Membangun sebuah Combinatory Categorial 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 Categorial Grammar (CCG) Supertagger Berbasis Maximum Entropy untuk Bahasa Indonesia

Building a Combinatory Categorial Grammar (CCG) Supertagger Based on the Maximum Entropy 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, 13 November 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 tata bahasa 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

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 tata bahasa 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 Categorial 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 *Tools* yang Tersedia																								
3	Studi Bahasa Pemrograman																								
4	Pengumpulan Data																								
5	Analisis dan Perancangan Sistem																								
6	Implementasi Sistem																								
7	Analisa Hasil Implementasi																								
8	Penulisan Laporan																								

Bab II

Kajian Pustaka

2.1 Categorial Grammar

Categorial Grammar (CG) merupakan sebuah istilah yang mencakup beberapa formalisme terkait yang diajukan untuk sintaks dan semantik dari bahasa alami serta untuk bahasa logis dan matematis [2]. Karakteristik yang paling terlihat dari CG adalah bentuk esktrim dari leksikalismenya di mana beban utama (atau bahkan seluruh beban) sintaksisnya ditanggung oleh leksikon. Konstituen tata bahasa dalam *categorial grammar* dan khususnya semua leksikal diasosiasikan dengan suatu *type* atau "*category*" (dalam *category theory*) yang mendefinisikan potensi mereka untuk dikombinasikan dengan konstituen lain untuk menghasilkan konstituen majemuk. *Category* tersebut adalah salah satu dari sejumlah kecil *category* dasar (seperti NP) atau *functor* (dalam *category theory*).

Ada beberapa notasi berbeda untuk *category* dalam merepresentasikan *directional*-nya. Notasi yang paling umum digunakan adalah "*slash notation*" yang dipelopori oleh Bar-Hilel, Lambek, dan kemudian dimodifikasi dalam kelompok teori yang dibedakan sebagai tata bahasa "*combinatory*" *categorial grammar* (CCG). Sebagai contoh, *category* (S\NP)/NP merupakan suatu *functor* yang memiliki dua buah notasi *slash* yaitu \ dan /. Masing-masing notasi *slash* tersebut merepresentasikan *directionality* yang berbeda. Notasi *forward slash*, /, mengindikasikan bahwa argumen dari suatu *functor* X/Y ada di bagian kanan atau dengan kata lain Y. Adapun *backward slash*, \, mengindikasikan bahwa argumen dari suatu *functor* X\Y ada di bagian kiri atau dengan kata lain X. Demikian itu, penggunaan notasi *slash* yang tepat sangat penting dikarenakan hal ini dapat mempengaruhi konstituen dari hasil "kombinasi" *category*-nya.

2.2 Combinatory Categorial Grammar

Combinatory Categorial Grammar (CCG) merupakan salah satu formalisme tata bahasa yang gaya aturannya diturunkan dari *categorial grammar* dengan beberapa penambahan aturan dan istilah baru. Di CCG, *category* dapat dipasangkan dengan *semantic representation*. Dalam hal ini, *semantic repre-*

$$
\begin{aligned}
\text{Pamungkas} &\vdash \text{NP} : \mathit{pamungkas}' \\
\text{Setyo} &\vdash \text{NP} : \mathit{setyo}' \\
\text{dan} &\vdash \text{CONJ} : \lambda x.\lambda y.\lambda f.\ (f\ x) \wedge (f\ y) \\
\text{menyukai} &\vdash (\text{S}\backslash\text{NP})/\text{NP} : \lambda x.\lambda y.\ \mathit{suka}(y, x) \\
\text{rendang} &\vdash \text{NP} : \mathit{rendang}'
\end{aligned}
$$

Gambar 2.1: Kamus yang memetakan token kata ke bentuk CCG *lexicon*-nya.

sentation yang dimaksud adalah abstraksi fungsi lambda (dalam *lambda calculus*, *lambda function*). Sebagai contoh, *category* (S\NP)/NP dapat dipasangkan dengan fungsi lambda $\lambda x.fx$ sehingga dapat ditulis menjadi (S\NP)/NP : $\lambda x.fx$. Adapun pemetaan dari suatu token kata ke *category*-nya menggunakan notasi $\vdash$. Sebagai contoh, anggap saja kita memiliki kamus pemetaan seperti pada Gambar 2.1. Apabila kita memiliki kalimat "Pamungkas dan Setyo menyukai rendang", maka kita dapatkan:

