CCGtown: Alat Anotasi Combinatory Categorial Grammar (CCG) Semi-otomatis Berbasis Web

Tugas Akhir
diajukan untuk memenuhi salah satu syarat
memperoleh gelar sarjana
dari Program Studi Informatika
Fakultas Informatika
Universitas Telkom

1301160479
Wisnu Adi Nurcahyo

Program Studi Sarjana Informatika
Fakultas Informatika
Universitas Telkom
Bandung

2021

LEMBAR PENGESAHAN

CCGtown: Alat Anotasi Combinatory Categorial Grammar (CCG) Semi-otomatis Berbasis Web

CCGtown: A Web-based Semi-automatic Combinatory Categorial Grammar (CCG) Annotation Tool

NIM: 1301160479

Wisnu Adi Nurcahyo

Tugas akhir ini telah diterima dan disahkan untuk memenuhi sebagian syarat memperoleh
gelar pada Program Studi Sarjana Informatika
Fakultas Informatika
Universitas Telkom

Bandung, 20 Januari 2021
Menyetujui

Pembimbing I

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

Ketua Program Studi
Sarjana Informatika,

Niken Dwi Wahyu Cahyani, S.T., M.Kom. PhD

NIP: 00750052

LEMBAR PERNYATAAN

Dengan ini saya, Wisnu Adi Nurcahyo, menyatakan sesungguhnya bahwa Tugas Akhir saya dengan judul "**CCGtown: Alat Anotasi Combinatory Categorial Grammar (CCG) Semi-otomatis Berbasis Web**" beserta dengan seluruh isinya adalah merupakan hasil karya sendiri, dan saya tidak melakukan penjiplakan yang tidak sesuai dengan etika keilmuan yang belaku dalam masyarakat keilmuan. Saya siap menanggung resiko/sanksi yang diberikan jika dikemudian hari ditemukan pelanggaran terhadap etika keilmuan dalam buku TA atau jika ada klaim dari pihak lain terhadap keaslian karya.

Bandung, 20 Januari 2021

Yang Menyatakan,

Wisnu Adi Nurcahyo

CCGtown: Alat Anotasi Combinatory Categorial Grammar (CCG) Semi-otomatis Berbasis Web

Wisnu Adi Nurcahyo[1], Ade Romadhony[2]

[1,2]Fakultas Informatika, Universitas Telkom, Bandung
[1]nurcahyo@student.telkomuniversity.ac.id, [2]aderomadhony@telkomuniversity.ac.id

Abstrak

Dokumen ini merupakan panduan penulisan jurnal Tugas Akhir (TA) di lingkungan Fakultas Informatika Universitas Telkom. Meskipun demikian, dimungkinan/dipersilahkan untuk pembimbing TA menggunakan struktur penulisan yang tidak sama persis dengan yang ada di dokumen ini. Panjang abstrak tidak lebih dari 200 kata dan diketik dalam ukuran huruf 10 pts. TA sebagai salah satu sarana latihan penulisan akademik dan memperjelas tulisan, abstrak dibagi menjadi empat paragraf atau sub-bagian. Setiap sub bagian bisa diberi judul yang digaris bawahi. Abstrak berisi apa, mengapa, bagaimana, dan hasil utama (kesimpulan).

***Apa permasalahan pada topik*. Yang juga menjelaskan latar belakang permasalahan topik. Sebaiknya tuliskan juga apa masukan dan keluaran secara sangat singkat.**

***Mengapa topik menarik atau penting*. Sebisa mungkin tuliskan contohnya secara sangat singkat. Pada bagian ini sebaiknya ditulis juga *apa masalah/kekurangan yang terjadi unt kondisi saat ini* (gap antara kondisi sekarang dengan yang diharapkan)?**

***Bagaimana solusinya*. Jelaskan secara garis besar sistem solusi yang telah dilakukan. Biasanya penjelasan solusi ini merupakan yang terpanjang pada abstrak.**

