BAB 1
PENDAHULUAN

1.1. Latar Belakang Masalah

Dalam Topik NLP (*Natural Language Processing*), kesalahan pada kata merupakan hal yang penting untuk diperhatikan. Berbagai topik dalam NLP seperti *question answering*, *information extraction*, *information retrieval* dan *machine translation* bergantung pada benarnya data masukan. Jika terdapat kesalahan dalam kata yang dimasukkan, maka proses-proses NLP tidak akan dapat berjalan dengan baik. Oleh karena itu, telah banyak dibuat program pemeriksa ejaan yang mendeteksi kesalahan-kesalahan ejaan kata dan kemudian mengoreksinya.

Ada dua jenis kesalahan ejaan, yaitu kesalahan ejaan *non-word* yang terjadi saat kata yang salah tidak ada didalam kamus, dan kesalahan ejaan *real-word* yang terjadi jika kata yang salah merupakan kata valid yang ada didalam kamus tetapi dianggap salah secara konteks atau makna. Kesalahan *real-word* lebih rumit untuk diatasi karena tidak dapat dideteksi oleh program pemeriksa ejaan pada umumnya. Biasanya program pemeriksa ejaan hanya mendeteksi kata yang tidak ada didalam kamus (kesalahan *non-word*).

Pada skripsi Muhammad Aburizal Siregar dari Universitas Komputer Indonesia telah mengatasi masalah kesalahan ejaan *real-word* dalam bahasa Indonesia berdasarkan metode n-gram lokal deteksi dan koreksi kesalahan kata *real-word* memiliki akurasi sebesar 11%[1] setelah mengubah perhitungan probabilitas n-gram dengan menggunakan metode *additive smoothing*, nilai yang sangat kecil jika dibandingkan dengan penelitian Samanta dan Chaudhuri sebesar 80%[2]. Oleh karena itu, dalam penelitian ini akan diambil beberapa metode *smoothing* untuk kasus deteksi dan koreksi kesalahan *real-word* dalam bahasa Indonesia yang diambil dari korpus PAN-BPPT dan UI *Tagged Corpus* dan akan ditambah dengan korpus bahasa Inggris dari *Wikipedia* tahun 2016 sebesar 20 juta kata yang di *translate* ke bahasa Indonesia. Metode yang diambil adalah metode bigram dan trigram lokal yang diajukan Samanta dan Chaudhuri. Metode ini dapat

mendeteksi lebih dari satu kesalahan dalam satu kalimat, dengan itu penelitian ini akan membandingkan metode *smoothing Good-Turing estimate, Jelinek-Mercer, Katz smoothing, Witten-Bell* dan *Absolute discounting* yang sudah diteliti oleh Stanley F. Chen dan Joshua Goodman[3], penelitiannya akan digunakan untuk melihat bagaimana hasil perbandingan dari perhitungan probabilitas n-gram untuk kasus koreksi kesalahan ejaan *real-word* bahasa Indonesia.

1.2. Identifikasi Masalah

Berdasarkan dari latar belakang yang telah diuraikan di atas, maka dapat diidentifikasi permasalahan yang dihadapi yaitu rendahnya akurasi penelitian yang dilakukan oleh Muhammad Aburizal Siregar untuk metode deteksi dan koreksi kesalahan ejaan *real-word* dalam bahasa Indonesia yang menggunakan perhitungan probabilitas n-gram dengan metode *additive smoothing*.

1.3. Maksud

Maksud dari penelitian ini adalah untuk membangun aplikasi deteksi dan koreksi kesalahan ejaan *real-word* yang menggunakan metode n-gram kata lokal pada teks bahasa Indonesia dan membandingkan perhitungan probabilitas n-gram menggunakan metode *smoothing Good-Turing estimate, Jelinek-Mercer, Katz smoothing, Witten-Bell* dan *Absolute discounting*.

1.4. Tujuan

Adapun tujuan dari penelitian ini adalah untuk mengukur akurasi yang dihasilkan oleh metode *smoothing Good-Turing estimate, Jelinek-Mercer, Katz smoothing, Witten-Bell* dan *Absolute discounting* untuk kasus koreksi kesalahan ejaan *real-word* bahasa Indonesia.

1.5. Batasan Masalah

Adapun batasan masalah yang diteliti ini lebih terarah dan dapat dipahami maka penelitian ini diberikan batasan-batasan masalah, yaitu:

Masukan

1. Bahasa masukan adalah bahasa Indonesia.
2. Format file masukan yaitu .txt.
3. Kamus yang digunakan diambil dari Kamus Besar Bahasa Indonesia.
4. Data latih diambil dari *UI Tagged Corpus*, *PAN-BPPT Localization* bahasa Indonesia dan korpus bahasa Inggris dari *Wikipedia* tahun 2016 sebesar 20 juta kata yang di *translate* ke bahasa Indonesia.
5. Data uji berasal dari artikel, cerita fiksi, majalah, berita surat kabar dan jurnal akademik yang bersumber dari *Wikipedia*.

