BAB 1
PENDAHULUAN

1.1 Latar Belakang

Kesalahan pengetikan kata adalah bagian dari *human error* dan merupakan hal yang perlu untuk diperbaiki karena akan mempengaruhi hal - hal yang membutuhkan benarnya suatu kata sebagai data masukan. Seperti berbagai topik dalam *Natural Language Processing* yaitu *question answering, information extraction, information retrieval* dan *machine translation.*

Telah dilakukan beberapa penelitian terkait kesalahan kata salah satunya adalah penelitian Muhammad Aburizal Siregar yang menggunakan metode n-gram. Penelitian tersebut memiliki nilai akurasi hanya sebesar 11%[1]. Oleh karena itu Fernando Sihole melakukan penelitian dengan mengambil beberapa metode *smoothing* lainnya untuk kasus deteksi dan koreksi kesalahan *real-word* untuk dibandingkan. Metode *smoothing* yang dibandingkan adalah metode *smoothing Good-Turing estimate, Jelinek-Mercer, Katz smoothing, Witten-Bell dan Absolute discounting.* Dari perbandingan metode *smoothing* yang telah dilakukan disimpulkan bahwa metode *smoothing* yang paling baik adalah metode *Absoulte Discounting* pada akurasi deteksi sebesar 80% dan akurasi koreksi sebesar 7 % dan *Witten Bell* pada akurasi deteksi sebesar 79% dan akurasi koreksi sebesar 7 % . [2] Hasil akurasi koreksi yang rendah tersebut dipengaruhi oleh kecilnya ukuran korpus dan kurangnya kelengkapan korpus yang digunakan. Hasil akurasi koreksi yang rendah juga dikarenakan dalam perhitungan probabilitas, metode *Absolute Discounting* dan *Witten Bell* menggunakan nilai probabilitas *Additive Smoothing* sedangkan pada metode *Kneser-Ney Smoothing* dikembangkan dengan menggunakan jumlah kemunculan kata pada *confusion set* data uji. Jauh sebelumnya Stanley dan Joshua Goodman [3] telah melakukan penelitian tentang metode *smoothing* yang menyimpulkan bahwa metode *Kneser-Ney Smoothing* dan variasinya menghasilkan performa yang konsisten dan lebih baik dibandingkan metode-metode *smoothing* yang lain. Oleh karena itu pada penelitian kali ini akan

menerapkan metode *Kneser-Ney Smoothing* dalam perhitungan probabilitas n-gram pada koreksi kesalahan kata bahasa Indonesia.

1.2 Identifikasi Masalah

Berdasarkan latar belakang yang telah diuraikan diatas maka dapat diidentifikasi permasalahan yang dihadapi yaitu kecilnya nilai akurasi koreksi pada penggunaan metode *Absolute Discounting* dan *Witten Bell* yang diakibatkan oleh kecilnya ukuran korpus, kurangnya kelengkapan korpus dan penggunaan nilai *additive smoothing* dalam perhitungan probabilitas serta belum adanya penggunaan metode *Kneser-Ney Smoothing* dimana metode ini dikembangkan dengan menggunakan jumlah kemunculan kata pada *confusion set* data uji dalam perhitungan probabilitas pada koreksi kesalahan kesalahan ejaan *real-word* dalam bahasa Indonesia yang menggunakan perhitungan probabilitas n-gram.

1.3 Maksud dan Tujuan

Maksud dari penelitian ini adalah untuk melakukan koreksi kesalahan kata bahasa Indonesia menggunakan metode *Kneser-Ney Smoothing*. Adapun tujuan dari penelitian ini adalah untuk mengetahui ketepatan metode *Kneser-Ney Smoothing* dalam melakukan koreksi kesalahan kata dalam bahasa Indonesia .

1.4 Batasan Masalah

Agar penelitian ini tidak menyimpang maka ditentukan batasan masalah sebagai berikut:

Masukan

1. Teks masukan berupa format .txt
2. Bahasa masukan adalah bahasa Indonesia
3. Kamus yang digunakan berasal dari Kamus Besar Bahasa Indonesia
4. Data latih diambil dari *UI Tagged Corpus, PAN-BPPT Localization* bahasa Indonesia, korpus berita tempo, korpus bahasa Inggris dari Wikipedia dan korpus berita *online* yang diterjemahkan
5. Data uji berasal dari artikel dan berita yang bersumber dari internet
6. Korpus yang digunakan merupakan korpus wikipedia berbahasa Inggris yang diterjemahkan menjadi bahasa Indonesia