Pamungkas	dan	Setyo	menyukai	rendang
NP	CONJ	NP	(S\NP)/NP	NP
$: \mathit{pamungkas}'$	$: \lambda x.\lambda y.\lambda f.\ (f\ x) \wedge (f\ y)$	$: \mathit{setyo}'$	$: \lambda x.\lambda y.\ \mathit{suka}(y, x)$	$: \mathit{rendang}'$

Ada beberapa operasi yang dapat dilakukan dalam CCG. *Operand* dari operasi yang dimaksud adalah *category*. Berdasarkan contoh di atas, akan ada tiga operasi yang dijalankan yaitu *coordination*, *forward application*, dan *type rising*. Untuk mendapatkan hasil yang diinginkan, kita lakukan *type rising* sebelum *forward application* di akhir. Sehingga, kita dapatkan:

Pamungkas	dan	Setyo	menyukai	rendang
NP	CONJ	NP	(S\NP)/NP	NP
$: \mathit{pamungkas}'$	$: \lambda x.\lambda y.\lambda f.\ (f\ x) \wedge (f\ y)$	$: \mathit{setyo}'$	$: \lambda x.\lambda y.\ \mathit{suka}(y, x)$	$: \mathit{rendang}'$
		$<\&>$		
	NP		S\NP	
	$: \lambda f.\ (f\ \mathit{pamungkas}') \wedge (f\ \mathit{setyo}')$		$: \lambda y.\ \mathit{suka}(y, \mathit{rendang}')$	
	$> T$			
	S/(S\NP)			
	$: \lambda f.\ (f\ \mathit{pamungkas}') \wedge (f\ \mathit{setyo}')$			
	$>$			
	S			
	$: \mathit{suka}(\mathit{pamungkas}', \mathit{rendang}') \wedge \mathit{suka}(\mathit{setyo}', \mathit{rendang}')$			

Berdasarkan hasil evaluasi tersebut, kita dapatkan *query* 2.1 yang diperoleh dari kalimat "Pamungkas dan Setyo menyukai rendang". Demikian itu, komputer dapat melakukan komputasi berdasarkan *query* yang telah diperoleh.

Kegiatan tersebut merupakan apa yang disebut dengan CCG *parsing*. Untuk dapat melakukan parsing, CCG *lexicon* diperlukan. Untuk mendapatkan CCG *lexicon* kita dapat menggunakan CCG *supertagger* yang akan melakukan pelabelan suatu token kata ke CCG *lexicon* berdasarkan pemetaannya.

$$suka(pamungkas', rendang') \wedge suka(setyo', rendang') \tag{2.1}$$

2.3 Category Theory

Category Theory (CT) merupakan formalisme yang dapat digunakan untuk memformalkan struktur matematis. CT mempelajari *category* yang merupakan sebuah representasi dari suatu abstraksi konsep matematis. Suatu *category* memiliki kumpulan *object* dan *morphism*. Untuk mempermudah pemahaman mengenai CT, kita akan gunakan *category of set* (kategori dari himpunan) sebagai contoh. Dalam *category of set*, *object*-nya adalah himpunan dan *morphism*-nya (terkadang disebut dengan *arrow*) adalah fungsi (*function*, sebuah pemetaan). Kemudian, pemetaan dari suatu *category* C ke *category* D yang dipetakan oleh F ($F : C \to D$) disebut sebagai *functor*.

2.4 Lambda Calculus

Lambda calculus (λ-*calculus*) merupakan sebuah formalisme yang dikembangkan oleh Alonzo Church sebagai alat yang digunakan untuk memahami konsep komputasi yang efektif [1]. Formalisme λ-*calculus* cukup populer dan bahkan dijadikan sebagai pondasi teori bagi paradigma pemrograman *functional programming*. Konsep utama dari λ-*calculus* adalah apa yang disebut dengan *expression*. Suatu *expression* dalam λ-*calculus* terdiri dari tiga bagian yaitu *lambda notation* (λ), *argument* (seperti a, b, c, x, dan lain-lain), dan *body* yang dipisahkan dengan tanda titik. Sebagai contoh, fungsi lambda $\lambda x.x$ merupakan sebuah fungsi identitas yang mengambil argumen x kemudian mengembalikan nilai x itu sendiri. Dalam hal ini, terlihat bahwa notasi λ merupakan sebuah penanda bagi suatu fungsi lambda. Kemudian, pengubah x setelah notasi λ merupakan argumen dari fungsi tersebut. Selanjutnya, tanda titik merupakan pemisah antara *head* dan *body* fungsi lambda. Terakhir, setelah tanda titik adalah *body* dari suatu fungsi lambda yang mana berupa *expression*.

