

**Pengembangan *Part of Speech Tagger* untuk Bahasa Indonesia
Berdasarkan Metode *Conditional Random Fields* dan
*Transformation Based Learning***

SKRIPSI

**Triastuti Chandrawati
1204000866**



UNIVERSITAS INDONESIA
FAKULTAS ILMU KOMPUTER
DEPOK
JULI, 2008

**Pengembangan *Part of Speech Tagger* untuk Bahasa Indonesia
Berdasarkan Metode *Conditional Random Fields* dan
*Transformation Based Learning***

SKRIPSI

Diajukan sebagai salah satu syarat untuk memperoleh gelar Sarjana Ilmu
Komputer

**Triastuti Chandrawati
1204000866**

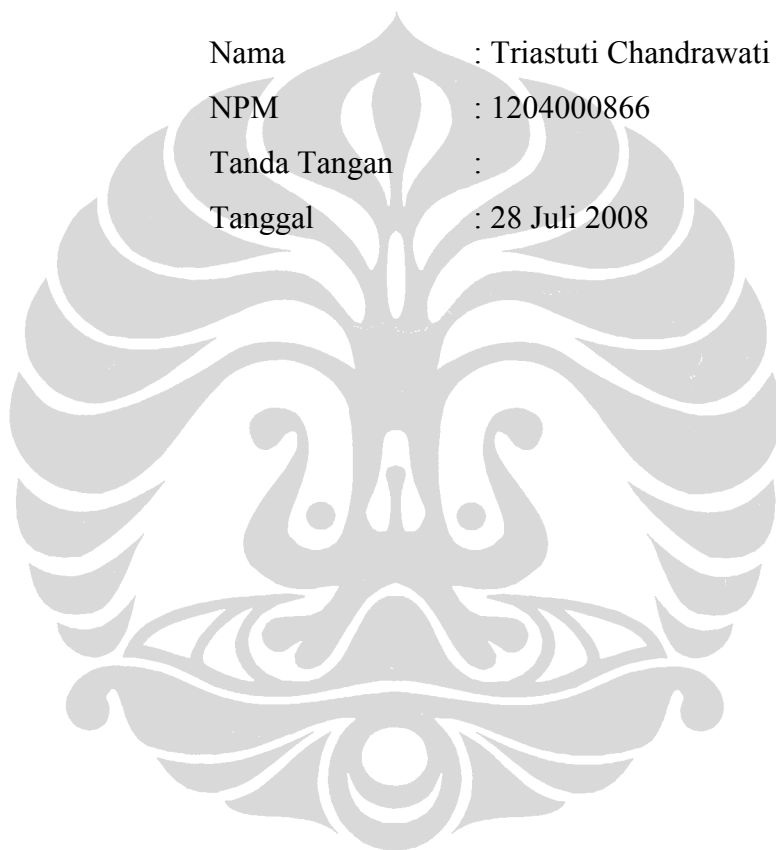


UNIVERSITAS INDONESIA
FAKULTAS ILMU KOMPUTER
DEPOK
JULI, 2008

HALAMAN PERNYATAAN ORISINALITAS

Skripsi ini adalah hasil karya saya sendiri, dan semua sumber baik yang dikutip maupun dirujuk telah saya nyatakan dengan benar.

Nama : Triastuti Chandrawati
NPM : 1204000866
Tanda Tangan :
Tanggal : 28 Juli 2008



HALAMAN PENGESAHAN

Skripsi ini diajukan oleh :
Nama : Triastuti Chandrawati
NPM : 1204000866
Program Studi : S1 Reguler
Judul Skripsi : Pengembangan Part of Speech Tagger untuk
Bahasa Indonesia Berdasarkan Metode Conditional
Random Fields dan Transformation Based
Learning

Telah berhasil dipertahankan di hadapan Dewan Penguji dan diterima sebagai bagian persyaratan yang diperlukan untuk memperoleh gelar Sarjana Ilmu Komputer pada Program Studi S1 Reguler Fakultas Ilmu Komputer, Universitas Indonesia.

