BAB 1
PENDAHULUAN

1.1. Latar Belakang Masalah

Dalam bidang *Natural Language Processing* (NLP), kesalahan kata merupakan hal yang penting diperhatikan. Berbagai topik dalam NLP seperti *information retrieval*, *POS-tagging*, 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 lalu mengoreksinya.

Terdapat dua jenis kesalahan ejaan, yaitu kesalahan ejaan *non-word* yang terjadi saat kata yang salah tidak terdapat dalam kamus, dan kesalahan ejaan *real-word* yang terjadi jika kata yang salah merupakan kata valid yang terdapat dalam kamus tetapi dianggap salah secara konteks. 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 terdapat dalam kamus (kesalahan *non-word*).

Beberapa penelitian telah dilakukan untuk mengatasi masalah kesalahan ejaan *real-word* dalam bahasa Inggris, di antaranya Mays *et al.* [8][1], Hirst dan Budanitsky [3], Peddler [4], dan Samanta dan Chaudhuri [2]. Dari metode-metode yang telah disebutkan, metode yang memiliki akurasi tertinggi adalah metode bigram dan trigram kata lokal yang diajukan Samanta dan Chaudhuri, dengan *precision* 71%-79% dan *recall* 81%-88% untuk deteksi kesalahan. Sementara itu penelitian tentang kesalahan ejaan dalam bahasa Indonesia kebanyakan hanya untuk kesalahan *non-word* yang tidak memperhatikan konteks kalimat.

Oleh karena itu, dibutuhkan suatu metode untuk mendeteksi dan mengkoreksi kesalahan ejaan *real-word* dalam bahasa Indonesia. Dalam penelitian ini akan diambil salah satu metode deteksi dan koreksi kesalahan *real-word* dalam bahasa Inggris dan menyesuaikannya untuk bahasa Indonesia.

Metode yang diambil adalah metode bigram dan trigram lokal yang diajukan Samanta dan Chaudhuri [2]. Metode ini dipilih karena memiliki akurasi yang sebanding atau lebih tinggi dengan metode-metode lainnya. Metode ini juga dapat mendeteksi lebih dari satu kesalahan dalam satu kalimat.

1.2. Identifikasi Masalah

Berdasarkan dari latar belakang yang telah diuraikan di atas, maka dapat dapat diidentifikasi permasalahan yang dihadapi yaitu belum adanya pengukuran akurasi untuk metode deteksi dan koreksi kesalahan ejaan *real-word* dalam bahasa Indonesia.

1.3. Maksud dan Tujuan

Maksud dari penelitian ini adalah untuk membangun aplikasi deteksi dan koreksi kesalahan ejaan *real-word* dengan metode n-gram kata lokal pada teks bahasa Indonesia. Adapun tujuan dari penelitian ini adalah untuk mengukur akurasi deteksi dan koreksi kesalahan ejaan *real-word* dalam bahasa Indonesia menggunakan metode n-gram kata lokal.

1.4. Batasan Masalah

Agar masalah yang diteliti tidak menyimpang, lebih terarah serta dapat dipahami dengan mudah maka dalam penelitian ini diberikan batasan-batasan masalah, yaitu:

Masukan

1. Format masukan adalah .txt.
2. Bahasa masukan adalah bahasa Indonesia.
3. Kamus yang digunakan berasal dari Kamus Besar Bahasa Indonesia [21].
4. Data latih berasal dari *UI Tagged Corpus* [22] dan *PAN-BPPT Localization* [23].
5. Data uji berasal dari artikel dan berita yang bersumber dari Internet.