Untuk mempermudah pemahaman, λ-*calculus* dapat diperlakukan seperti fungsi tanpa nama. Sebagai contoh, fungsi lambda ($\lambda x.x+5$) apabila diberikan nilai 2 sehingga menjadi ($\lambda x.x + 5)2$ akan dievaluasi menjadi $\lambda(2).(2) + 5$. Demikian itu, nilai yang dikembalikan oleh fungsi tersebut adalah 7. Sama seperti fungsi pada umumnya, konsep ini bernama *substition* (substitusi). Memahami λ-*calculus* dirasa perlu berhubung dalam tugas akhir ini λ-*calculus* digunakan sebagai bentuk formal di *category* dalam konteks CCG *lexicon*. Meskipun λ-*calculus* tidak sesederhana yang dijelaskan sebelumnya, setidaknya memahami

λ-*calculus* seperti ini sudah cukup untuk dapat membangun *supertagger* yang ada di tugas akhir ini.

2.5 Supertagging

Supertagging merupakan proses yang memetakan suatu token kata ke bentuk *supertag*-nya. CCG *supertagging* artinya proses pemetaan dari suatu token kata ke dalam bentuk CCG *supertag* atau dapat juga disebut sebagai CCG *lexicon*. *Supertag* mirip dengan POS *tag*. Perbedaannya, *supertag* memiliki bentuk formalisme yang lebih kompleks dari POS *tag*. Hal ini dikarenakan *supertag* menyimpan informasi lain selain tanda gramatikalnya (NP, NN, VB, dan sebagainya). Sebagai contoh, CCG *supertag* memiliki bentuk dengan format:

$$< \text{categorial grammar tag} >: < \text{semantic representation} >$$

Kita dapat menggunakan notasi $\vdash$ untuk memetakan token kata ke bentuk CCG *supertag*-nya. Anggap kita memiliki kata kerja "menyukai" yang akan dipetakan ke (S\NP)/NP : $\lambda x.\lambda y.\ suka(y, x)$, kita dapatkan:

$$\text{menyukai} \vdash (S\backslash NP)/NP : \lambda x.\lambda y.\ suka(y, x)$$

Adapun sebuah *tool* yang melakukan proses *supertagging* ini dinamakan *supertagger*. *Supertagger* sederhananya mengambil daftar token kata yang kemudian untuk setiap token kata tersebut "dilabelkan" dengan *supertag*-nya.

2.6 Maximum Entropy Model

Maximum Entropy (MaxEnt) merupakan model statistik yang dapat digunakan untuk melakukan *train* korpus teranotasi dengan Part-Of-Speech (POS) *tag*. Salah satu aplikasi MaxEnt adalah POS *tagger* yang mana memiliki hasil akurasi yang lebih baik ketimbang *state-of-the-art*. Adapun model probabilitasnya didefinisikan dalam $\mathcal{H} \times \mathcal{T}$, dimana $\mathcal{H}$ adalah himpunan dari kemungkinan kata dan konteks *tag*, atau "*history*" (riwayat), dan $\mathcal{T}$ adalah himpunan dari *tag* yang diizinkan. Model probabilitas dari suatu *history* h bersama dengan *tag* t didefinisikan dalam persamaan 2.2.

$$p(h, t) = \pi\mu \prod_{j=1}^{k} a_j^{f_j(h,t)} \tag{2.2}$$

dimana π merupakan konstan normalisasi, $\{\mu, a_1, \ldots, a_k\}$ merupakan parameter model positif, dan $\{f_1, \ldots, f_k\}$ merupakan apa yang kita sebut sebagai "*feature*", dimana $f_j(h, t) \in \{0, 1\}$.