DEWAN PENGUJI

Pembimbing : Dra. Mirna Adriani, Ph.D. ()
Penguji : Dr. Indra Budi ()
Penguji : Dr. Petrus Mursanto ()

Ditetapkan di : Depok
Tanggal : 28 Juli 2008

KATA PENGANTAR

Puji syukur kepada Tuhan atas segala kekuatan dan berkat yang daripada-Nya untuk memampukan penulis menyelesaikan pelaksanaan Tugas Akhir dan penulisan laporan ini. Bersamaan dengan itu penulis juga hendak menyampaikan penghargaan yang sebesar-besarnya kepada pihak-pihak yang telah memampukan penulis dalam pelaksanaan dan penyelesaian Tugas Akhir ini.

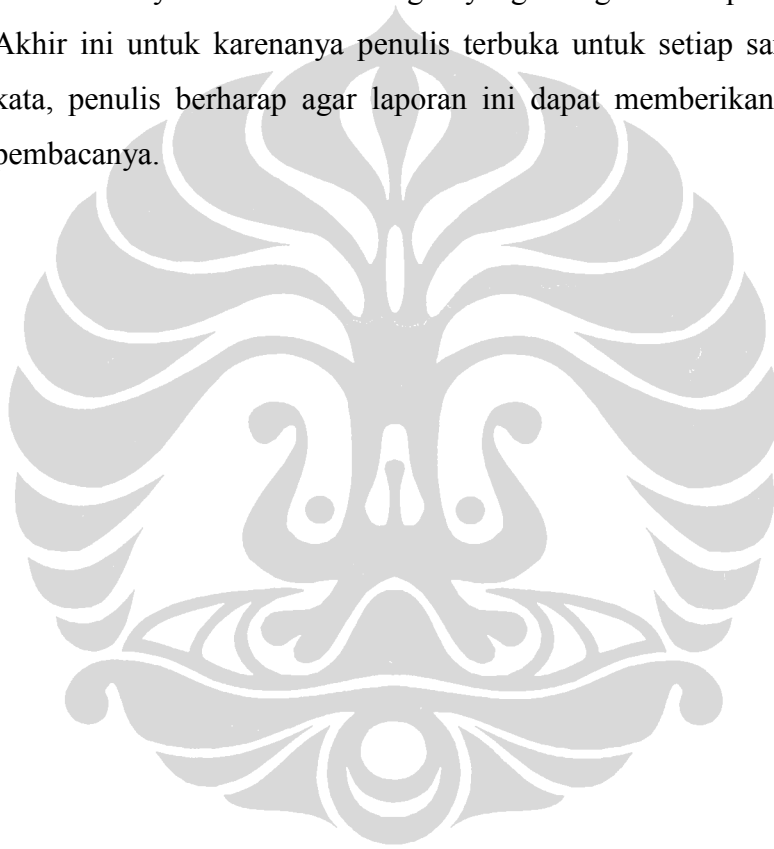
1. Untuk yang terhormat Ibu Dra. Mirna Ariani, PhD. selaku pembimbing Tugas Akhir atas segala bimbingan yang diberikan selama pelaksanaan Tugas Akhir ini, atas segala ilmu dan saran, dan atas kesediaan waktu untuk membantu penulis selama diskusi lab Information Retrieval.
2. Untuk yang terhormat Bapak Dana Indra S., MLIS, PhD. selaku Pembimbing Akademik atas segala bimbingan yang telah diberikan.
3. Untuk yang terhormat Bapak Hisar Maruli M. atas segala ilmu yang diberikan selama kelas Pemrosesan Bahasa Natural dan atas diskusi yang membantu penulis selama pelaksanaan Tugas Akhir.
4. Untuk yang terkasih seluruh anggota keluarga penulis, Bapak, Ibu, dan kakak-kakak, atas segala dukungan dan doa yang tanpa henti, atas kepercayaan dan kekuatan yang tetap diberikan selama pelaksanaan Tugas Akhir ini.
5. Untuk yang terkasih Mbak Syandra Sari dan Mbak Herika Hayurani atas sebagian korpus yang diizinkan untuk dikembangkan dalam Tugas Akhir ini dan atas segala bantuan dalam diskusi selama pelaksanaan Tugas Akhir.
6. Untuk yang terkasih seluruh rekan di lab Information Retrieval—Amalia, Aprilia, Ananda, Arfan, Aurora, Baskoro, Eliza, Femphy, Franky, Ibrahim, Joji, Rama, Rahmad, Wisnu—atas segala diskusi dan dukungan yang diberikan.
7. Untuk yang terkasih Laverdy P, M. Reza B, Candra Adhi W, dan Abe Mitsuteru yang selalu dapat memberikan semangat bagi penulis dan selalu memberikan dukungan moril yang sangat berarti.