Proses

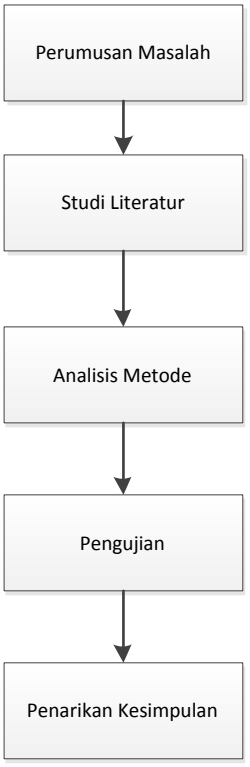
1. Metode N-Gram yang digunakan adalah unigram (n=1), bigram (n=2) dan trigram (n=3) pada level kata.
2. Deteksi dan koreksi kesalahan ejaan akan dibatasi hanya pada kesalahan ejaan *real-word*. Setiap data yang dimasukkan dipastikan sudah tidak ada kesalahan ejaan *non-word*.
3. Perhitungan probabilitas n-gram menggunakan metode *smoothing Good-Turing estimate, Jelinek-Mercer, Katz smoothing, Witten-Bell* dan *Absolute discounting*.

Keluaran

Keluaran yang dihasilkan adalah daftar kesalahan kata *real-word* dan daftar sugesti kata koreksi untuk tiap-tiap kesalahan dan hasil perbandingan metode *smoothing Good-Turing estimate, Jelinek-Mercer, Katz smoothing, Witten-Bell* dan *Absolute discounting*.

1.6. Metode Penelitian

Metode penelitian yang dipakai adalah metode penelitian deskriptif. Metode ini digunakan karena dalam penerapan metode deteksi dan koreksi kesalahan ejaan *real-word* dilakukan berdasarkan fakta-fakta yang didapat dalam pengumpulan data. Dalam metode penelitian ini digunakan teknik-teknik analisis klasifikasi masalah, survei, studi literatur terhadap masalah-masalah yang berhubungan dengan penelitian yang dijalankan, wawancara dengan narasumber dan teknik tes terhadap metode deteksi dan koreksi kesalahan ejaan yang telah ada. Dengan metode deskriptif, data yang telah dikumpulkan mula-mula disusun, dijelaskan, dianalisis, dan kemudian diimplementasikan dalam sebuah sistem/perangkat lunak. Berikut adalah skema metode penelitian deskriptif.



Gambar 1.1 Skema Metode Penelitian Deskriptif

Tahapan dalam metodologi penelitian diambil sebagai acuan dalam penyusunan tugas akhir ini dan berikut adalah penjelasannya.

1.6.1. Perumusan Masalah

Identifikasi permasalahan yang dihadapi yaitu kecilnya akurasi untuk metode deteksi dan koreksi kesalahan ejaan *real-word* dalam bahasa Indonesia dibandingkan dengan penelitian yang menggunakan bahasa Inggris.

1.6.2. Studi Literatur

Pada penelitian ini metode yang digunakan yaitu dari studi literatur dan wawancara. Studi literatur yang diperoleh dari sumber-sumber tertulis, baik yang tercetak maupun elektronik, seperti buku, *e-book*, jurnal, *paper*, dan sumber-sumber yang berhubungan dengan n-gram dan koreksi kesalahan kata *real-word* dan metode *smoothing*.

1.6.3. Analisis Metode

Analisis yang dilakukan yaitu mulai penerimaan input teks dokumen hingga melakukan proses deteksi dan koreksi kesalahan *real-word* menggunakan metode *smoothing good-turing, jelinek mercer, katz, witten-bell* dan *absolute discounting*.

1.6.4. Pengujian Sistem

Metode yang digunakan untuk pengujian sistem yaitu menggunakan metode *black box* yang digunakan untuk mengetahui semua hasil dari akurasi yang dihasilkan oleh metode yang telah diterapkan.

1.6.5. Penarikan Kesimpulan

Pada penarikan kesimpulan ini, akan menggambarkan atau menyimpulkan dari semua hasil pengujian akurasi yang telah dibangun pada aplikasi deteksi dan koreksi kesalahan *real-word* menggunakan metode *smoothing good-turing, jelinek mercer, katz, witten-bell* dan *absolute discounting*.

1.7. Sistematika Penulisan

Adapun sistematika penulisan dalam penyusunan penulisan ini adalah sebagai berikut:

BAB 1 PENDAHULUAN

Menguraikan tentang latar belakang permasalahan, mengidentifikasi

permasalahan dan menentukan tujuan penelitian, yang kemudian diikuti dengan pembatasan masalah, metode penelitian, serta sistematika penulisan.

BAB 2 TINJAUAN PUSTAKA

Bab ini memuat berbagai dasar teori yang mendukung dan mendasari penulisan penelitian ini, di antaranya mengenai konsep kesalahan ejaan, penelitian-penelitian kesalahan ejaan real-word terdahulu, dan metode n-gram kata lokal, metode *smoothung*, beserta pemodelan, bahasa pemrograman, dan perangkat lunak yang digunakan.

BAB 3 ANALISIS DAN PERANCANGAN SISTEM

Pada bab ini berisi tentang tahapan analisis masalah dan solusi, analisis data masukan, analisis proses, analisis data keluaran, analisis basis data, dan analisis kebutuhan *fungsional* dan *nonfungsional*, serta perancangan sistem yang mencakup perancangan antar muka, struktur menu dan pesan.

BAB 4 IMPLEMENTASI PENGUJIAN SISTEM

Bab ini membahas tentang penerapan metode yang telah dianalisis. Dalam bab ini juga dilakukan pengujian untuk mengukur tingkat *precision*, *recall* metode dan akurasi.

BAB 5 KESIMPULAN DAN SARAN

Pada bab ini membahas tentang kesimpulan dari penelitian ini dan saran yang dapat dijadikan masukan untuk pengembangan penelitian NLP selanjutnya.