REPOSITOR, Vol. 2, No. 7, Juli 2020, Pp. 897-904
ISSN : 2714-7975
E-ISSN : 2716-1382

Parsing Twitter Menggunakan Metode Left-Corner Parsing Dengan Memanfaatkan POS Tagger

Dyah Anitia[*1], Yuda Munarko[2], Yufis Azhar[3]
[1,2,3]Teknik Informatika/Universitas Muhammadiyah Malang
dyahanitia@gmail.com[*1], yuda@umm.ac.id[2], yufis@umm.ac.id[3]

Abstrak

Pada penelitian ini dilakukan investigasi parser dengan pendekatan left-corner untuk data tweet bahasa Indonesia. Total koleksi tweet sebanyak 850 tweet yang dibagi menjadi tiga kumpulan data, yakni data train POS Tagger, data train dan data uji. Left-corner menggabungkan dua metode yakni top-down dan bottom-up. Dimana top-down digunakan pada proses pengenalan kelas kata dan bottom-up digunakan pada proses pengenalan struktur kalimat. Adapun jenis tag yang digunakan dalam proses top-down berjumlah 23 tagset dan frasa yang digunakan untuk menentukan struktur kalimat frasa yakni frasa nomina, frasa verbal, frasa adjektiva, frasa adverbia dan frasa preposisional. Hasilnya adalah untuk pendekatan left corner mencapai nilai precision 88,29%, nilai recall 68,3% dan F1 measure 77,02%. Nilai yang diperoleh dengan pendekatan left-corner lebih besar dibandingkan nilai dengan pendekatan bottom-up. Hasil dari nilai yang diperoleh dengan bottom up mencapai nilai precision 68,79%, nilai recall 47,12% dan F1 measure 55,9%. Hal ini disebabkan penggunaan kelas kata pada proses top-down berpengaruh pada sturuktur kalimat pada proses bottom up.

Kata Kunci: *Stanford Parser, Left Corner Parsing,Tagging, Parsing, Parse Tree*

Abstract

In this research, we investigated parser with left-corner parser approach for data tweet in Indonesian language. The data used was consisted of 850 tweets which divided for into three data set, that is data train for POS Tagger, data train for parser and data test. The left-corner combines two methods, top-down and bottom-up methods. Top-down used for processes a sequence of words, and attaches a part of speech tag to each and bottom-up used for processes a sentence structure. We used 41 tags and the pharse used to define the sentence structure is noun phrase, verbal phrase, adjective pharse, adverd phrase and prepositional pharse. The result was that precision 88,29%, recall 68,3% and F1 measure 77,02% of left-corner approach. The value obtained by the left-corner approach is greater than the value with the bottom-up approach. The result was that precision 68,29%, recall 47,12% and F1 measure 55,9% of bottom-up approach. This is because the use of word class in top-down process affect the sentence structure in the bottom up process. that is because the use of word class in top-down process affect the sentence structure in the bottom up process.

Keywords: *Stanford Parser, Left Corner Parsing,Tagging, Parsing, Parse Tree*

1. Pendahuluan

Twitter adalah salah satu jejaring sosial yang hingga saat ini masih aktif dengan jumlah pengguna yang meningkat setiap tahunnya. Jejaring sosial yang membatasi penggunanya untuk mengirim 140 karakater dalam setiap ini mengalami pertumbuhan tercepat sejak 2006. Indonesia sendiri menempati posisi ketiga terbanyak di didunia dalam menulis tweet(kicauan) [1]. Salah satu keunggulan penggunaan twitter adalah mampu memberikan informasi secara real-time. Hanya saja informasi yang diberikan terkadang bersifat ambigu, sehingga mengakibatkan penerima salah dalam menafsirkan informasi yang diberikan. Oleh karena itu, diperlukan sebuah sistem untuk memeriksa kebenaran struktur kalimat berdasarkan grammar dan lexicon sehingga penerima informasi dapat memahami makna informasi yang sebenarnya.

