BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Kesalahan pengetikan kata adalah hal yang penting untuk diperhatikan, karena akan merubah arti atau makna dan bisa saja tidak memiliki arti. Dalam topik NLP (*Natural Language Processing*) seperti *essay scoring, information extraction, information retrieval* dan *machine translation* bergantung pada benarnya suatu kata sebagai data masukan agar proses-proses dapat berjalan dengan baik. Oleh karena itu, telah banyak dibuat program pemeriksa ejaan yang mendeteksi kesalahan-kesalahan ejaan kata dan kemudian mengoreksinya.

Penelitian terkait kesalahan kata telah dilakukan diantaranya penelitian Muhammad Aburizal Siregar dengan menggunakan metode n-gram lokal dalam teks bahasa Indonesia. Penelitian tersebut memiliki nilai rata-rata akurasi sebesar 11%. Kecilnya nilai akurasi disebabkan oleh kecilnya ukuran korpus n-gram yang digunakan [1], Dilakukan pengembangan oleh Fernando Sihole menggunakan metode n-gram lokal dengan membandingkan beberapa metode *smoothing* yaitu *good-turing estimate, jelinek-mercer, katz smoothing, witten-bell* dan *absolute discounting* untuk kasus deteksi dan koreksi kesalahan ejaan *real-word*. Hasil dari perbandingan metode *smoothing* dapat disimpulkan bahwa metode *smoothing* yang paling baik adalah metode *absolute discounting* pada akurasi deteksi sebesar 80% dan akurasi koreksi sebesar 7% dan *witten bell* pada akurasi deteksi sebesar 79% dan akurasi koreksi sebesar 7%. Hasil akurasi koreksi yang rendah dipengaruhi oleh kecilnya ukuran korpus dan kurangnya kelengkapan korpus yang digunakan [2]. Kemudian pengembangan dilanjutkan oleh Ken Wite Ariing Cahyu dengan menggunakan metode *kneser-ney smoothing* untuk kesalahan ejaan *real-word* kata bahasa Indonesia dengan menggunakan jumlah kemunculan kata pada *confusion set*. Didapatkan hasil akurasi deteksi sebesar 19,8% dengan parameter D paling optimal yaitu 0,0001. Hasil akurasi deteksi yang rendah dipengaruhi oleh korpus yang terdapat sebagian besar merupakan hasil terjemahan yang tidak sesuai dengan tata bahasa Indonesia sehingga mengakibatkan kualitas

korpus yang kurang baik untuk digunakan [3]. Pada penelitian yang sudah diteliti oleh Stanley F. Chen dan Joshua Goodman mengenai metode *smoothing* yang mana pada *kneser-ney smoothing* memiliki modifikasi yaitu *modified kneser-ney smoothing*, dimana dikatakan memiliki performa yang sangat baik dengan mengungguli algoritma lain yang dievaluasi. Pada metode ini tidak menggunakan diskon tunggal tetapi menggunakan 3 tingkatan diskon untuk n-gram yang lebih dari nol [4]. Metode ini telah diterapkan pada penelitian yang dilakukan oleh Ronja dengan melakukan perbandingan algoritma *smoothing* untuk menerjemahkan dari bahasa Indonesia ke bahasa daerah dengan menggunakan mesin penerjemah statistik dengan didapatkan hasil yang terbaik dimiliki oleh *modified kneser-ney smoothing* dengan 3-gram nilai akurasi sebesar 68,04% dan 5-gram memiliki nilai akurasi 67,8% [5].

Berdasarkan latar belakang yang telah dijelaskan dan beberapa penelitian yang sudah dilakukan sebelumnya, maka pada penelitian ini akan menggunakan metode *modified kneser-ney smoothing* dalam perhitungan probabilitas n-gram untuk deteksi dan koreksi kesalahan kata bahasa Indonesia dan juga menambahkan korpus untuk meningkatkan nilai akurasi pendeteksian dan pengoreksian, dimana korpus yang ditambahkan adalah korpus berbahasa Indoneisa dari *Leipzig Corpora Collection* [6].

1.2 Rumusan Masalah

Berdasarkan uraian yang terdapat pada latar belakang masalah, maka dapat dirumuskan masalah yang ada yaitu bagaimana performansi metode *modified kneser-ney smoothing* pada deteksi dan koreksi kesalahan kata bahasa Indonesia.

1.3 Maksud dan Tujuan

Berdasarkan permasalahan yang telah dirumuskan, maka maksud dari penelitian ini adalah untuk melakukan deteksi dan koreksi kesalahan kata bahasa Indonesia menggunakan metode *modified kneser-ney smoothing*. Adapun tujuan dari penelitian ini adalah untuk mengetahui ketepatan metode *modified kneser-ney smoothing* dalam melakukan deteksi dan koreksi kesalahan kata dalam bahasa Indonesia.