8. Untuk yang terkasih Eka Gatari yang bersedia selalu berbagi rasa dan saling memberikan dukungan dalam pelaksanaan Tugas Akhir ini
9. Untuk seluruh teman-teman yang selalu ada untuk penulis dan bersedia terus untuk memberikan dukungan moral maupun teknis.
10. Untuk seluruh keluarga besar Fakultas Ilmu Komputer Universitas Indonesia yang telah memberikan dukungan baik langsung maupun tidak langsung.

Penulis menyadari atas kekurangan yang mungkin terdapat dalam laporan Tugas Akhir ini untuk karenanya penulis terbuka untuk setiap saran dan kritik. Akhir kata, penulis berharap agar laporan ini dapat memberikan manfaat bagi setiap pembacanya.

Depok, Juli 2008

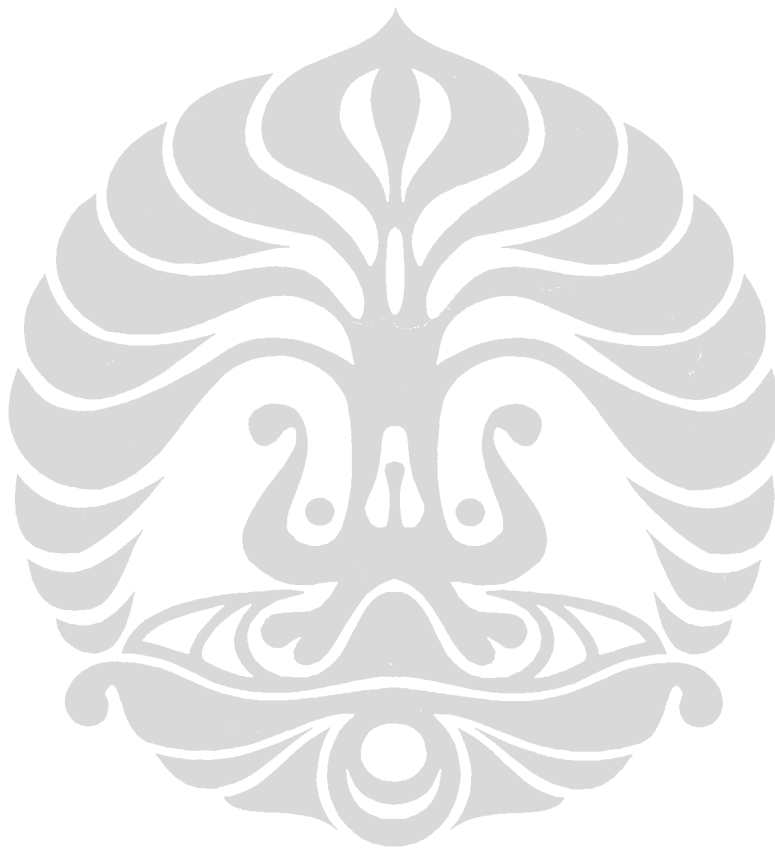
Triastuti Chadrawati



DAFTAR ISI

HALAMAN PERNYATAAN ORISINALITAS	ii
HALAMAN PENGESAHAN	iii
KATA PENGANTAR	iv
ABSTRAK	vi
DAFTAR ISI	viii
DAFTAR TABEL	x
DAFTAR GAMBAR	xi
BAB 1 PENDAHULUAN	1
1.1. Latar Belakang	1
1.2. Permasalahan	5
1.3. Tujuan	6
1.4. Metodologi Penelitian	7
1.5. Sistematika Penulisan	7
BAB 2 LANDASAN TEORI	9
2.1. <i>Part of Speech Tag</i>	9
2.1.1. Terminologi	9
2.1.2. Fungsi dan Tujuan <i>Part of Speech Tag</i>	10
2.1.3. <i>Part of Speech Tagset</i>	11
2.2. Metode-Metode <i>Part of Speech Tagger</i>	14
2.2.1. Conditional Random Fields	15
2.2.2. Transformation Based Learning	21
2.3. Penelitian <i>Part of Speech Tagging</i> yang Pernah Dilakukan	26
BAB 3 PENELITIAN	28
3.1. Aplikasi	28
3.1.1. CRF++-0.50	28
3.1.1.1. Fitur-Fitur dalam CRF++-0.50	28
3.1.1.2. Pelatihan-Pengujian dalam CRF++-0.50	29
3.1.1.3. Penghitungan Akurasi dalam CRF++-0.50	33
3.1.2. RBT V1.14	33
3.1.2.1. Fitur-Fitur dalam RBT V1.14	33
3.1.2.2. Pelatihan-Pengujian dalam RBT V1.14	34
3.1.2.3. Penghitungan Akurasi dalam RBT V1.14	37
3.2. Korpus Bahasa Indonesia	37
3.3. Cetakan Fitur	38
3.3.1. Format Cetakan Fitur	39
3.3.2. Pembuatan Cetakan Fitur	41
