

Laporan Tugas Akhir

**Penemuan Jawaban pada Sistem Tanya
Jawab Bahasa Indonesia-Inggris dengan
Pembobotan Kata dan Informasi dari Internet**



Oleh:

Septian Adiwibowo

1203001036

Fakultas Ilmu Komputer

Universitas Indonesia

Depok, Indonesia

Februari 2008

Lembar Pengesahan

Nama : Septian Adiwibowo
NPM : 1203001036
Judul Tugas Akhir : Penemuan Jawaban pada Sistem Tanya Jawab Bahasa
Indonesia-Inggris dengan Pembobotan Kata dan Informasi
dari Internet

Tugas Akhir ini telah diperiksa dan disetujui.

Depok, 4 Februari 2008

Mirna Adriani, PhD.
Dosen Pembimbing

Kata Pengantar

Puji Tuhan penulis telah berhasil menyelesaikan Tugas Akhir ini dengan perjuangan yang panjang. Oleh karena itu, pada kesempatan yang sangat berbahagia ini penulis ingin mengucapkan terima kasih kepada pihak-pihak berikut ini:

1. Allah Tuhan Ibrahim, Ishak, dan Ismail.
2. Bapak dan Ibu atas segala kebaikan dalam mendukung, membantu, dan menyemangati (hampir) semua kegiatan yang penulis lakukan.
3. Bu Mirna Adriani. Terima kasih atas semua nasihat, petunjuk, bimbingan, dan ilmu yang telah diberikan. Terima kasih atas kesempatan yang diberikan untuk mengikuti QA@CLEF 2007.
4. Pak Ruli Manurung. Terima kasih atas semua saran, ide, inspirasi, bantuan teknis, dan hiburan yang telah diberikan selama ini.
5. Pak Lim Yohanes Stefanus. Terima kasih atas bimbingan selama empat setengah tahun masa perkuliahan penulis di Fakultas Ilmu Komputer Universitas Indonesia ini.
6. Tina, Okky, Markus, Mbak Lily, Herika, Amel, Mbak Syandra, Mamad, Nasikhin, dan masih banyak lagi teman-teman lainnya yang telah membantu keberhasilan tugas akhir penulis.

Penulis menyadari bahwa masih terdapat kekurangan dalam penulisan Laporan Tugas Akhir ini, karena itulah penulis sangat terbuka terhadap segala masukan, kritik, saran, pertanyaan, dan kesan-kesan. Pembaca yang terhormat dapat menghubungi penulis melalui alamat surat elektronik septian.adiwibowo@gmail.com. Akhir kata, semoga laporan ini bermanfaat bagi yang membacanya.