1.4 Batasan Masalah

Batasan masalah dalam penelitian ini adalah sebagai berikut:

Masukan

1. Teks masukan adalah text file berformat .txt
2. Bahasa masukan adalah bahasa Indonesia
3. Kamus yang digunakan berasal dari Kamus Besar Bahasa Indonesia
4. Data latih diambil dari PANL-BPPT *Localization* bahasa Indonesia, korpus berita Tempo dan korpus *Leipzig Corpora Collection* dengan total keseluruhan kurang lebih 22 juta kata.
5. Data uji berasal dari artikel dan berita yang bersumber dari internet, terdapat 30 data uji, dengan masing-masing data uji berisi satu atau lebih paragraf yang bertemakan politik, pertanian, pencurian, dan berita internasional.

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*
3. Perhitungan probabilitas n-gram

Keluaran

Keluaran yang dihasilkan adalah daftar kesalahan kata *real-word* dan daftar saran kata koreksi setiap kesalahan kata.

1.5 Metodologi Penelitian

Metode penelitian yang digunakan pada penelitian ini adalah metode penelitian deskriptif karena penerapan sistem deteksi dan koreksi kesalahan kata dilakukan berdasarkan fakta-fakta yang didapatkan dari proses pengumpulan data dan penelitian yang dilakukan secara berkelanjutan sehingga diperoleh pengetahuan yang menyeluruh. Dalam metode penelitian ini digunakan teknik-teknik analisis perumusan masalah, pengumpulan data untuk data masukan yang

berkaitan dengan penelitian yang dikerjakan dan melakukan pengujian terhadap metode deteksi dan koreksi kesalahan ejaan. Tahapan penelitian yang akan dilakukan dapat dilihat pada gambar 1.1.

Gambar 1.1 Tahapan Metode Penelitian

Berikut merupakan penjelasan dari tahapan penelitian yang akan dilakukan.

1.5.1 Perumusan Masalah

Perumusan masalah yaitu kecilnya nilai akurasi deteksi pada metode *kneser-ney smoothing* yang diakibatkan oleh kualitas korpus yang kurang baik dikarenakan korpus hasil terjemahan tidak sesuai dengan tata bahasa Indonesia dalam kesalahan ejaan *real-word*, untuk itu berapa besar tingkat akurasi metode

modified kneser-ney smoothing dalam perhitungan probabilitas n-gram pada deteksi dan koreksi kesalahan kata bahasa Indonesia.

1.5.2 Pengumpulan Data

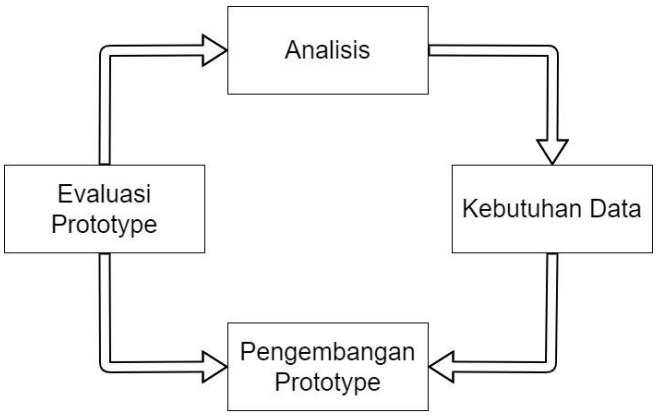
Metode pengumpulan data yang digunakan pada penelitian ini adalah studi literatur. Studi literatur diperoleh dari sumber-sumber tertulis, baik yang tercetak maupun elektronik, seperti buku, *e-book*, *paper*, jurnal dan sumber-sumber yang berhubungan dengan n-gram dan metode *smoothing* untuk deteksi dan koreksi kesalahan kata *real-word* dan juga dilakukan pengumpulan dataset yang terdiri dari KBBI yang bersumber dari DataHub.io dan juga penambahan korpus dari *Leipzig Corpora Collection*.

1.5.3 Analisis Metode

Analisis metode yang dilakukan yaitu dimulai dari penerimaan input teks dokumen, kemudian dilakukan *preprocessing* yang akan menghasilkan n-gram dan token data uji, sehingga dapat melakukan proses deteksi dan koreksi kesalahan *real-word* menggunakan metode *modified kneser-ney smoothing*.

