

BAB I PENDAHULUAN

1.1 Latar Belakang

Cross-Language Information Retrieval (CLIR) atau Temu Kembali Informasi Lintas Bahasa merupakan subbidang Temu Kembali Informasi atau *Information Retrieval* (IR) yang berkaitan dengan pengambilan informasi yang disimpan dalam bahasa yang berbeda dari bahasa permintaan yang diberikan pengguna. Tujuan CLIR adalah untuk memberikan manfaat kepada pengguna dalam menemukan dan mengakses informasi tanpa dibatasi oleh hambatan bahasa. Sebagai contoh pengguna mencari artikel berbahasa Inggris tetapi menggunakan inputan berbahasa Indonesia[1].

Pada CLIR terdapat berbagai metode terjemahan, metode yang digunakan untuk CLIR adalah *Dictionary-based*, *Parallel corpora based*, *Comparable corpora based*, *Machine translator based*, *Ontology based*, *Transitive based*. Langkah-langkahnya menerjemahkan *query* ke dalam bahasa target, tahap stemming merupakan proses penguraian suatu term ke term asli atau menjadi bentuk dasarnya, lalu perhitungan kemiripan[2] [3].

M.Saravanan dkk pada tahun 2016 melakukan penelitian *Cross Lingual Information Retrieval* bahasa Tamil Inggris dalam pencarian dokumen tentang pertanian menggunakan Metode *Query Translation Word Sense Disambigution* (WSD) dan *Word by Word Translation* (WBWT) menggunakan VSM. Tingkat precission yang di peroleh translasi WSD 0,97 dan WBWT 0,72 [15]

Namun, permasalahannya terdapat pada efisiensi waktu dan akurasi dalam penerjemahan *query* dari banyaknya metode di CLIR pada Bahasa Indonesia dan Inggris. Pada tahap stemming mempunyai metode masing-masing dari kedua Bahasa tersebut. Metode stemming sudah banyak dikembangkan contohnya seperti *Porter Stemmer*. Untuk stemming bahasa Indonesia terdapat kata dasar dan kata imbuhan. Ada variasi imbuhan termasuk *prefix* (imbuhan yang terletak di awal kata dasar), *suffix* (imbuhan yang terletak di akhir kata dasar), *infiks*(imbuhan yang disisipkan di tengah kata dasar), dan *confixes*(imbuhan yang terletak di awal dan akhir kata dasar sekaligus), sedangkan stemming bahasa Inggris hanya kata imbuhan *suffix* [4][5][6]. struktur morfologi dalam bahasa Indonesia memiliki tingkat kerumitan yang lebih tinggi daripada bahasa Inggris. Seperti misalnya,

kesulitan dalam membedakan suatu kata yang mengandung imbuhan baik prefix maupun sufiks dengan suatu kata dasar yang salah satu suku katanya merupakan bagian dari imbuhan, sehingga membutuhkan kamus kata dasar dan memakan memori yang menyebabkan proses yang lama [7].

Pada penelitian sebelumnya Muhammad Alif [8] menggunakan stemming *Porter Stemmer* bahasa Indonesia dokumen uji coba diambil dari penelitian Tala dengan uji coba 300 dokumen, porter hanya mampu menghasilkan 61%[8]. Untuk melakukan perhitungan persamaan dokumen, digunakanlah metode cosine similarity dan pembobotan menggunakan TF-IDF. Metode cosine similarity digunakan karena metode ini mempunyai nilai akurasi yang tinggi dimana kelebihan utamanya adalah tidak terpengaruh pada panjang pendeknya suatu dokumen dan pembobotan TF-IDF digunakan karena metode ini merupakan metode yang paling umum digunakan pada information retrieval untuk perhitungan bobot suatu kata. Metode ini juga terkenal efisien, mudah dan memiliki hasil yang akurat [9] [10] .

