1. Pendahuluan

Latar Belakang

Bahasa merupakan salah satu cara manusia untuk dapat mengungkapkan sebuah gagasan, ide atau menyampaikan perasaan kepada orang lain. Dengan adanya bahasa, dua individu atau lebih dapat mengespresikan berbagai ide, perasaan, arti dan pengalaman [1]. Indonesia memiliki sangat banyak bahasa yang beragam, dari daerah timur hingga barat. Salah satu contohnya adalah bahasa Minang, bahasa ini berasal dari daerah Sumatera Barat.

Bahasa Minang sudah mulai kurang digunakan, menurut Lembaga Kerapatan Adat Alam Minangkabau (LKAAM) banyak warga yang berdarah Minangkabau tidak mengajari anaknya sendiri untuk berbahasa Minang, tetapi mengajarkan bahasa Indonesia dan Inggris [2]. Hal ini dapat membuat bahasa Minang menjadi bahasa daerah yang terancam punah.

Dalam penelitian ini mesin penerjemah berperan sebagai solusi dalam masalah ini. Penelitian yang telah dilakukan dalam kasus menerjemahkan bahasa Minang dan bahasa Indonesia menggunakan metode *rule-based*, di dalam penelitian tersebut diketahui bahwa ketepatan hasil terjemahan adalah 97% [3]. Hasil terjemahan pada *rule-based* memiliki akurasi yang baik dikarenakan bahasa terjemahannya berdasarkan dari seorang ahli bahasa, sehingga hasil mesin terjemahan ini bahasanya sangat alami. Selain metode *rule-based,* ada juga metode lain yang dapat melakukan penerjemahan yaitu metode statistik.

Mesin penerjemah statistik tidaklah membutuhkan seorang ahli bahasa karena hasil terjemahan dihasilkan berdasarkan probabilitas yang dihitung di dalam mesin terjemahan, sehingga mesin terjemahan statistik jauh lebih murah dari segi biaya dibadingkan dengan metode *rule-based* yang menggunakan seorang ahli bahasa untuk membentuk mesin penerjemahnya.

Topik dan Batasannya

Berdasarkan latar belakang diatas, terdapat permasalahan yang akan diangkat dalam penelitian ini yaitu melihat efisiensi mesin penerjemahan statistik dalam menerjemahkan bahasa Minang dan bahasa Indonesia. Pada penelitian ini akan digunakan *parallel corpus* yang berjumlah 1300 *corpus* dimana 700 *corpus* digunakan sebagai data *training* dan 600 *corpus* digunakan sebagai data *testing*. Data *parallel corpus* bersumber dari *Wikipedia* bahasa Minang. Untuk topik yang digunakan dalam *parallel corpus* ini berkaitan tentang sejarah, kuliner, kesehatan dan teknologi. Selain *parallel corpus,* digunakan juga *monolingual corpus* yang bersumber dari *website* berita bahasa Indonesia yaitu kompas.com dan cnnindonesia.com yang dikumpulkan sebanyak 3000 *corpus*. Pengumpulan data dilakukan secara manual seperti mengambil teks artikel dari Wikipedia dan menerjemahkan setiap kalimat dalam *corpus* bahasa Minang yang telah dikumpulkan.

Tujuan

Tujuan dari penelitian Tugas Akhir ini adalah bagaimana menghasilkan hasil terjemahan statistik bahasa Minang - Indonesia terbaik dengan hanya menggunakan 1300 *parallel corpus* dan 3000 *monolingual corpus* dan beberapa skenario *test case* untuk pengujian.