Proses

1. Metode N-Gram yang digunakan menggunakan bigram (n=2) dan trigram (n=3) pada level kata.

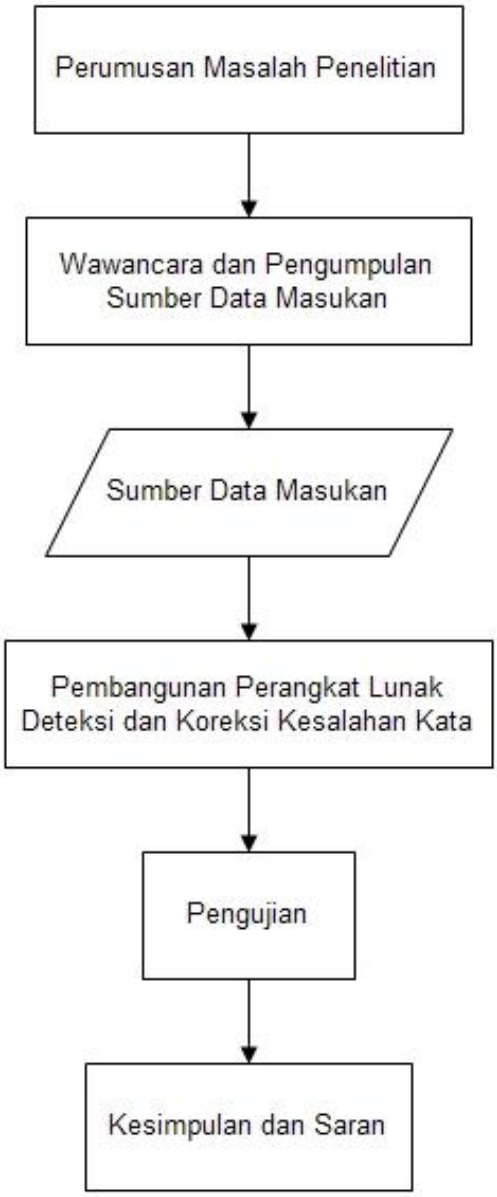
2. Deteksi dan koreksi kesalahan ejaan dibatasi hanya pada kesalahan ejaan *real-word*. Setiap data yang dimasukkan dipastikan sudah bebas dari kesalahan ejaan *non-word*.

Keluaran

Keluaran yang dihasilkan adalah daftar kesalahan kata *real-word* dan daftar sugesti kata koreksi untuk tiap-tiap kesalahan.

1.5. Metodologi 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, observasi, 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 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.

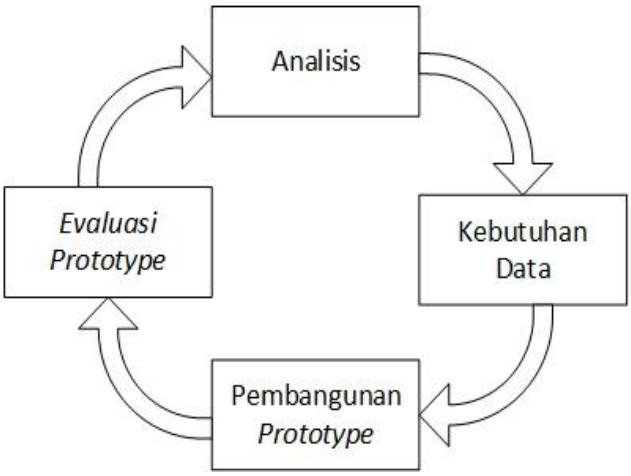
a. Metode Pengumpulan Data

Pada penelitian ini metode yang digunakan adalah 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*.

b. Metode Pembangunan Perangkat Lunak

Pembangunan perangkat lunak deteksi dan koreksi kesalahan ejaan *real-word* ini menggunakan model *prototype* karena dalam pembangunan perangkat lunak ini didasarkan pada ide pengembangan awal, diuji keberhasilannya, dan menyempurnakannya dengan kembali ke pembangunan. Tahapan-tahapan dari metode prototype dapat dilihat pada Gambar 1.2.



Gambar 1.2 Model *Prototype* [5]

1. Analisis

   Analisis yang dilakukan yaitu mulai penerimaan input teks dokumen hingga melakukan proses deteksi dan koreksi kesalahan *real-word*.

2. Pengumpulan Data

   Pada tahap ini peneliti akan mengumpulkan data berupa artikel dan berita berbahasa Indonesia untuk data masukan sistem.

3. Pembangunan *Prototype*

   Merupakan tahap pengimplementasian dari proses analisis dan kebutuhan sistem yang sudah didapatkan. Pada tahap ini pengembang mencoba mengimplementasikan metode n-gram kata lokal untuk deteksi dan koreksi kesalahan ejaan *real-word* ke dalam logika-logika program.

4. Evaluasi *Prototype*

Program akan diuji, dimana uji coba dilakukan untuk mengetahui kekurangan pada program. Jika masih ada kekurangan, maka *prototype* direvisi dengan tahapan-tahapan yang sebelumnya telah dilakukan. Proses ini akan terus berulang sampai *prototype* yang dihasilkan mendekati harapan yang diinginkan [5].

1.6. Sistematika Penulisan

Sistematika penulisan penelitian ini adalah sebagai berikut.

BAB 1 PENDAHULUAN

Bab ini akan membahas kerangka penelitian atau percobaan dalam tugas akhir, meliputi latar belakang masalah, perumusan masalah, tujuan, batasan masalah, metode penyelesaian masalah, dan sistematika penulisan dari penelitian koreksi kesalahan ejaan *real-word* berbahasa Indonesia menggunakan metode n-gram.

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, beserta pemodelan, bahasa pemrograman, dan perangkat lunak yang digunakan.

BAB 3 ANALISIS DAN PERANCANGAN SISTEM

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 DAN PENGUJIAN

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

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.