Depok, 4 Februari 2008

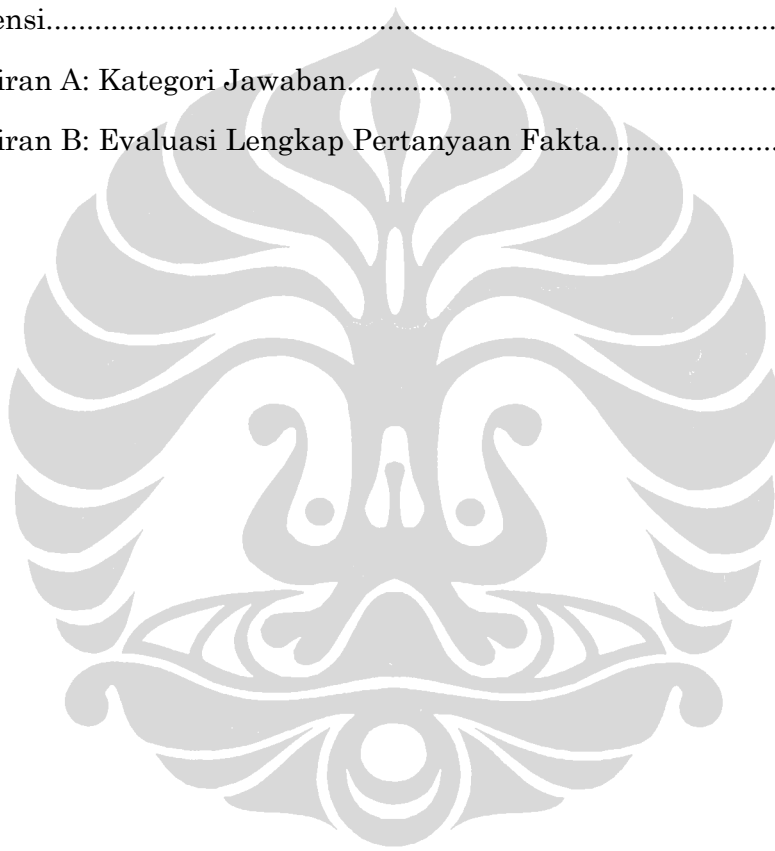
Penulis

Daftar Isi

Lembar Pengesahan.....	ii
Abstrak.....	iii
Kata Pengantar.....	iv
Daftar Isi.....	v
Daftar Gambar.....	viii
Daftar Tabel.....	x
Bab 1 Pendahuluan.....	1
1.1 Latar Belakang.....	1
1.2 Rumusan Masalah.....	3
1.3 Tujuan dan Ruang Lingkup.....	4
1.4 Metodologi Penelitian.....	4
1.5 Sistematika Penulisan.....	5
Bab 2 Landasan Teori.....	6
2.1 Perolehan Informasi.....	6
2.2 Model Perolehan Informasi.....	7
2.2.1 Model Inference Networks.....	8
2.2.2 Model Bahasa.....	12
2.3 Sistem Perolehan Informasi.....	15
2.3.1 Identifikasi Jenis Dokumen.....	16
2.3.2 Pembuangan Stopwords.....	16
2.3.3 Stemming.....	17
2.3.4 TF-IDF.....	17
2.3.5 Pembuatan Indeks.....	18
2.4 Evaluasi Perolehan Informasi.....	20
2.5 Teknik-Teknik Lain.....	22
2.5.1 Perluasan Kueri.....	22
2.5.2 Pengenalan Entitas Bernama.....	25
2.6 Cross-Language Evaluation Forum (CLEF).....	26
2.7 Sistem Tanya Jawab.....	27

2.7.1 Analisis Pertanyaan.....	28
2.7.2 Perolehan Berbasis Passage.....	29
2.7.3 Ekstraksi Jawaban.....	30
2.8 Sistem Tanya Jawab Bilingual.....	32
2.9 Evaluasi Sistem Tanya Jawab.....	32
2.10 Constituency Tree.....	34
2.11 Penelitian-Penelitian di Bidang Sistem Tanya Jawab.....	38
2.11.1 Korea University Question Answering System (KUQA).....	38
2.11.2 DFKI Language Techology Lab.....	41
2.11.3 Tim Fakultas Ilmu Komputer Universitas Indonesia.....	44
Bab 3 Eksperimen.....	47
3.1 Pemrosesan Awal.....	48
3.1.1 Perangkat Lunak Pendukung.....	48
3.1.2 Pemrosesan Koleksi Dokumen.....	49
3.2 Analisis Pertanyaan.....	52
3.3 Penerjemahan.....	54
3.4 Perolehan Passage.....	56
3.5 Anotasi Entitas Bernama.....	60
3.6 Ekstraksi Jawaban Entitas Bernama.....	61
3.6.1 Pencarian Dari Google.....	62
3.6.2 Skor Perolehan Lemur.....	63
3.6.3 Average Distance Weight.....	65
3.6.4 Pembobotan Kata dengan TF-IDF.....	67
3.6.5 Nilai Akhir Kandidat Jawaban.....	68
3.7 Ekstraksi Jawaban Definisi.....	70
Bab 4 Evaluasi dan Analisis.....	74
4.1 Evaluasi Sistem Tanya Jawab.....	74
4.1.1 Evaluasi Jawaban Fakta.....	75
4.1.2 Evaluasi Jawaban Definisi.....	77
4.2 Mengapa Skornya Begitu Rendah?.....	79
4.2.1 Bagian Analisis Pertanyaan.....	79
4.2.2 Penerjemahan.....	81
4.3 Analisis Jawaban Definisi.....	82
4.4 Analisis Jawaban Fakta.....	83