Meskipun MaxEnt dipernalkan untuk POS *tagging*, MaxEnt dapat pula digunakan untuk *supertagging*. Stephen Clark 2002 mempublikasikan literatur

supertagging untuk CCG yang mana MaxEnt merupakan model yang digunakan. Adapun persamaan modelnya dapat dilihat di persamaan 2.3.

$$p(c|h) = \frac{1}{Z(h)} e^{\Sigma_i \lambda_i f_i(c,h)} \tag{2.3}$$

dimana c merupakan *category*, h merupakan *context*, fungsi $f_i(c, h)$ merupakan "*feature*" dari suatu *category* dan *context*, dan $Z(h)$ merupakan konstan normalisasinya. Adapun contoh "*feature*" yang dimaksud dapat dilihat di persamaan 2.4.

$$f_j(c, h) = \begin{cases} 1 & , \text{if } merupakan\_kata\_yang(h) = true \ \& \ c = \text{NP/N} \\ 0 & , \text{selainnya} \end{cases} \tag{2.4}$$

Bab III

Metodologi dan Desain Sistem

3.1 Pembuatan *Dataset*

Dataset untuk melakukan *train* CCG sayangnya belum tersedia. Karenanya, kita akan membuatkan *dataset* secara semi otomatis dengan memanfaatkan POS *tagger* untuk bahasa Indonesia. Secara formal, *flowchart* untuk proses pembuatan *dataset* dapat dilihat pada Gambar 3.1.

3.1.1 *Input* Teks Bahasa Indonesia

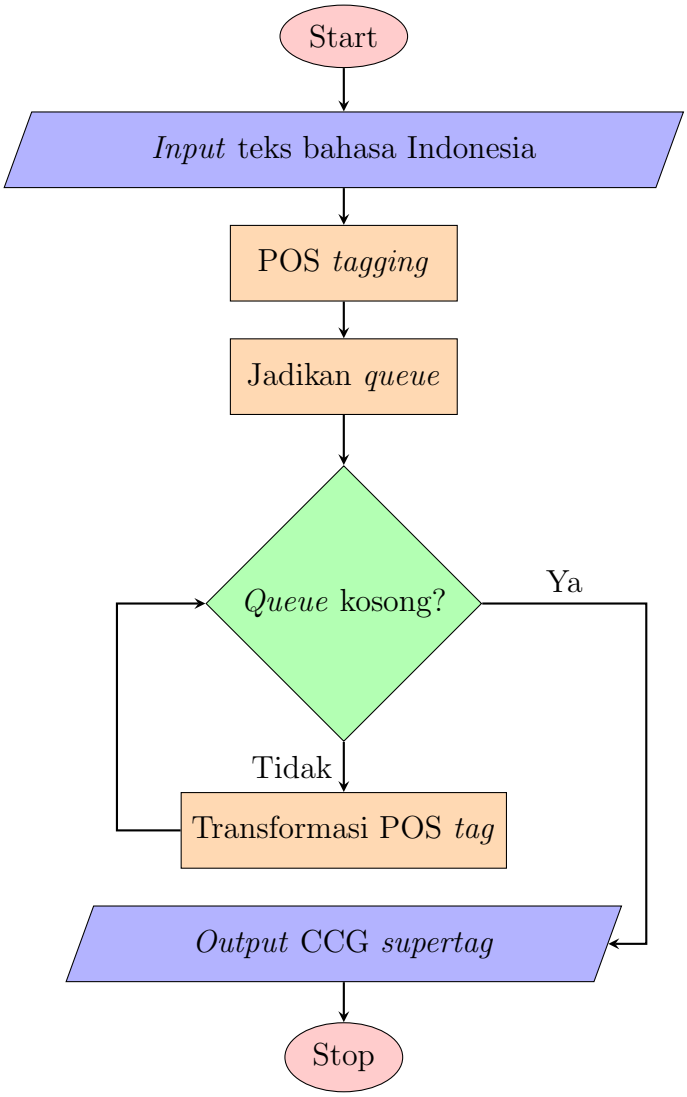
Pada bagian *input* dalam pembuatan *dataset*, kita dapat mengambil teks bahasa Indonesia dari Indonesian Treebank[1] dan/atau dari beberapa contoh artikel yang terdapat di *website* Wikipedia Indonesia[2]. Indonesian Treebank terdapat setidaknya 1000 kalimat yang dirasa cukup untuk melakukan *training* menggunakan model MaxEnt. Apa yang perlu dilakukan setelah mengambil *input* adalah membersihkan bentuk *tree*-nya untuk kemudian diambilkan kalimat yang sebenarnya. Caranya cukup sederhana yaitu dengan mengambil *leaf* dari *tree* tersebut kemudian disatukan di suatu pengubah dengan tipe data *string*.

