RANCANG BANGUN APLIKASI INFORMATION RETRIEVAL UNTUK MENGKOLEKSI DATA PARALEL KORPUS TEKS BAHASA INGGRIS – BAHASA INDONESIA

Edy Septiandri

Program Studi Teknik Informatika
Jurusan Teknik Elektro Fakultas Teknik Universitas Tanjungpura
edy.kumkum09@gmail.com

Abstrak - Paralel korpus merupakan dua dokumen teks yang saling berhubungan dimana dokumen teks pertama berisi kumpulan kalimat sumber dan dokumen teks kedua berisi kumpulan kalimat terjemahannya. Paralel korpus berfungsi sebagai sumber utama dalam mengembangkan sebuah Mesin Penerjemah Statistik. Hasil terjemahan dari sebuah Mesin Penerjemah Statistik tergantung pada jumlah dari paralel korpus yang tersedia. Pembuatan paralel korpus secara manual tidaklah mudah, karena akan memakan waktu yang lama, memerlukan biaya yang tidak sedikit dan jumlah dokumen yang terbatas. Sistem temu balik informasi atau information retrieval dapat membantu dalam hal mengelola dan menemukan kembali dokumen secara cepat dan efektif. Sistem ini dibuat untuk mengkoleksi data parlel korpus khususnya bahasa Indonesia dan Inggris, dimana output dari aplikasi ini akan menambah perbendaharaan paralel korpus bahasa Indonesia-Inggris. Sistem ini mampu untuk mengumpulkan dokumen secara otomatis dari sebuah website yang telah ditentukan, dan sasarannya hanya dokumen yang berbahasa Indonesia dan Inggris. Dokumen yang telah terkumpul akan dibersihkan dari semua tanda baca yang tidak diperlukan dengan metode tokenization, setelah itu dokumen tersebut akan diproses kembali untuk memisahkan antara kalimat sumber dan kalimat terjemahannya dengan metode parse. Setelah semua proses selesai maka akan menghasilkan sebuah paralel korpus. Aplikasi information retrieval ini akan mempermudah dalam pembuatan sebuah paralel korpus dan akan memperkaya perbendaharaan paralel korpus bahasa Indonesia-Inggris.

Kata kunci: paralel korpus, mesin penerjemah statistik, sistem temu balik informasi, tokenization, parse.

1. Pendahuluan

Semakin canggihnya teknologi di bidang komputasi dan telekomunikasi khususnya internet membuat informasi mudah untuk didapatkan oleh banyak orang. Kemudahan ini mendorong pertambahan jumlah informasi digital menjadi semakin banyak dan beragam. Informasi dapat berupa berita, dokumen, surat, cerita, laporan penelitian, data keuangan, dan lain-lain. Tidak dapat dipungkiri lagi informasi telah menjadi komoditi yang paling penting dalam dunia modern masa kini.

Bertambahnya jumlah dokumen teks yang dapat diakses di internet diikuti dengan meningkatnya kebutuhan pengguna akan perangkat pencarian informasi yang efektif dan efesien. Sistem temu balik informasi bertujuan mengahasilkan dokumen yang paling relevan dan sesuai dengan keinginan pengguna. Tujuan dari sistem temu balik informasi ini adalah untuk memenuhi kebutuhan informasi pengguna dengan me-*retrieve* semua dokumen yang mungkin relevan, pada waktu yang sama me-*retrieve* sesedikit mungkin dokumen yang tidak relevan. Dokumen-dokumen tersebut akan memiliki banyak manfaat bagi penggunanya, sebagai salah satu contoh yaitu sebagai bahan dasar dalam pembuatan paralel korpus.

Paralel korpus merupakan suatu sumber utama dalam mengembangkan sebuah sistem penerjemah statistik dan aplikasi NLP (*Natural Language Processing*) lainnya. Hasil terjemahan dari sebuah sistem penerjemah statistik tergantung pada jumlah dari paralel korpus yang tersedia, karena jika semakin banyak jumlah paralel korpus yang dimiliki, maka akan semakin baik pula hasil terjemahannya. Jumlah paralel korpus *identic* untuk bahasa Indonesia – bahasa Inggris saat ini sudah tersedia sebanyak 27.326 kalimat [6]. Pembuatan korpus secara manual akan memakan waktu yang lama, memerlukan biaya yang tidak sedikit, dan jumlah dokumen yang terbatas. Oleh karena itu diperlukan sebuah sistem yang dapat membuat sebuah paralel korpus dengan cepat dan efektif agar dapat menambah perbendaharaan korpus yang sudah ada.

