Universitas Indonesia Library >> UI - Skripsi Open

Automatic english to Indonesia lexical mapping using latent semantic analysis

Eliza Margaretha, author

Deskripsi Lengkap: https://lib.ui.ac.id/detail?id=123805&lokasi=lokal

---

Abstrak

WordNet (Fellbaum, 1998) adalah suatu lexical resource yang kaya akan informas linguistik yang sangat bermanfaat bagi berbagai macam aplikasi, khususnya aplikasiaplikasi yang berhubungan dengan linguistik, pemrosesan bahasa alami, dan kecerdasan buatan. Dewasa ini, WordNet telah dibangun untuk lebih dari 40 bahasa, tetapi WordNet untuk bahasa Indonesia belum tersedia. Oleh karena pengembangan WordNet secara manual membutuhkan sumber daya yang tidak sedikit, penelitian yang dipaparkan dalam laporan tugas akhir ini bermaksud untuk membangun WordNet secara otomatis.

Penelitian ini mencoba untuk membuat synset (synonym set) untuk bahasa Indonesia dengan melakukan pemetaan konsep dwibahasa secara otomatis antara konsep bahasa Inggris yang diambil dari Princeton WordNet dan konsep bahasa Indonesia yang diambil dari Kamus Besar Bahasa Indonesia (KBBI). Tugas lain, yaitu pemetaan kata dwibahasa, diperkenalkan untuk memetakan kata-kata bahasa Inggris ke kata-kata bahasa Indonesia secara otomatis. Kedua pemetaan tersebut dilakukan dengan mengaplikasikan metode Latent Semantic Analysis (Landauer, Foltz, & Laham, 1998) pada korpora paralel berupa teks.

Awalnya, pemetaan kata dwibahasa dimaksudkan untuk melakukan verifikasi proses di balik pemetaan konsep dwibahasa. Namun, hasil pemetaan kata tidak memuaskan karena performa model kemiripan vektor lebih baik dari pada model LSA. Di sisi lain, hasil dari pemetaan konsep dwibahasa, menunjukkan kemampuan LSA untuk menangkap informasi semantik yang terkandung secara implisit dalam suatu korpus parallel. Walaupun LSA belum berhasil mencapai tingkat yang setara dengan pemetaan yang dilakukan manusia, secara umum LSA lebih baik dari pada random baseline.
<hr>
WordNet (Fellbaum, 1998) is a lexical resource containing rich linguistic knowledge, which is very useful for a wide variety of applications, especially for applications related to linguistics, natural language processing, and artificial intelligence. Recently, WordNets have been built for more than 40 languages, but not yet in Indonesian. Since building a WordNet manually is complex and expensive, the work presented in this thesis considers building an Indonesian WordNet automatically.

This work attempts to construct Indonesian synsets (synonym set) by conducting automatic bilingual concept mapping between English concepts derived from Princeton WordNet and Indonesian concepts derived from Kamus Besar Bahasa Indonesia (KBBI). Another task, namely bilingual term mapping, is introduced to map English terms to their Indonesian analogues automatically. Both mappings are conducted by applying LatentSemantic Analysis (Landauer, Foltz, & Laham, 1998) on parallel corpora of text.

Bilingual term mapping was intended to verify the underlying process of bilingual concept mapping.

However, the results are unsatisfactory suggesting that vector model similarity performs better than the LSA model. The results of bilingual concept mapping, on the other hand, show some capability of LSA to capture some semantic information implicit within a parallel corpus. Although LSA is not yet able to attain levels comparable to human judgements, it is generally better than random baseline.