***Hasil utama*. Hasil utama dari eksperimen ditulis singkat dua-tiga kalimat. Akan lebih baik (optional), kalau dituliskan secara eksplisit kontribusi yang telah dihasilkan. Kontribusi bisa dituliskan diantara bagian solusi dan hasil eksperimen.**

Pastikan abstrak pada jurnal TA tidak copas dari abstrak proposal TA. Pada abstrak proposal kadang ada kata *akan*, seperti misalnya *yang akan dilakukan*; sedangkan pada abstrak Jurnal TA tidak ada kata *akan* spt itu. Tidak boleh ada sitasi pada abstrak. Pada abstrak tidak menggunakan penamaan, simbol atau istilah yang teknis, misalnya *minsup* untuk menyatakan nilai support minimal.

Kata kunci : merupakan kata-kata kunci yang menjelaskan isi tulisan, biasanya bisa diambil dari judul dan abstrak. Maksimal enam buah dan ditulis dengan huruf kecil, kecuali singkatan

Abstract

The abstract should state briefly the general aspects of the subject and the main concolusions. The length of abstract should bo no more than 200 word and should be typed be with 10 pts.

Keywords: keyword should be chosen that they best describe the contents of the paper and should be typed in lower-case, except abbreviation. Keyword should be no more than 6 word

1. Pendahuluan

Latar Belakang

Riset pemrosesan bahasa alami untuk bahasa Indonesia saat ini masih terbilang sedikit. Bahkan, masih banyak area riset yang belum tersentuh seperti contohnya *combinatory categorial grammar* (CCG). Sementara itu, riset mengenai CCG untuk bahasa Inggris sudah cukup matang. Adapun untuk bahasa lainnya (seperti bahasa Vietnam) sudah mulai menggunakan CCG di dalam penelitiannya [3]. Agar dapat menerapkan CCG di dalam aplikasi yang dibangun, *tools* seperti CCG *parser* dan CCG *supertagger* harus tersedia terlebih dahulu. Masing-masing dari *tools* tersebut memerlukan *dataset* agar dapat memberikan hasil yang akurat.

Umumnya terdapat dua cara yang paling sering digunakan untuk mengembangkan CCG *supertagger* maupun CCG *parser* bahasa lokal yaitu (1) membangun *dataset* CCG *supertag* secara manual maupun semi-otomatis atau (2) melakukan transfer *dataset* dari CCGbank (atau dari sumber lainnya) ke dalam bahasa lokal dengan cara melakukan alih bahasa dan bila perlu melakukan penyesuaian untuk *supertag*-nya [2]. Proses pembangunan *dataset* umumnya menggunakan bantuan *annotation tool* agar proses anotasinya menjadi lebih mudah. Salah satu *annotation tool* yang dapat digunakan adalah CCGweb [1].

Tugas akhir ini berusaha untuk membangun alat anotasi CCG baru dengan UI/UX yang lebih baik dari CCGweb. Selain itu, dengan bantuan NLTK alat anotasi ini dapat melakukan *generate* untuk CCG *derivation*-nya kemudian pengguna juga dapat mengubah *derivation*-nya apabila diperlukan. Tujuan dari dibangunnya alat anotasi CCG ini adalah untuk mempermudah proses anotasi yang repetitif. Selanjutnya, *dataset* CCG pertama untuk bahasa Indonesia diharapkan dapat dipublikasikan.

