

**Membangun sebuah Combinatory Categorical
Grammar (CCG) Supertagger berbasis
Maximum Entropy untuk Bahasa Indonesia**

Proposal Tugas Akhir

Kelas TA NLP

Wisnu Adi Nurcahyo

NIM: 1301160479



Program Studi Sarjana Informatika

Fakultas Informatika

Universitas Telkom

Bandung

2019

Lembar Persetujuan

**Membangun sebuah Combinatory Categorical Grammar (CCG)
Supertagger berbasis Maximum Entropy untuk Bahasa Indonesia**

***Building a Maximum Entropy based Combinatory Categorical
Grammar (CCG) Supertagger for Bahasa Indonesia***

**Wisnu Adi Nurcahyo
NIM: 1301160479**

Proposal ini diajukan sebagai usulan pembuatan tugas akhir pada
Program Studi Sarjana Informatika
Fakultas Informatika Universitas Telkom

Bandung, 24 Oktober 2019
Menyetujui

Calon Pembimbing 1

Dr. Ade Romadhony, S.T., M.T.
NIP: 06840042

Abstrak

Dalam pemrosesan bahasa alami, combinatory categorial grammar (CCG) merupakan salah satu formalisme tatabahasa yang dapat digunakan untuk membangun sebuah *parser* yang umumnya dikenal sebagai CCG *parser*. CCG *parser* dapat digunakan untuk berbagai macam keperluan dalam pemrosesan bahasa alami. Sebagai contoh, CCG *parser* dapat digunakan untuk memperoleh informasi (*information extraction*) dari suatu kalimat yang kemudian membentuk sebuah *query*. Agar dapat bekerja, CCG *parser* membutuhkan CCG *lexicon*. CCG *lexicon* diperoleh dari proses yang bernama *supertagging*. *Supertagging* adalah proses pelabelan suatu token kata terhadap *supertag*-nya. Perangkat lunak yang melakukan *supertagging* disebut sebagai *supertagger*. Demikian itu, *supertagging* merupakan langkah pertama yang perlu dilakukan sebelum membangun sebuah CCG *parser*. *Supertagger* yang dibangun dalam tugas akhir ini dimaksudkan sebagai produsen CCG *lexicon* bahasa Indonesia untuk riset-riset yang berkenaan dengan CCG di masa yang akan datang.

Kata Kunci: natural language processing, combinatory categorial grammar, supertagger, maximum entropy model, bahasa indonesia, haskell

Daftar Isi

Abstrak	i
Daftar Isi	ii
I Pendahuluan	1
1.1 Latar Belakang	1
1.2 Perumusan Masalah	1
1.3 Tujuan	2
1.4 Batasan Masalah	2
1.5 Rencana Kegiatan	2
1.6 Jadwal Kegiatan	2
Daftar Pustaka	4
Lampiran	5

Bab I

Pendahuluan

1.1 Latar Belakang

Riset pemrosesan bahasa natural untuk bahasa Indonesia saat ini terbilang sedikit. Bahkan, masih banyak area riset yang belum tersentuh seperti contohnya *combinatory categorial grammar* (CCG). CCG merupakan formalisme tatabahasa yang salah satu manfaatnya adalah untuk memperoleh informasi (*information extraction*) dari suatu kalimat. Informasi tersebut diperoleh setelah melakukan *parsing* berdasarkan formalisme CCG dengan menggunakan perangkat lunak bernama CCG *parser*. Untuk dapat melakukan *parsing*, CCG *parser* membutuhkan CCG *lexicon* yang mengandung bentuk formal dari suatu token kata. Bentuk formal yang dimaksud adalah *category* dalam *category theory*. CCG *lexicon* diperoleh dari proses pelabelan suatu token kata terhadap bentuk formalnya yang mana dikenal sebagai *supertagging*. Proses *supertagging* akan menghasilkan *supertag* yang kemudian disebut sebagai CCG *supertag* karena formalisme yang digunakan adalah formalisme CCG. Dalam hal ini, CCG *supertag* adalah CCG *lexicon* itu sendiri.

Tugas akhir dengan judul “Membangun sebuah Combinatory Categorical Grammar (CCG) Supertagger berbasis Maximum Entropy untuk Bahasa Indonesia” berusaha untuk membangun versi awal dari CCG *supertagger* untuk bahasa Indonesia yang mana harapannya dapat menjadi inisiator riset pemrosesan bahasa natural dengan tema CCG sehingga ke depannya akan ada lebih banyak riset mengenai CCG yang tersedia. *Supertagger* yang dimaksud dalam tugas akhir ini akan dibangun dengan menggunakan model Maximum Entropy (MaxEnt) dan implementasinya akan ditulis dalam bahasa pemrograman Haskell. Model MaxEnt digunakan karena keterbatasan *dataset* untuk melakukan *learning*. Adapun bahasa pemrograman Haskell digunakan karena abstraksi bahasanya yang sangat mendekati *category theory* serta kemampuannya yang sangat baik dalam pemrosesan data.

1.2 Perumusan Masalah

Rumusan masalah yang akan diangkat yaitu:

1. Mengapa CCG *supertagger* diperlukan?

2. Apa saja yang harus dipersiapkan untuk membangun CCG *supertagger*?
3. Bagaimana proses pembangunan CCG *supertagger*?

1.3 Tujuan

Tujuan yang diharapkan dapat tercapai oleh tugas akhir ini yaitu:

1. Mengenalkan alternatif metode yang dapat digunakan dalam pemrosesan bahasa alami untuk bahasa Indonesia.
2. Merilis CCG *supertagger* pertama untuk bahasa Indonesia.
3. Membuka peluang riset untuk CCG *parser* bahasa Indonesia.

1.4 Batasan Masalah

Hipotesis dari tugas akhir ini yaitu:

1. Memberikan label CCG untuk proses *learning* merupakan permasalahan utama dari tugas akhir ini.
2. *Supertagger* yang akan dibangun kemungkinan besar memiliki akurasi yang cenderung rendah.
3. CCG *lexicon* sudah dapat digunakan oleh CCG *parser* bahasa Indonesia (apabila ada).

1.5 Rencana Kegiatan

Rencana kegiatan yang akan dilakukan adalah sebagai berikut:

- Studi literatur
- Studi *tools* yang tersedia
- Studi bahasa pemrograman yang akan digunakan
- Perancangan sistem *supertagger*
- Membangun *supertagger*
- Memeriksa hasil

1.6 Jadwal Kegiatan

Laporan proposal ini akan dijadwalkan sesuai dengan tabel 1.1.

Tabel 1.1: Jadwal kegiatan proposal tugas akhir.

No	Kegiatan	Bulan ke-																							
		1				2				3				4				5				6			
1	Studi Literatur																								
2	Studi <i>Tools</i> yang Tersedia																								
3	Studi Bahasa Pemrograman																								
4	Pengumpulan Data																								
5	Analisis dan Perancangan Sistem																								
6	Implementasi Sistem																								
7	Analisa Hasil Implementasi																								
8	Penulisan Laporan																								

Daftar Pustaka

Lampiran