1. Pendahuluan

Latar Belakang

Natural Language Processing (NLP) adalah salah satu cabang ilmu sains, teknik informasi, dan kecerdasan buatan yang mempelajari hubungan interaksi antara komputer dengan bahasa manusia [1]. Ketika kita bicara bahasa manusia berarti kita juga bicara tentang kata yang jumlahnya sangat banyak dan bahasa yang bermacam - macam. Manusia mungkin dapat mengerti kata tersebut tergantung dari bahasa apa yang digunakan tetapi berbeda dengan komputer, untuk mengerti suatu kata berarti komputer harus mengenali ciri – ciri dari kata tadi. Oleh karena itu, terdapat istilah *Word Embedding* dari NLP, yaitu cabang ilmu yang mempelajari bagaimana komputer merepresentasikan suatu kata ke dalam bentuk vektor bilangan real sehingga vektor disini akan mewakili makna kata tersebut.

Pada penelitian sebelumnya [4], *Word Embedding* ini mendapatkan akurasi yang tinggi dalam kasus data set yang berjumlah enam milyar kata dengan 640 dimensi vektor dan *test set* yang diujikan berupa *semantic* (contoh: *brother;sister*, France;Paris) dan *syntatic* (contoh: *great;greater, possibly;impossibly*) dalam bahasa Inggris. Jika metod yang sebelumnya memperoleh total *semantic* dan *syntactic* dengan akurasi 35% dan 47%, *Word Embedding* ini mendapat akurasi 61% dan 51%. Juga untuk kasus dimensi vektor yang berbeda, metod sebelumnya menghabiskan 14 hari sedangkan *Word Embedding* ini menghabiskan hanya 2 – 2.5 hari saja. Hal ini disebabkan karena kompleksitas sistem yang dimiliki dari metod yang lama dan *Word Embedding* yang baru berbeda [4]. Jika metod yang lama masih menggunakan *projection* dan *hidden layer*, tetapi *Word Embedding* ini hanya menggunakan *projection layer* saja sehingga hasilnya mempengaruhi kecepatan kompleksitas dalam mengelola data yang besar. Oleh karena itu, penulis memutuskan akan menggunakan *Word Embedding* dari penelitian sebelumnya yang sekarang lebih dikenal dengan nama word2vec sebagai metod dalam tugas akhir ini.

Topik dan Batasannya

Untuk memahami sebuah bahasa diharuskan mengenal terlebih dahulu kosakata bahasa yang akan digunakan. Jadi kosakata memiliki peran penting dalam menjelaskan makna suatu kata. Sehingga kita bisa tahu kedekatan antara kata satu dengan kata yang lain melalui makna katanya. Di dalam NLP istilah tadi disebut kesamaan semantik [1]. Contoh kata “tonton” dan “baca” dibandingkan “tonton” dan “bioskop”, kata tonton dan baca dikatakan memiliki nilai kesamaan semantik karena dua kata tadi berarti melihat sedangkan tonton dan bioskop hanya saling berkaitan saja. Dalam penelitian yang dijelaskan di sub-bab sebelumnya, kata – kata yang diinputkan tadi masih dalam satu bahasa saja sehingga kita tidak akan tahu apa yang akan dihasilkan jika menggunakan bahasa yang berbeda. Contoh dalam bahasa berbeda (*Cross-Lingual*) : Inggris-Spanyol (Jupiter;Mercurio;3.25) dan Spanyol-Itali (estrealla;pianeta;2.83). Oleh karena itu, penulis tertarik untuk menggunakan data set dan *test set* dari bahasa yang berbeda serta word2vec sebagai metodnya sehingga diharapkan bisa mengetahui bagaimana nilai yang dihasilkan. Nilai yang dimaksud mengindikasikan seberapa dekat dua kata tadi, seperti di contoh sebelumnya.

Pada penelitian tugas akhir ini berfokus untuk mengimplementasikan sistem yang mampu mengukur kesamaan semantik kata antar bahasa menggunakan data set berbahasa Inggris dan Spanyol dan mengetahui korelasi dari pengukuran sistem dengan *gold standard*.

Batasan dari penelitian tugas akhir ini di antaranya :

1. Data set menggunakan data berbahasa Inggris dan Spanyol dari Europarl korpus.
2. Pengukuran kesamaan semantik menggunakan skala 1 sampai 5.
3. Korelasi menggunakan *Pearson Correlation.*

Tujuan

Tujuan dari penelitian tugas akhir ini adalah mengimplementasikan word2vec menggunakan korpus untuk menghitung kesamaan semantik kata antar bahasa dan memberikan analisis dari hasil korelasi sistem dengan *gold standard* yang ada.