Topik dan Batasannya

Sub-bagian ini bisa juga dinamakan Perumusan Masalah atau Identifikasi Masalah. Untuk nama dalam Bahasa Inggris nama yang populer adalah *Problem Statement* atau *Problem Identification.*

Sub-bagian ini mempunyai fungsi sebagai penjelasan tentang topik TA yaitu apa isu/permasalahan yang akan dikerjakan. Untuk lebih memperjelas bisa juga disampaikan definisi atau pengertian. Penyampaian definisi dan penjelasan pada sub-bagian ini sebaiknya dilakukan dalam tulisan naratif dan informal (tanpa formula matematis) apa topik permasalahan yang telah dikerjakan untuk TA. Untuk mempermudah dalam menuliskan sub-bagian ini, dapat dipandang membuat penjelasan kata-kata kunci (pada abstrak) dan judul TA. Dengan penjelasan di sub-bagian ini, maka topiknya menjadi jelas bagi pembaca. Kalau digambarkan dalam sebuah algoritma, maka salah satu materi utama pada sub-bagian ini menjelaskan apa input dan output dari algoritma tersebut. Oleh karena itu, sangat dianjurkan untuk menerangkan apa input dan output, serta sebuah contoh kasusnya secara sangat singkat.

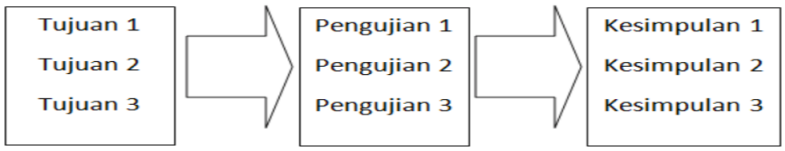
Sebutkan batasan pekerjaan yang ada. Batasan adalah kondisi-kondisi penyederhaan permasalahan, sehingga membuat pekerjaan semakin jauh dari ideal. Batasan masalah berisi pembatasan-pembatasan permasalahan agar menjadi lebih sederhana sehingga bisa/layak dikerjakan sebagai TA yang empat SKS dalam satu semester. Batasan diperlukan karena keterbatasan sumber daya saat pengerjaa TA, misalnya keterbatasan waktu pengerjaan yang hanya satu semester, keterbatasan data pendukung (misalnya tidak tersedianya korpus pengetahuan yang diperlukan) dan keterbatas kemampuan (misalnya untuk implementasi algoritma yang kompleks, dalam implementasinya diimplementasikan bentuk penyederhanaan). Salah satu ciri batasan yang bisa dipakai adalah bila bisa digunakan pada sub-bagian Saran (pada bagian Kesimpulan) agar TA berikutnya melonggarkan atau meniadakan batasan tersebut. Penyederhanaan yang dituliskan untuk batasan, antara lain meliputi data yang ditangani/digunakan, misalnya jumlah data yang digunakan relatif sedikit, dan proses yang dikerjakan, misalnya ada satu subproses yang dikerjakan secara manual. Sebaiknya setiap batasan diberi alasan, misalnya jumlah data yang digunakan hanya 500 buah (relatif sedikit dibandingkan banyak penilitian unt topik sejenis) karena keterbatasan kemampuan komputer yang tersedia. Contoh lain, misalnya proses pelabelan peran semantik pada kalimat Bahasa Indonesia dilakukan secara manual, karena saat ini belum ditemukan alat bantu otomatis unt pelabelan peran semantik untuk Bahasa Indonesia yang efektif. Contoh batasan masalah yang tidak perlu misalnya sudah jelas tercerminkan pada judul.

Tujuan

Sub-bagian Tujuan ini menerangkan kondisi apa yang hendak dicapai atau pertanyaan yang hendak dicari jawabannya. Sebisa mungkin tuliskan kondisi yang hendak dicapai yang terukur (bisa diukur dengan metrik evaluasi yang ditetapkan). Penulisan diupayakan dalam bentuk narasi (bukan berupa poin-poin).

Tujuan-tujuan yang ditetapkan menjadi bahan untuk menentukan skenario eksperimen yang dilakukan. atau dengan kata lain eksperimen dilakukan sesuai dengan tujuannya. Kemudian, kesimpulan pada jurnal TA harus selaras dengan tujuan. Hal ini bisa diilustrasikan pada Gambar 1 atau Tabel 1.