Berdasarkan pada permasalahan di atas, penulis akan melakukan analisis, perancangan dan pembuatan aplikasi yang difokuskan untuk menghasilkan paralel korpus dengan

menggunakan metode *information retrieval*. Pada penelitian ini, *website* yang akan di *crawling* adalah sebuah *website* berita milik BBC Indonesia. Dengan adanya sistem ini diharapkan dapat memperkaya perbendaharaan korpus dalam bahasa Indonesia dan Inggris.

2. Tinjauan Pustaka

2.1 *Information Retrieval*

Pencarian informasi atau yang dikenal dengan istilah sistem temu balik informasi (Information Retrieval) digunakan untuk menemukan kembali (retrieve) secara otomatis informasi- informasi yang relevan terhadap kebutuhan pengguna dari suatu kumpulan informasi [3]. Tujuan dari sistem sistem temu balik informasi adalah memenuhi kebutuhan informasi pengguna dengan me-*retrieve* semua dokumen yang mungkin relevan, pada waktu yang sama me-*retrieve* sesedikit mungkin dokumen yang tak relevan.

2.2 *Crawler*

Crawler merupakan program yang berjalan secara otomatis, berisi *script* program yang melakukan crawling melalui halaman *web*site untuk mengumpulkan data berdasarkan indeks dari halaman *web* yang ditemukan [4]. Tujuan dari *crawler* adalah dengan cepat dan efisien mengumpulkan banyak informasi dari halaman *web* yang berguna, berikut dengan struktur link yang terkoneksi dengan halaman *web* tersebut.

2.3 *Parser*

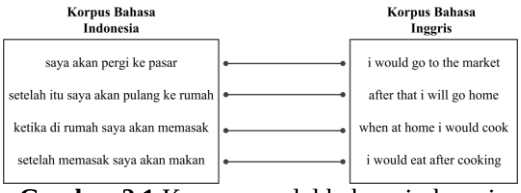
Parser adalah sebuah sistem dasar yang memungkinkan pemahaman yang lebih baik terhadap kalimat dalam bahasa tertentu [2]. Proses yang dilakukan oleh parser disebut *parsing*. *Parsing* merupakan proses pengambilan kata-kata dari kumpulan dokumen. Tujuan utama dari *parsing* adalah memeriksa apakah urutan *token* yang dihasilkan sesuai dengan tata bahasa dari bahasa yang bersangkutan.

2.4 Tokenisasi

Tokenisasi merupakan proses pemisahan suatu rangkaian karakter berdasarkan karakter spasi, dan mungkin pada waktu yang bersamaan dilakukan juga proses penghapusan karakter tertentu, seperti tanda baca [1]. Proses ini cukup rumit untuk sebuat program komputer karena beberapa karakter dapat dijadikan sebagai pembatas (*delimiter*) dari token-token itu sendiri. Pembatas dari token tersebut antara lain spasi, tab dan baris baru, sedangkan karakter ( ) < > ! ? “ . , terkadang dianggap sebagai pembatas dan juga bukan pembatas tergantung pada kondisi pemakainya.

2.5 Korpus

Korpus adalah repositori dari kumpulan materi bahasa alami, seperti teks, paragraf, dan kalimat dari satu atau banyak bahasa [5]. Paralel korpus terdiri dari teks yang sama di lebih dari satu bahasa. Kesejajaran paralel korpus dijelaskan untuk menunjukkan secara tepat kalimat dari bahasa sumber sesuai dengan kalimat dari teks sasaran.

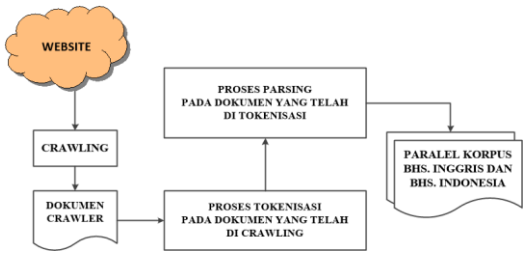


Gambar 2.1 Korpus paralel bahasa indonesia dan bahasa inggris

3. Perancangan Sistem

3.1 Perancangan Arsitektur Sistem

Gambaran umum sistem yang akan dibuat digambarkan melalui desain arsitektur sistem dapat dilihat pada Gambar 3.1.



