Penambangan Kamus Dwibahasa: Studi Percontohan Pada Bahasa Indonesia dan Bahasa-Bahasa Daerah = Bilingual Dictionary Mining: A Pilot Study on Indonesian and Local Languages in Indonesia

Intan Fadilla Andyani, author

Abstrak

Pengembangan NLP di Indonesia terbilang lambat, terutama penelitian terkait bahasa daerah Indonesia. Alasannya adalah sumber data bahasa daerah tidak terdokumentasikan dengan baik sehingga sumber daya NLP yang ditemukan juga sedikit. Penelitian ini membahas metode ekstraksi kamus-kamus bahasa daerah di Indonesia untuk menghasilkan suatu sumber daya NLP yang dapat dibaca oleh mesin. Tahap penelitian dimulai dari pengumpulan data kamus, perancangan dan eksperimen metode ekstraksi, serta evaluasi hasil ekstraksi. Hasil penelitian berupa korpus paralel, leksikon bilingual, dan pasangan kata dasar-kata berimbuhan dalam format CSV dari beberapa kamus dwibahasa di Indonesia. Beberapa bahasa di antaranya adalah bahasa Minangkabau, Sunda, Mooi, Jambi, Bugis, Bali, dan Aceh. Perancangan metode ekstraksi berfokus pada kamus Minangkabau yang kemudian dilakukan eksperimen pada kamus-kamus bahasa daerah lainnya. Evaluasi dilakukan terhadap hasil ekstraksi kamus Minangkabau dengan melakukan anotasi data. Perhitungan akurasi dilakukan terhadap penempatan kelompok kata dari hasil anotasi. Hasil perhitungan menunjukkan 99% hasil ekstraksi sudah tepat untuk penentuan kelompok kata pada leksikon bilingual dan 88% untuk korpus paralel. Tim peneliti menemukan bahwa struktur dalam kamus bahasa daerah Indonesia sangat beragam, sehingga menuntut perlakuan yang berbeda pada setiap kamus, seperti perihal penomoran halaman. Selain itu, tim peneliti menemukan banyak kamus bahasa daerah Indonesia dengan kualitas yang kurang baik. Kualitas yang kurang baik ditunjukan dengan banyaknya kesalahan baca akibat noise yang terdapat pada tampilan berkas kamus.

.....The development of NLP in Indonesia is relatively slow, especially for Indonesian local languages. Indonesian local language data sources are not well-documented so that there are only few NLP resources found. This study discusses the extraction method of Indonesian local language dictionaries to produce a machine-readable NLP resource. Starting from collecting dictionary data, designing and experimentation of the extraction method, and evaluating the extraction results. The extraction results are parallel corpus, bilingual lexicon, and words' morphological form in CSV format from several Indonesian Local Language bilingual dictionaries that are Baso Minangkabau, Sundanese, Moi, Jambinese, Buginese, Balinese, and Acehnese. The designed method is also applied to some other local language dictionaries. Data annotation has been done to evaluate the extraction results so that we can calculate its accuracy of word classification for parallel corpus and bilingual lexicon. Extraction method design focuses on the Minangkabau dictionary which is then applied to other dictionaries. Data annotation has been done to evaluate the extraction results.The evaluation results show that 99% of the extraction results are correct for word classifying in the bilingual lexicon and 88% correct for parallel corpus. We found that the structure of dictionaries varies, so it requires different approaches for each dictionary, for example regarding page numbering. We also found many dictionaries with poor quality. The poor quality is indicated by the number of reading errors due to noise contained in the original dictionary file.