Organisasi Tulisan



Gambar 1. **Keterkaitan antara tujuan, pengujian dan kesimpulan**

Tabel 1. **Keterkaitan antara tujuan, pengujian dan kesimpulan**

No	Tujuan	Pengujian	Kesimpulan
1	Tujuan 1	Pengujian 1	Kesimpulan 1
2	Tujuan 2	Pengujian 2	Kesimpulan 2
3	Tujuan 3	Pengujian 3	Kesimpulan 3

Pada sub-bagian ini dituliskan bagian-bagian selanjutnya (setelah Pendahuluan) pada jurnal TA ini, disertai penjelasan sangat singkat.

2. Studi Terkait

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 [6]. Karakteristik yang paling terlihat dari CG adalah bentuk ekstrim 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*). Dalam hal ini, *category* dapat diartikan sebagai *syntactic type* dari suatu kata.

Secara formal, *syntactic type* didefinisikan sebagai himpunan bagian dari suatu *semigroup* M yang tunduk pada tiga operasi yaitu 1, 2, dan 3 dimana A, B, dan C merupakan himpunan bagian dari M [**?**]. Adapun $A \cdot B$ dibaca A *times* B, C/B dibaca C *over* B, dan $A\backslash C$ dibaca A *under* C. Selanjutnya, dapat dilihat bahwasannya untuk semua $A, B, C \subseteq M$ sehingga kita dapatkan 4 dan 5. Terakhir, persamaan 6 dapat diabaikan apabila dihadapkan dengan *multiplicative system* yang tidak asosiatif. Sementara itu, apabila *semigroup*-nya merupakan sebuah *monoid* dengan identitas 1 maka kita dapatkan 7 dimana $I = \{1\}$.

$$A \cdot B = \{x \cdot y \in M \mid x \in A \wedge y \in B\} \tag{1}$$
$$C/B = \{x \in M \mid \forall_{y \in B} x \cdot y \in C\} \tag{2}$$
$$A\backslash C = \{y \in M \mid \forall_{x \in A} x \cdot y \in C\} \tag{3}$$

$$A \cdot B \subseteq C \quad \text{jika dan hanya jika} \quad A \subseteq C/B \tag{4}$$
$$A \cdot B \subseteq C \quad \text{jika dan hanya jika} \quad B \subseteq A\backslash C \tag{5}$$

$$(A \cdot B) \cdot C = A \cdot (B \cdot C) \tag{6}$$
$$I \cdot A = A = A \cdot I \tag{7}$$

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

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

Gambar 2. Kamus yang memetakan token kata ke bentuk CCG *lexicon*-nya.

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.

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 [7]. Di CCG, *category* dapat dipasangkan dengan *semantic representation*. Dalam hal ini, *semantic representation* 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. Apabila kita memiliki kalimat "Pamungkas dan Setyo menyukai rendang", maka kita dapatkan:

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

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:

Berdasarkan hasil evaluasi tersebut, kita dapatkan *query* 8 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.

$$
\mathit{suka}(\mathit{pamungkas}', \mathit{rendang}') \wedge \mathit{suka}(\mathit{setyo}', \mathit{rendang}') \tag{8}
$$

3. Sistem yang Dibangun

Setelah bagian Pendahuluan dan bagian Studi Terkait, dijelaskan rancangan dan sistem atau produk yang dihasilkan. Penjelasan rancangan dan sistem/produk dituliskan dalam satu atau lebih bagian. Judul untuk bagian-bagian ini bisa menyesuaikan dengan topik TA. Bagian-bagian di sini tidak memuat teori secara umum, namun berisi rancangan dan sistem yang benar-benar telah dibuat atau dipakai.

Sebaiknya judul tidak generik, seperti misalnya Sistem yang Dibangun; namun spesifik sesuai dengan topiknya. Contohnya untuk topik seputar deteksi plagiat, judul bagian-bagian ini misalnya bagian Praproses dan bagian Seeding, Extension dan Filtering.