Proses

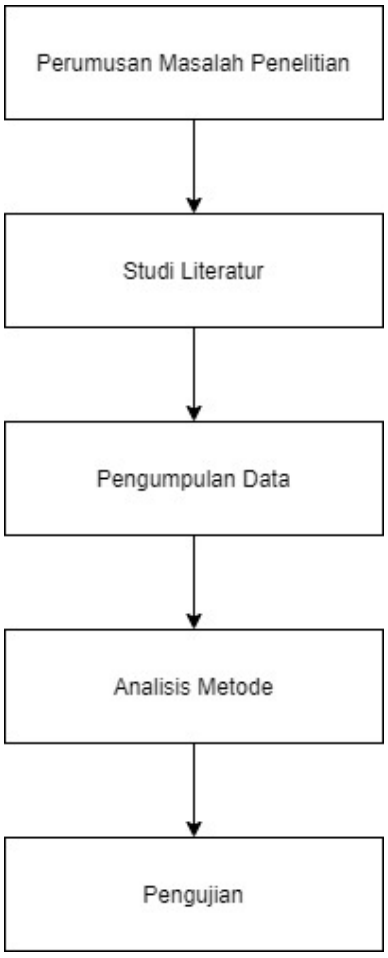
1. Metode N-Gram yang digunakan adalah unigram (n=1), bigram (n=2) dan trigram (n=3) pada level kata.
2. Deteksi dan koreksi kesalahan ejaan akan dibatasi hanya pada kesalahan ejaan *real-word.*
3. Perhitungan probabilitas n-gram

Keluaran

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

1.5 Metodologi Penelitian

Metode yang dipakai di penelitian kali ini adalah metode penelitian deskriptif. Metode penelitian deskriptif digunakan karena penerapan sistem deteksi dan koreksi kesalahan kata dilakukan berdasarkan fakta-fakta yang didapatkan dari proses pengumpulan data dan penelitian ini dilakukan secara berkesinambungan sehingga diperoleh pengetahuan yang menyeluruh mengenai masalah, fenomena, jika hubungan-hubungan fenomena tersebut dikaji dalam periode yang lama. Dalam metode penelitian deskriptif digunakan teknik – teknik analisis perumusan masalah, studi literatur terhadap permasalahan yang berhubungan dengan penelitian ini seara intensif dan melakukan pengujian terhadap metode deteksi dan metode koreksi kesalahan ejaan yang sudah ada.



Gambar 1. 1 Skema Metode Penelitian Deskriptif

Berikut merupakan penjelasan dari tahapan penelitian yang terdapat pada penelitian ini

1.5.1 Perumusan Masalah Penelitian

Mengidentifikasi masalah yaitu kecilnya nilai akurasi koreksi pada penggunaan metode *Absolute Discounting* dan *Witten Bell* yang diakibatkan oleh kecilnya ukuran korpus, kurangnya kelengkapan korpus dan penggunaan nilai *additive smoothing* dalam perhitungan probabilitas serta belum adanya penggunaan metode *Kneser-Ney Smoothing* dimana metode ini menggunakan nilai probabilitas jumlah kemunculan kata dalam perhitungan probabilitas pada koreksi kesalahan kesalahan ejaan *real-word* dalam bahasa Indonesia yang menggunakan perhitungan probabilitas n-gram.

1.5.2 Studi Literatur

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

1.5.3 Pengumpulan Data

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

1.5.4 Analisis Metode

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

1.5.5 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 Sistematika Penulisan

Sistematika penulisan bab-bab yang akan ditulisan dalam tugas akhir yang akan diambil adalah sebagai berikut :

BAB 1 PENDAHULUAN

Pada bab ini berisi uraian tentang latar belakang masalah, identifikasi masalah, maksud dan tujuan penelitian, batasan masalah, metodologi penelitian yang digunakan dan sistematika penulisan

BAB 2 LANDASAN TEORI

Pada bab ini menguraikan terkait dengan teori-teori yang berhubungan dengan topik penelitian.

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 yang dibuat.

BAB 4 IMPLEMENTASI DAN PENGUJIAN SISTEM

Bab ini menguraikan implementasi dari analisis dan perancangan yang sudah dilakukan dan juga melakukan pengujian dari sistem yang dibuat apakah sesuai dengan tujuan awal atau tidak.

BAB 5 KESIMPULAN DAN SARAN

Bab ini menjelasakan hasil dari penelitian yang sudah dilakukan dan juga disertai dengan saran yang diberikan untuk penelitian kedepannya