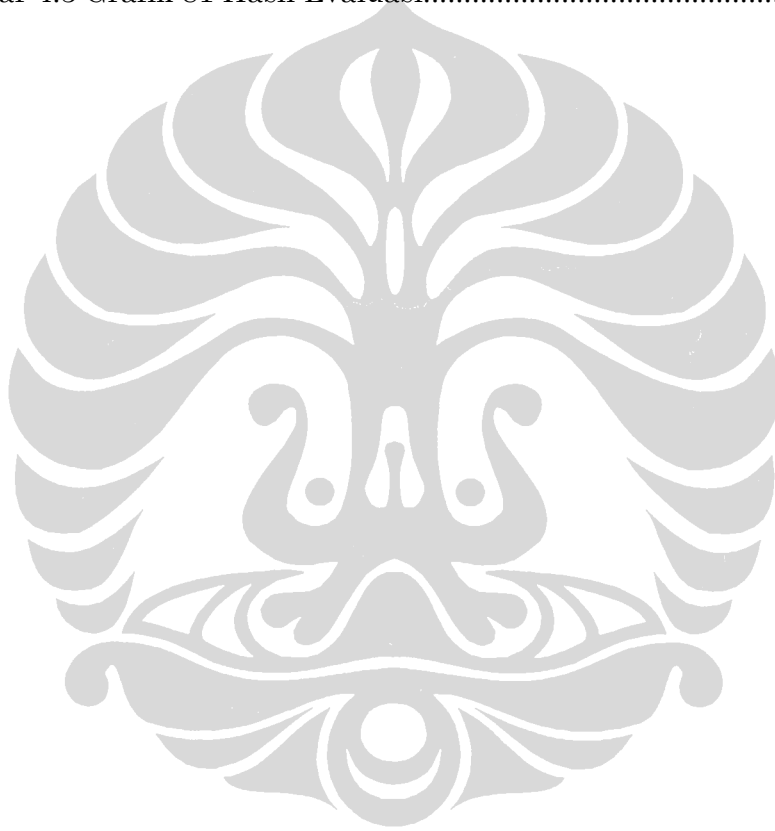
4.4.1 Kesalahan Pengenalan Entitas Bernama.....	84
4.4.2 Meneliti Daftar Kandidat Jawaban.....	85
4.4.3 Analisis Formula G, R, W, dan T.....	87
4.4.4 Modifikasi Skor T.....	90
4.5 Perbandingan Dengan Peserta CLEF.....	91
Bab 5 Kesimpulan, Saran, dan Penutup.....	94
5.1 Kesimpulan.....	94
5.2 Saran Untuk Pengembangan Selanjutnya.....	95
5.3 Penutup.....	96
Referensi.....	97
Lampiran A: Kategori Jawaban.....	101
Lampiran B: Evaluasi Lengkap Pertanyaan Fakta.....	106



Daftar Gambar

Gambar 2.1 Contoh Inference Networks.....	9
Gambar 2.2 Diagram Sistem Perolehan Informasi.....	15
Gambar 2.3 Relevan, Diperoleh, dan Relevan Diperoleh.....	21
Gambar 2.4 User Relevance Feedback.....	23
Gambar 2.5 Pseudo Relevance Feedback.....	24
Gambar 2.6 Contoh Dokumen.....	25
Gambar 2.7 Contoh Dokumen yang Telah Dianotasi.....	25
Gambar 2.8 Passage Untuk Penghitungan ADW.....	31
Gambar 2.9 Contoh Constituency Tree.....	35
Gambar 2.10 Pemberian Anotasi Part-of-Speech.....	35
Gambar 2.11 Kalimat Contoh dengan Pengelompokan Frasa Kata Benda.....	37
Gambar 2.12 Anotasi Sintaktik untuk “This apple pie looks good and a real treat”.....	37
Gambar 3.1 Diagram Alur Sistem.....	47
Gambar 3.2 Struktur Dokumen TREC Text.....	50
Gambar 3.3 Contoh Dokumen TREC Text.....	50
Gambar 3.4 Menjalankan IndriBuildIndex.....	51
Gambar 3.5 Berkas Konfigurasi Untuk IndriBuildIndex.....	51
Gambar 3.6 Contoh Pola Pertanyaan.....	52
Gambar 3.7 Sandi Semu Modul Analisis Pertanyaan.....	53
Gambar 3.8 Sandi Semu Modul Penerjemahan.....	56
Gambar 3.9 Sandi Semu Modul Perolehan Passage.....	59
Gambar 3.10 Contoh Anotasi.....	60
Gambar 3.11 Sandi Semu Modul Anotasi Entitas Bernama.....	61
Gambar 3.12 Contoh Pencarian Dari Google.....	63
Gambar 3.13 Contoh Isi Passage Teratas.....	65
Gambar 3.14 Contoh Passage.....	66

