa. Proses ekstraksi dilakukan dengan memperhatikan pola entri sehingga perlu dilakukan tahap standardisasi entri terlebih dahulu sebelum sumber daya dibentuk. Selain pembentukan sumber daya, dilakukan pula perbaikan ejaan khusus untuk sumber daya korpus paralel. Dalam mengevaluasi hasil ekstraksi, diambil beberapa kamus dwibahasa sebagai sampel. Evaluasi dilakukan dengan memperhatikan ketepatan peletakan setiap komponen entri di dalam hasil ekstraksi. Tim peneliti menemukan bahwa sistem yang dibangun telah berhasil mengekstrak sumber daya NLP berupa leksikon bilingual, kamus morfologi, dan korpus paralel dengan optimal pada 32 kamus dwibahasa Indonesia dan bahasa daerah. Masih terdapat beberapa kekurangan pada sistem yang berhasil dibangun karena proses ekstraksi sangat bergantung dengan ketepatan pendeteksian font sehingga kualitas kamus masih memberikan pengaruh yang besar pada kualitas hasil ekstraksi.

.....The development of regional language NLP in Indonesia is still relatively slow. There are several factors behind this, such as poor language documentation, a small number of speakers of the language, and lack of the resources needed to study regional language NLP. This research aims to develop a general extraction method for Indonesian and regional bilingual dictionaries to produce NLP resources. The resulting system is able to process multiple bilingual dictionaries at once into NLP resources. Dictionaries are converted to machine readable form and processed to the form of a corpus of entries in advance before extraction is carried out. A corpus of entries means corpus that contains full information of each entry in the dictionary as well as font style, font size, and the position of each word of the entry in the bilingual dictionary. The extraction process is carried out by observing the entry's pattern resulting in the entry standardization phase having to be done prior before resources are produced. Besides resource production, spell checking is also carried out specifically for parallel corpus resources. In order to evaluate the extraction results, several bilingual dictionaries are taken to be samples. Evaluation process is carried out by observing the accuracy of each entry component’s placement in the extraction results. Research team found that the resulting system has succeeded in extracting NLP resources optimally in the form of bilingual lexicon, morphology, and parallel corpus on 32 Indonesian and regional bilingual dictionaries. There are still some deficiencies in the developed system since the extraction process is highly dependent on the accuracy of font detection such

that the qualities of dictionaries still have a big impact on the quality of extraction results.