Sebagai contoh, salah satu *tree* yang terdapat di Indonesian Treebank dapat dilihat pada Gambar 3.2. Apa yang dimaksud dengan *leaf* pada *tree* tersebut adalah $(Kera)$, $(untuk)$, $(*)$, $(amankan)$, $(pesta\ olahraga)$. Terkhusus untuk *leaf* dengan bentuk spesial, seperti $(*)$, kita hilangkan sehingga teks yang diperoleh dari *tree* tersebut adalah "Kera untuk amankan pesta olahraga". Selanjutnya, contoh teks yang akan digunakan agar konsisten yaitu "Pamungkas dan Setyo menyukai rendang".

3.1.2 POS *Tagging*

Berdasarkan teks bahasa Indonesia yang telah diambil, kita manfaatkan *tool* POS *tagger* bahasa Indonesia untuk mendapatkan *lexical category* atomik untuk masing-masing token. Dengan memanfaatkan POS *tag* kita dapat membuat *dataset* untuk CCG *supertag* lebih mudah dibandingkan dengan

[1] github.com/famrashel/idn-treebank/blob/master/Indonesian_Treebank.bracket
[2] id.wikipedia.org



Gambar 3.1: Alur kerja pembuatan *dataset* untuk CCG *supertag*.

memberikan *tag* CCG secara manual. Ide dasarnya yaitu kita akan mentransformasikan POS *tag* yang didapatkan menjadi CCG *supertag* berdasarkan aturan-aturan khusus yang telah ditentukan. Sebagai contoh, kita dapat membuat aturan seperti mentransformasikan VB menjadi (S\NP)/NP. Dengan menggunakan contoh kalimat yang sama seperti di bagian sebelumnya, yaitu "Pamungkas dan Setyo menyukai rendang", setelah menggunakan POS *tagger* bahasa Indonesia kita dapatkan hasil sesuai dengan Gambar 3.3.

3.1.3 Transformasi POS *Tag*

Pada bagian proses transformasi POS *tag* ke bentuk CCG *supertag*-nya, kita buatkan aturan-aturan transformasinya. Aturan transformasi tersebut

$$
\begin{array}{l}
(NP \\
\quad (NN\ (Kera)) \\
\quad (SBAR \\
\quad\quad (SC\ (untuk)) \\
\quad\quad (S\ (NP-SBJ\ (*)) \\
\quad\quad\quad (VP \\
\quad\quad\quad\quad (VB\ (amankan)) \\
\quad\quad\quad\quad (NP\ (NN\ (pesta\ olahraga)))))))
\end{array}
$$

Gambar 3.2: Salah satu contoh *tree* dalam Indonesian Treebank.

Pamungkas	dan	Setyo	menyukai	rendang
NNP	CC	NNP	VB	X

Gambar 3.3: Kalimat contoh dengan POS *tag*-nya.

merupakan pemetaan berbasis aturan. Sebagai contoh, kata "menyukai" memiliki POS *tag* VB. Anggap saja dalam aturan transformasi terdapat pemetaan VB $\vdash$ (S\NP)/NP. Sehingga, kita dapatkan CCG *supertag* untuk "menyukai" yaitu (S\NP)/NP. Kendati demikian, masih ada bagian yang belum kita dapatkan yaitu *semantic representation*-nya. Kita dapat menggunakan *stemmer* bahasa Indonesia agar mendapatkan *root words* dari kata "menyukai" yaitu "suka". Langkah terakhirnya adalah membuatkan *semantic representation*-nya berdasarkan *root words* yang telah diperoleh sehingga kita dapatkan fungsi lambdanya yaitu $\lambda x.\lambda y.$ suka(y, x).

Selain menggunakan *stemmer*, kita dapat menggunakan *morphological analyzer*. Untuk bahasa Indonesia, kita dapat menggunakan *tool* bernama MorphInd[3]. *Morphological analyzer* salah satu kegunaannya yaitu dapat menghasilkan *root words* sehingga dapat kita manfaatkan untuk membuat fungsi lambda. Namun, kita dapat menggunakan MorphInd sebagai pelengkap POS *tagger* untuk bahasa Indonesia. Hal ini agar *dataset* yang dibuatkan secara semi-otomatis ini dapat memiliki kualitas yang baik.

[3]septinalarasati.com/morphind/

Daftar Pustaka

[1] Raul Rojas. A tutorial introduction to the lambda calculus. *CoRR*, abs/1503.09060, 2015.

[2] Mark Steedman. Categorial grammar. Technical report, 1992.

Lampiran