1.5.4 Pengembangan Perangkat Lunak

Metode pengembangan perangkat lunak yang akan digunakan dalam penelitian ini adalah model *prototype*. Model *prototype* digunakan untuk mengembangkan sistem dari penelitian sebelumnya dengan menguji keberhasilan sistem, dan pengembangan disesuaikan dengan penelitian yang akan dikerjakan. Tahapan–tahapan pada metode *prototype model* dapat dilihat pada gambar 1.2.



Gambar 1.2 Model *Prototype* [7]

a. Analisis

Analisis dilakukan berdasarakan masalah yang akan diteliti, dimana pada tahap ini dianalisis bahwa pembuatan *confusion set* agar tidak berulang seperti penelitian sebulumnya maka hasil dari pembuatan *confusion set* disimpan ke dalam database daftar *confusion set*, kemudian data masukan yang digunakan merupakan data latih yang diambil dari korpus yang telah dijelaskan pada batasan masalah data uji di atas. dengan data uji yang digunakan untuk melakukan pengujian kemudian akan diproses oleh sistem untuk mengetahui deteksi dan koreksi kesalahan ejaan *real-word* dalam bahasa Indonesia menggunakan metode *modified kneser-ney smoothing*.

b. Kebutuhan data

Pada tahap ini akan dilakukan adalah mengumpulkan beberapa data, yaitu: data kamus, data daftar *confusion set* yang dibentuk dari data kamus, data uji dan beberapa korpus sebagai data latih seperti yang telah dijelaskan pada batasan masalah di atas.

c. Pengembangan *Prototype*

Pada tahap pengembangan dilakukan implementasi dari proses analisis dan kebutuhan sistem yang telah didapatkan untuk dilakukan pengembangan dari sistem yang telah dikerjakan sebelumnya, dengan mengimplementasikan pembuatan *confusion set* dengan sistem yang tersendiri dari deteksi dan koreksi kesalahan ejaan *real-word* dalam bahasa Indonesia yang mana pada penelitian ini menggunakan metode *smoothing modified kneser-ney*.

d. Evaluasi *Prototype*

Pada tahap ini program akan melalui tahap pengujian pada sistem, pengujian dilakukan untuk menemukan kesalahan yang mungkin terjadi dalam sistem untuk nantinya diperbaiki. Sehingga sistem diharapkan dapat bekerja dengan baik.

1.5.5 Pengujian Akurasi

Pengujian akurasi dilakukan untuk menguji nilai akurasi dari metode *modified kneser-ney smoothing*. Prose pengujian nilai akurasi dilakukan dengan melakukan 2 tahapan pengujian, tahap pengujian pertama menggunakan data uji sebagai data masukan, sedangkan pada tahap kedua menggunakan hasil dari pengujian tahap pertama sebagai data masukan dalam melakukan pendeteksian dan pengoreksian kesalahan kata dalam bahasa Indonesia.

1.5.6 Penarikan Kesimpulan

Penarikan kesimpulan akan menyimpulkan hasil dari pengujian akurasi yang telah dilakukan pada sistem deteksi dan koreksi kesalahan *real-word* menggunakan metode *modified kneser-ney smoothing*.

1.6 Sistematika Penulisan

Sistematika penulisan tugas akhir dijabarkan dalam bab-bab yang menggambarkan penelitian secara umum. Sistematika penulisan tugas akhir sebagai berikut :

BAB 1 PENDAHULUAN

Pada bab ini berisikan uraian tentang latar belakang masalah, rumusan masalah, maksud dan tujuan, batasan masalah, metodologi penelitian dan sistematika penulisan.

BAB 2 LANDASAN TEORI

Pada bab ini menguraikan terkait dengan dasar teori yang mendukung dan berhubungan dengan topik penelitian yang akan dikerjakan, diantaranya mengenai konsep kesalahan ejaan *real-word*, penelitian-penelitian terdahulu terkait kesalahan ejaan *real-word* dan metode n-gram yang digunakan.

BAB 3 ANALISIS DAN PERANCANGAN

Pada bab ini berisi penjelasan tentang analisis kebutuhan yang diperlukan baik dari kebutuhan fungsional maupun kebutuhan non fungsional serta menjelaskan perancangan dari sistem.

BAB 4 IMPLEMENTASI DAN PENGUJIAN

Pada bab ini membahas implementasi dari analisis dan perancangan yang telah dilakukan dan juga melakukan pengujian untuk mengukur tingkat akurasi deteksi dan koreksi.

BAB 5 KESIMPULAN DAN SARAN

Bab ini menjelaskan hasil dari kesimpulan penelitian yang telah dilakukan dan saran yang dapat dijadikan pengembangan penelitian kedepannya.