3.4. Pelaksanaan Penelitian	46
3.4.1. Proses Pelatihan	47
3.4.2. Proses Pengujian	48
BAB 4 HASIL PENELITIAN DAN ANALISIS	49
4.1. Hasil Penelitian	49
4.2. Analisis Hasil Penelitian	52
4.2.1. Analisis Kesalahan	53
4.2.2. Analisis Perbandingan Metode <i>Part of speech Tagger</i>	59
4.2.3. Analisis Perbandingan Fitur	60

BAB 5 PENUTUP	62
5.1. Kesimpulan.....	62
5.2. Saran	63
DAFTAR PUSTAKA.....	64
LAMPIRAN A <i>TAGSET</i>	68
LAMPIRAN B <i>FEATURE TEMPLATE</i>	69
LAMPIRAN C TABEL KESALAHAN <i>TAGGING</i>	73



DAFTAR TABEL

Table 2.1: Penn Treebank dan Brown Corpus Tagset.....	11
Table 2.2: POS <i>Tagset</i> untuk Bahasa Indonesia.....	12
Table 3.1: Fitur Masing-Masing Dokumen.....	45
Table 4.2: Akurasi Hasil <i>Tagging</i> CRF.....	49
Table 4.3: Akurasi Hasil <i>Tagging</i> RBT.....	51
Table 4.4: Akurasi Hasil <i>Tagging</i> CRF, TBL, dan CRF-TBL.....	51
Table 4.5: Kesalahan <i>Tagging</i>	53
Table 4.6: Rata-Rata Persentase Kesalahan <i>Tagging</i>	56



DAFTAR GAMBAR

Gambar 2.1: CRF [Lafferty et al, 2001]	17
Gambar 2.2: Contoh Data CRF	18
Gambar 2.3: Proses Tagging TBL [Brill, 1994].....	22
Gambar 2.4: Contoh Data TBL	24
Gambar 2.5: Data Pelatihan Leksikal.....	24
Gambar 2.6: Data Pelatihan Kontekstual	24
Gambar 2.7: Leksikon	24
Gambar 2.8: Hasil Tagging dengan Leksikon dan <i>Lexical Rule</i>	25
Gambar 3.1: Proses Pelatihan-Pengujian CRF.....	30
Gambar 3.2: Bentuk Data Pelatihan CRF++-0.50.....	32
Gambar 3.3: Proses Pelatihan-Pengujian RBT V1.14.....	35
Gambar 3.4: Bentuk Data Pelatihan RBT V1.14	37
Gambar 3.5: Contoh Dokumen Pelatihan.....	40
Gambar 3.6: Contoh Fitur	40
Gambar 3.7: Dokumen Cetakan Fitur 1	42
Gambar 3.8: Dokumen Cetakan Fitur 2	43
Gambar 3.9: Dokumen Cetakan Fitur 3	44
Gambar 3.10: Dokumen Cetakan Fitur 4	44
Gambar 3.11 Contoh Data Untuk Pembuatan Fungsi Fitur	45
Gambar 4.1: Sebaran Probabilitas	59