Gambar 3.1 Desain arsitektur sistem

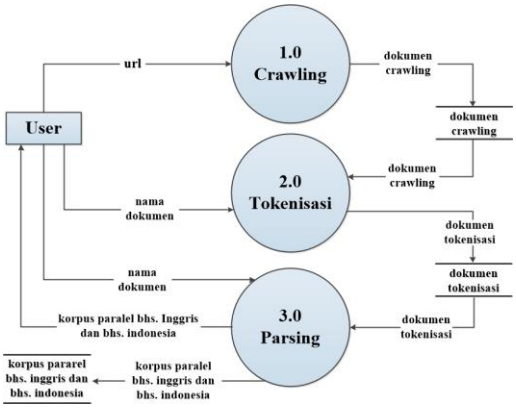
3.2 Perancangan Diagram Konteks

Diagram konteks sistem temu bakik informasi untuk mengkoleksi data paralel korpus teks bahasa Inggris dan bahasa Indonesia digambarkan pada gambar 3.2. Berdasarkan gambar diagram tersebut, entitas luar yang terlibat langsung dalam sistem adalah *user*. Sistem ini merupakan sistem *single user*, karena hanya satu *user* yang dapat menggunakannya.

Gambar 3.2 Diagram konteks sistem

3.3 Perancangan Diagram *Overview*

Diagram overview berisi penjelasan urutan-urutan proses dari diagram konteks digambarkan pada gambar 3.3. Pada diagram ini proses dibagi menjadi 3 proses, yaitu *crawling*, tokenisasi dan *parsing*.



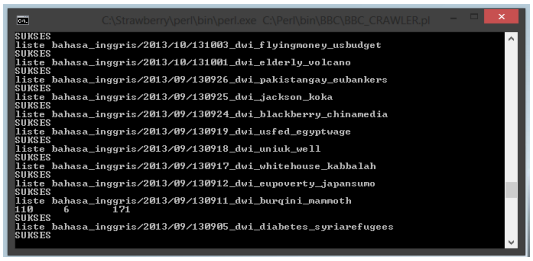
Gambar 3.3 Diagram *overview* sistem

4. Hasil Perancangan

4.1 Tampilan Proses *Crawler*

Proses ini dirancang menggunakan PERL sebagai bahasa pemrograman, Strawberry PERL sebagai *tools* untuk menjalankan bahasa pemrograman PERL di Windows dan Padre sebagai PERL *editor*. Cara menjalankan proses *crawling* adalah sebagai berikut:

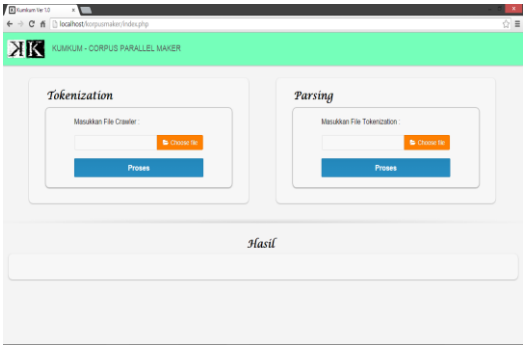
1. Buka *file script crawler* di Padre PERL *editor*.
2. Setelah *file script* terbuka, klik tombol pada *toolbar* Padre PERL *editor* untuk menjalankan *script crawler*.
3. Setelah tombol di klik maka proses *crawling* akan berjalan pada *tools* command prompt.



Gambar 4.1 Tampilan proses *crawling* di command prompt

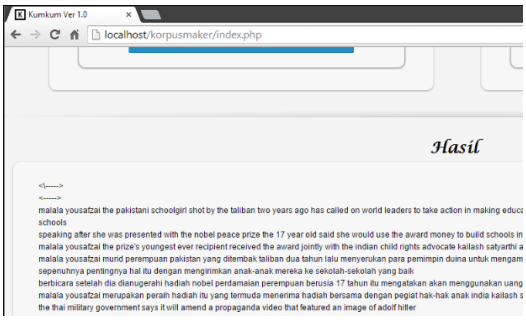
4.2. Halaman Tokenisasi dan *Parsing*

Halaman *tokenization* dan *parsing* berisi *form* untuk melakukan dua proses yaitu proses *tokenization* dan proses *parsing*. Proses yang harus dilakukan pertama kali di antara kedua proses itu adalah proses *tokenization*, setelah itu proses *parsing*. Antarmuka halaman *tokenization* dan *parsing* dapat dilihat pada gambar 4.2.



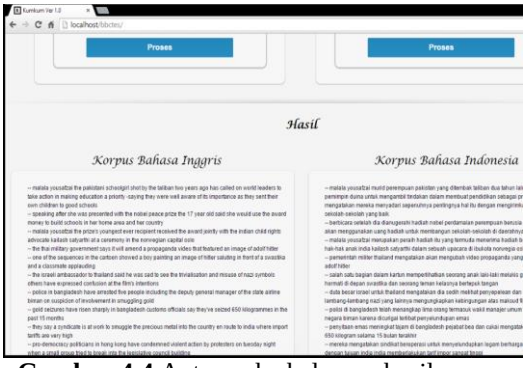
Gambar 4.2 Antarmuka halaman tokenisasi dan *parsing*

Proses *tokenization* dilakukan dengan cara memasukkan *file* hasil *crawling* yang berasal dari direktori komputer, jika sudah tahap selanjutnya klik tombol proses pada *form tokenization* dan hasilnya akan terlihat di bagian *form* hasil seperti pada gambar 4.3.



