ABSTRAK

Klasifikasi argumen semantik adalah proses menganalisa kalimat untuk menyelidiki pola WHO did WHAT to WHOM, WHEN, WHERE, WHY dan HOW dari struktur data teks. Penelitian terkait klasifikasi argumen semantik memerlukan data yang telah diberi label semantik dalam jumlah yang besar, yang disebut korpus. Pada penelitian awal, telah dibangun dua jenis korpus yaitu FrameNet dan Propbank, keduanya merupakan domain atau aliran berita. Karena membangun korpus membutuhkan biaya yang besar dan waktu yang lama, maka beberapa tahun terakhir telah banyak penelitian yang memanfaatkan korpus FrameNet dan Propbank sebagai data pelatihan untuk melakukan penelitian klasifikasi argumen semantik pada domain yang baru tanpa perlu membangun korpus untuk domain baru tersebut.

Thesis ini akan melakukan penelitian terkait klasifikasi argumen semantik pada domain baru yaitu translasi AlQuran dalam bahasa Inggris dengan memanfaatkan korpus Propbank sebagai data latih. AlQuran dalam bahasa Inggris adalah alih bahasa dari AlQuran asli berbahasa Arab. Oleh karena itu dalam tata bahasa dan struktur kalimat, bahasa dalam translasi Quran berbahasa Inggris masih dipengaruhi oleh bahasa aslinya, yaitu bahasa Arab. Di dalamnya bahasa aslinya yaitu bahasa Arab, AlQuran memiliki perbedaan yang signifikan dari domain berita, lebih dekat dengan bahasa puitis, memiliki ekspresi linguistik yang lebih kreatif, dan memiliki banyak variasi kosa kata dan struktur kalimat.

Peneitian terdahulu telah membuktikan bahwa terjadi penurunan performansi secara signifikan ketika melakukan klasifikasi argumen semantik pada domain yang berbeda antara data latih dan data uji. Masalah utamanya adalah karena ditemukan argumen baru yang terdapat pada data uji namun tidak ditemukan dalam data latih. Untuk mengenali argumen baru ini dalam data latih, salah satu solusinya adalah dengan memperluas fitur argumen dalam data latih untuk mengakomodasi fitur baru dari argumen baru tersebut. Thesis ini mengusulkan penambahan empat fitur baru pada sistem baseline untuk meningkatkan kinerja sistem.

Dengan menggunakan SVM Linear, percobaan telah membuktikan bahwa performansi klasifikasi argumen semantik pada domain Quran menggunakan data Propbank sebagai data training dapat ditingkatkan dengan penambahan fitur yang diusulkan pada sistem baseline dengan dengan beberapa pilihan kombinasi. Ketika diuji pada auto labeled data, penambahan fitur PTO+SP dapat meningkatkan akurasi sebesar 1.25% dan F-Measure sebesar 1.30%. Ketika diuji pada hand-labeled data, penambahan fitur PO+PTO dapat meningkatkan akurasi sebesar 0.47% dan F-Measure sebesar 0.40%.

Keywords: semantic argument classification, semantic role labeling, shallow semantic parsing, classification algorithm, Support Vector Machine classifier.