Word sense disambiguation (WSD) untuk bahasa indonesia menggunakan cross-lingual WSD dengan korpus paralel dan wordnet = Word sense disambiguation WSD for Indonesian language using cross lingual WSD with parallel corpora and wordnet

Heninggar Septiantri, author

Abstrak

Ambiguitas adalah masalah yang seringkali ditemui dalam pemrosesan bahasa alami oleh komputer. Word Sense Disambiguation (WSD) adalah upaya untuk menentukan makna yang tepat dari sebuah kata yang ambigu. Berbagai penelitian tentang WSD telah banyak dikerjakan, namun penelitian WSD untuk bahasa Indonesia belum banyak dilakukan. Ketersediaan korpus paralel berbahasa Inggris-Indonesia dan sumber pengetahuan bahasa berupa WordNet bahasa Inggris dan bahasa Indonesia dapat dimanfaatkan untuk menyediakan data pelatihan untuk WSD dengan metode Cross-Lingual WSD (CLWSD). Data pelatihan ini kemudian dijadikan input untuk klasifikasi dengan algoritma Naive Bayes, sehingga model klasifikasinya dapat digunakan untuk melakukan monolingual WSD untuk bahasa Indonesia.

Evaluasi klasifikasi menunjukkan rata-rata akurasi hasil klasifikasi lebih tinggi dari baseline. Penelitian ini juga menggunakan stemming dan stopwords removal untuk mengetahui bagaimana efeknya terhadap klasifikasi. Penggunaan stemming menaikkan rata-rata akurasi, sedangkan penerapan stopwords removal menurunkan rata-rata akurasi. Namun pada kata yang memiliki dua makna dalam konteks yang cukup jelas berbeda, stemming dan stopwords removal dapat menaikkan rata-rata akurasi.
<hr>
Ambiguity is a problem we frequently face in natural languange processing. Word Sense Disambiguation (WSD) is an attempt to decide the correct sense of an ambiguous word. Various research in WSD have been conducted, but research in WSD for Indonesian Language is still rare to find. The availability of parallel corpora in English and Indonesian language and WordNet for both language can be used to provide training data for WSD with Cross-Lingual WSD (CLWSD) method. This training data can be used as input to the classification process using Naive Bayes classifier.

The model resulted by the classification process is then used to do monolingual WSD for Indonesian language. The whole process in this research results in higher accuracy compared to baseline. This research also includes the use of stemming and stopwords removal. The effect of stemming is increasing the average accuracy, whereas stopwords removal is decreasing average accuracy. Nevertheless, for ambiguous words that have distinct context of usage, the use of stemming and stopwords removal can increase average accuracy.