Uraikan data yang digunakan, sebaiknya disertai sampel data. Jelaskan juga metrik evaluasi yang dipakai serta alasan mengapa menggunakan/memilih metrik tersebut.

Bila diperlukan, informasi lebih detil tentang sistem atau produk yang dibangun bisa disampaikan pada lampiran.

4. Evaluasi

Bagian ini berisi dua sub-bagian, yaitu Hasil Pengujian dan Analisis Hasil Pengujian. Pengujian dan analisis yang dilakukan selaras dengan tujuan TA sebagaimana dinyatakan dalam Pendahuluan.

4.1 Hasil Pengujian

Pertama, tampilkan hasil pengujian yang paling utama. Kemudian hasil-hasil yang lebih detil ditampilkan setelah hasil yang utama. Mengingat tinggi atau rendah, baik atau jeleknya hasil pengujian bersifat relatif, maka sangat dianjurkan ada pembanding (baseline) yang membandingkan dengan algoritma atau pendekatan yang dipilih untuk TA. Pembanding dijalankan pada lingkungan (termasuk data set) yang sama.

Pilih tabel atau jenis diagram yang sesuai untuk menampilkan hasil pengujian.

4.2 Analisis Hasil Pengujian

Analisis merupakan salah satu bagian yang penting untuk TA. Pada TA S1 tidak dituntut untuk mendapatkan hasil performasi yang lebih bagus dibandingkan dengan baseline yang populer, yang dituntut adalah membuat analisis yang lengkap. Menganalisis pengaruh kondisi-kondisi yang berbeda (seperti parameter, jenis data, threshold, dan sub-sistem) yang digunakan.

Cara sitasi adalah sebagai berikut: [8] untuk buku, [4] untuk *paper*, dan [5] untuk website.

5. Kesimpulan

Bagian Kesimpulan memuat kesimpulan dan Saran (*Future Work*), bisa dituliskan dalam poin-poin ataupun paragraf-paragraf. Semua poin kesimpulan diambil dari hasil pengujian dan analisis hasil pengujian sehingga tidak ada kesimpulan dari teori ataupun nalar semata. Sebagaimana sudah disebutkan pada bagian sebelumnya, pengujian dan analisis harus sesuai dengan tujuan TA. Jadi kesimpulan-kesimpulan yang dituliskan selaras dengan seluruh tujuan TA.

Daftar Pustaka

[1] K. Evang, L. Abzianidze, and J. Bos. CCGweb: a new annotation tool and a first quadrilingual CCG treebank. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 37–42, Florence, Italy, Aug. 2019. Association for Computational Linguistics.

[2] J. Hockenmaier and M. Steedman. CCGbank: A corpus of CCG derivations and dependency structures extracted from the Penn treebank. *Computational Linguistics*, 33(3):355–396, 2007.

[3] K. V. Nguyen and N. L.-T. Nguyen. Vietnamese transition-based dependency parsing with supertag features, 2019.

[4] H. Ochoa, K. Rao, and C. Juárez. A hybrid dwt-svd image-coding system (hdwtsvd) for color images. *Systemics. Cybernetics and Informatics*, 1:2–64, 2003.

[5] B. Rahadjo. Pola akses internet yang bursty. `http://rahard.wordpress.com/2011/04/04/pola-akses-internet-yang-bursty/`, 2008. Online; Accessed 3 March 2011.

[6] M. Steedman. Categorial grammar. Technical report, 1992.

[7] M. Steedman. A very short introduction to ccg. Technical report, 1996.

[8] J. Van de Vegte and Y. Xiaoli. *Fundamentals of digital signal processing*. Prentice Hall, 2002.

Lampiran

Lampiran dapat berupa detil data dan contoh lebih lengkapnya, data-data pendukung, detail hasil pengujian, analisis hasil pengujian, detail hasil survey, surat pernyataan dari tempat studi kasus, screenshot tampilan sistem, hasil kuesioner dan lain-lain.