Metode yang digunakan peneliti di atas menggunakan metode penerjemahan CLIR dengan stemming dua Bahasa Indonesia dan Inggris, dimana memiliki kelemahan yaitu efisien waktu terlalu lama. Untuk mengatasi masalah tersebut, di usulkan menggunakan *Google Translate API* dalam penerjemahan query dengan maksimal 5000 karakter dan menggunakan satu stemming bahasa Inggris dengan metode Porter Stemmer yang merupakan metode stemming terbaik dari beberapa penelitian dengan akurasi terbaik. Untuk perhitungan kemiripan menggunakan *Cosine Similarity*. penelitian ini diharapkan dapat menghasilkan sistem yang lebih akurat dan dapat mempercepat waktu dalam proses dan pencarian dokumen lintas bahasa Indonesia Inggris.

1.2 Rumusan Masalah

1.2.1 Permasalahan

Proses CLIR perlu tahapan penerjemahan, *preprocessing*, dan *retrieval* yang memungkinkan membutuhkan proses panjang. Stemming dari kedua bahasa tersebut masih memiliki kekurangan terutama teks bahasa Indonesia tingkat akurasi, sehingga proses untuk melakukan *retrieval* lintas bahasa memakan waktu yang cukup lama dan tidak efisien.

1.2.2 Usulan Solusi

Pemanfaatan *Google Translate API* untuk penerjemahan dokumen dan *query* bahasa Indonesia ke bahasa Inggris dan menerapkan stemming bahasa Inggris menggunakan metode *Porter Stemmer* dengan perhitungan kemiripan menggunakan *Cosine Similarity*

1.2.3 Pertanyaan Penelitian

1. Seberapa tepat hasil translasi Indonesia-Inggris menggunakan *Google Translate API*?
2. Berapa presisi untuk sistem temu kembali informasi lintas bahasa dengan menggunakan perhitungan kemiripan *Cosine Similarity*?

1.3 Tujuan & Manfaat

1.3.1 Tujuan Penelitian

Penerapan *Google Translate API* dalam penerjemahan *query* dan menggunakan satu stemming bahasa Inggris dengan metode *Porter Stemmer* untuk memperoleh hasil yang lebih akurat dan dapat mempercepat waktu dalam pencarian dokumen lintas bahasa Indonesia Inggris

1.3.2 Manfaat

Dihasilkan aplikasi pencarian lintas bahasa yang mampu mencari dokumen yang lebih cepat dan akurat daripada penelitian sebelumnya

1.4 Batasan-Batasan Masalah

Batasan masalah pada penelitian ini adalah sebagai berikut :

1. Dokumen yang akan diuji merupakan dokumen teks
2. Teks yang digunakan sebagai dokumen uji merupakan dokumen yang menggunakan bahasa Indonesia dan bahasa Inggris
3. Data dokumen yang akan digunakan hanya mengambil isi berita pariwisata di madura.
4. Hasil akhir dalam penelitian skripsi ini adalah menemukan artikel dokumen yang cocok dan sesuai dengan kata kunci
5. Translasi menggunakan *Google Translate API* versi gratis dan batas maksimal karakter *Google Translate API* sebanyak 15000 karakter
6. Proses *preprocessing* menggunakan library NLTK python

1.5 Sistematika Proposal

Sistematika penulisan Proposal ini terdiri dari beberapa bab, yaitu :

BAB I PENDAHULUAN

Pada bab ini membahas tentang latar belakang, rumusan masalah, tujuan, manfaat, batasan masalah, serta sistematika penulisan laporan.

BAB II KAJIAN PUSTAKA

Pada bab ini membahas tentang landasan teori dan penelitian sebelumnya.

BAB III METODE USULAN

Pada bab ini membahas tentang metode yang diusulkan (algoritma), arsitektur (Input, Proses, Ouput), Dataset, tahapan penelitian, serta rencana pengujian.

BAB IV HASIL DAN PEMBAHASAN

Bab ini menjelaskan hasil dari penelitian yang dilakukan. Hasil yang diperoleh berupa kekurangan, kelebihan, dan uji coba yang telah dilakukan.

BAB V PENUTUP

Bab ini berisi kesimpulan dari hasil uji coba serta saran yang dapat digunakan untuk pengembangan penelitian berikutnya.