Gambar 4.3 Antarmuka halaman hasil proses tokenisasi

Proses *parsing* dilakukan dengan cara memasukkan *file* hasil *tokenization* yang berasal dari direktori komputer, jika sudah tahap selanjutnya klik tombol proses pada *form parsing* dan hasilnya akan terlihat di bagian *form* hasil seperti pada gambar 4.4.



Gambar 4.4 Antarmuka halaman hasil proses *parsing*

Hasil dari proses *tokenization* dan *parsing* akan disimpan dalam sebuah file dengan format txt yang berada di *folder* yang sudah ditentukan sebelumnya. Data hasil dari proses *tokenization* tersimpan di dalam file yang bernama Hasil_Tokenisasi.txt dan data hasil dari proses *parsing* tersimpan di dalamdua file, yaitu Korpus_Bhs_Indonesia.txt dan Korpus_Bhs_Inggris.txt

5. Kesimpulan dan Saran

5.1 Kesimpulan

Berdasarkan hasil analisis dan pengujian terhadap Sistem Temu Balik Informasi untuk menghasilkan paralel korpus bahasa Inggris-Indonesia, maka dapat ditarik kesimpulan sebagai berikut:

1. Sistem mampu mengumpulkan dokumen artikel berita melalui proses *crawling website* dari situs milik BBC dengan alamat URL http://www.bbc.co.uk/indonesia/topik/dwibahasa/, sehingga menghasilkan sebuah dokumen yang berisi kumpulan artikel berbahasa Inggris sebagai bahasa sumber dan berbahasa Indonesia sebagai bahasa terjemahan.
2. Sistem dapat melakukan proses tokenisasi untuk menghilangkan semua tanda baca yang tidak diperlukan, dan proses *parsing* untuk menghilang semua dokumen yang tidak relevan dan memisahkan antara kalimat bahasa Inggris dan bahasa Indonesia ke dalam dua dokumen yang berbeda yaitu, dokumen korpus bahasa Inggris dan dokumen korpus bahasa Indonesia. Kedua dokumen tersebut merupakan paralel korpus teks bahasa Inggris-Indonesia.
3. Sistem temu balik ini telah menghasilkan paralel korpus bahasa Inggris dan bahasa Indonesia sebanyak 1541 kalimat, sehingga dapat menambah perbendaharaan paralel korpus bahasa Inggris dan bahasa Indonesia yang sudah ada, yaitu dari 27.326 kalimat menjadi 28.867 kalimat.
4. Pembuatan paralel korpus menggunakan aplikasi *Information Retrieval* jauh lebih cepat dibanding dengan pembuatan paralel korpus dengan cara manual.

5.2 Saran

Adapun beberapa hal yang perlu ditambahkan dalam pengembangan sistem ini adalah pada proses *crawling*, yaitu perlu pemberian inisial untuk setiap kalimat berbahasa Inggris dan kalimat berbahasa Indonesia.

Referensi

[1] Amin, Fakhtul. 2012. Sistem Temu Balik Informasi Dengan Metode Vector Space Model. Jurnal Sistem Informasi Bisnis. Unduh: http://ejournal.undip.ac.id/index.php/jsinbis/article/downloadSuppFile/37/303

[2] Gusmita, R. H. dan Manurung, R. 2008. Some initial experiments with indonesian probabilistic parsing. Malaysia: MALINDO Workshop.

[3] Hadhiatma, Agung. 2010. Pencarian Dokumen Berdasarkan Kombinasi Antara Model Ruang Vektor Dan Model Domain Ontologi. Yogyakarta: semnasIF.

[4] Sasongko, Jati. 2010. Aplikasi untuk Membangun Corpus dari Data Hasil Crawling dengan Berbagai Format Data Secara Otomatis. Jurnal Teknologi Informasi Dinamik. Unduh: http://www.unisbank.ac.id/ojs/index.php/fti1/article/download/107/102

[5] Siagian, Adelina Irmadewita. 2012. Implementasi Corpus Generator Dengan Parallel Text. Unduh: http://repository.usu.ac.id/bitstream/123456789/33897/4/Chapter%20II.pdf

[6] Sujaini, Herry. 2012. Korpus Paralel Indonesia - Inggris. Unduh: http://herrysujaini.blogspot.com/2012/05/korpus-paralel-indonesia-inggris.html