Gambar 3.15 Dokumen Untuk Ilustrasi Pembobotan Dengan TF-IDF.....	67
Gambar 3.16 Sandi Semu Modul Ekstraksi Jawaban Entitas Bernama.....	69
Gambar 3.17 Contoh Passage Untuk Pertanyaan Definisi.....	71
Gambar 3.18 Hasil Analisis Sintaktik APP.....	71
Gambar 3.19 Struktur Constituency Tree.....	71
Gambar 3.20 Sandi Semu Modul Ekstraksi Jawaban Definisi.....	73
Gambar 4.1 Contoh Jawaban Right Yang Dievaluasi Sebagai Unsupported (1).....	76
Gambar 4.2 Contoh Jawaban Right Yang Dievaluasi Sebagai Unsupported (2).....	77
Gambar 4.3 Grafik 81 Hasil Evaluasi.....	89



Daftar Tabel

Tabel 2.1 TF-IDF Untuk Setiap Kata (1).....	10
Tabel 2.2 TF-IDF Untuk Setiap Kata (2).....	10
Tabel 2.3 Nilai Kepercayaan Kata Terhadap Dokumen.....	11
Tabel 2.4 Pembobotan Kata (Link Matrix).....	11
Tabel 2.5 Contoh Indeks Terbalik.....	19
Tabel 2.6 Indeks Terbalik dengan Pembuangan Stopwords dan Stemming.....	20
Tabel 2.7 Dokumen Diperoleh oleh Sistem A.....	22
Tabel 2.8 Contoh Pertanyaan Kategori Fakta.....	28
Tabel 2.9 Contoh Pertanyaan Kategori Definisi.....	29
Tabel 2.10 Contoh Pertanyaan Kategori Daftar.....	29
Tabel 2.11 Contoh Evaluasi Jawaban.....	33
Tabel 2.12 Kategori Pertanyaan.....	38
Tabel 2.13 Evaluasi KUQA.....	41
Tabel 2.14 QUANTICO Subtopik Jerman-Jerman.....	43
Tabel 2.15 QUANTICO Subtopik Inggris-Jerman.....	43
Tabel 2.16 QUANTICO Subtopik Jerman-Inggris.....	43
Tabel 2.17 Kata Tanya untuk Klasifikasi Pertanyaan.....	44
Tabel 2.18 Evaluasi Sistem Tanya Jawab Tim Fasilkom UI.....	46
Tabel 3.1 Tujuh Kombinasi Mesin Penerjemah.....	55
Tabel 3.2 Prosentase Kemiripan Hasil Penerjemahan Mesin dengan Manusia.....	55
Tabel 3.3 Variasi Perluasan Kueri dengan Pseudo Relevance Feedback.....	57
Tabel 3.4 Variasi Perluasan Kueri dengan Google.....	58
Tabel 3.5 Recall dengan Pseudo Relevance Feedback.....	58
Tabel 3.6 Recall dengan Google.....	59
Tabel 3.7 Contoh Skor Penilaian.....	64
Tabel 3.8 Jarak Kandidat Jawaban Dengan Kata-Kata Kueri.....	66

Tabel 3.9 Pola Frasa Untuk Pertanyaan Definisi.....	70
Tabel 4.1 Lima Hasil Terbaik Untuk Pertanyaan Fakta.....	75
Tabel 4.2 Dua Jawaban Definisi yang Sebetulnya Benar.....	78
Tabel 4.3 Hasil Evaluasi Sistem Tanya Jawab.....	79
Tabel 4.4 Pertanyaan yang Tidak Bisa Diklasifikasikan.....	80
Tabel 4.5 Pertanyaan yang Diklasifikasikan secara salah.....	80
Tabel 4.6 Kesalahan Pada Analisis Pertanyaan.....	81
Tabel 4.7 Contoh Penerjemahan.....	81
Tabel 4.8 Contoh Kesalahan Jawaban untuk Pertanyaan Definisi.....	83
Tabel 4.9 Evaluasi Pengenalan Entitas Bernama.....	84
Tabel 4.10 Kandidat Jawaban Di Urutan Non-Pertama.....	86
Tabel 4.11 Peranan G, R, W, dan T.....	87
Tabel 4.12 Lima Hasil Terbaik Untuk Pertanyaan Fakta.....	88
Tabel 4.13 Lima Hasil Terendah Untuk Pertanyaan Fakta.....	89
Tabel 4.14 Lima Hasil Terbaik Untuk Pertanyaan Fakta.....	91
Tabel 4.15 Evaluasi 7 Hasil Terbaik Topik Tanya Jawab CLEF.....	92
Tabel B.1 Evaluasi Lengkap.....	106