Bahasa natural adalah bahasa yang diucapkan, ditulis atau diisyaratkan sebagai bahasa komunikasi untuk komputer dan manusia. Komunikasi ini dalam ilmu komputer dikenal dengan Natural Language Processing (NLP). Natural Language (NLP) adalah pembuatan program untuk memahasi bahasa manusia [2]. Salah satu pengolahan bahasa natural dalam NLP adalah

parsing. Terdapat 2 (dua) kategori pada metode parsing yaitu top-down dan bottom-up. Top-down parsing bekerja dengan menguraikan sebuah kalimat dengan contituent terbesar sampai menghasilkan contituent terkecil dalam kalimat yaitu kata. Bottom-up parsing bekerja sebaliknya, yaitu dengan mengambil satu demi satu kata lalu merangkainya menjadi contituent terbesar yaitu kalimat.

Left-corner parsing adalah sebuah metode gabungan dimana prosesnya dimulai secara bottom-up dan diakhiri dengan top-down. Adanya penggabungan kedua metode ini dikarenakan terdapat kekurangan pada kedua metode sebelumnya. Misalnya pada metode top-down parsing, metode ini tidak dapat menganggani grammar dengan left recursion. Sedangkan bottom-up parsing, tidak dapat menanggani empty production [3].

Penelitian dalam bidang ini sudah banyak dilakukan, teruatama di luar negeri. Sementara di Indonesia sendiri masih jarang dilakukan. Hal inilah yang menjadi kontribusi utama pada penelitian ini.

2. Metode penelitian

2.1 Penelitian Terkait

Sebagai referensi serta pembanding penelitian ini, peneliti mengumpulkan beberapa penelitian sejenis yang berhubungan dengan POS Tagger dan Parsing. Berikut adalah penelitian sebelumnya:

1. Susi Setyowati (2015). Penelitian ini berjudul POS-Tagger twitter bahasa Indonesia Menggunakan Stanford NLP. Penelitian ini tentang bagaimana membuat data training dari tweet bahasa Indonesia dan bagaimana melakukan POS-Tagger twitter berbahasa Indonesia menggunakan Stanford NLP. Hasil dari penelitian ini adalah POS Tagger Stanford NLP dapat digunakan untuk melatih dan mengenali tweet berbahasa Indonesia tetapi tidak dapat memberikan jenis tag berbeda pada suatu kata yang sama tetapi memiliki arti yang berbeda.
2. Fachry Khusaini dan Fachrul Kurniawan (2013). Penelitian ini berjudul Implementasi Left-corner parsing Untuk Pembelajaran Grammar Bahasa Inggris Pada Game 3d Adventure “Go To London”. Penelitian ini tentang bagaimana menerapkan algoritma left-corner parsing sebagai pemeriksa grammar pada sebuah permainan. Hasil dari penelitian ini adalah left-corner parsing mampu mengenali pola grammar pada permainan “Go to London” dengan baik. Hanya saja untuk hal memilih kata yang sesuai dengan definisinya berdasarkan sebuah kalimat belum bisa, sehingga dalam grammar dapat dikenali dengan benar akan tetapi secara pelafalan dirasa kurang tepat.
3. Sri Susanti (2016). Penelitian ini berjudul Analisis Perbandingan Algoritma LCP (Left-corner Parsing) Dan Algoritma CYK (Cocke-Younger-Kasami) Untuk Memeriksa Pola Kalimat Baku Bahasa Indonesia. Penelitian ini tentang analisa perbandingan tingkat akurasi antara algoritma LCP (Left-corner Parsing) dengan algoritma CYK (Cocke-Younger-Kasami) menggunakan aplikasi simulator sebagai pemeriksa pola kalimat bahasa baku. Hasil dari penelilitian ini adalah tingkat akurasi algoritma CYK lebih besar dibandingkan dengan LCP. Penelitian ini masih terbatas pada POS Tag yang digunakan.

2.2 Data dan Sumber Data

Data yang digunakan adalah postingan pengguna twitter yang dikenal sebagai kicauan (tweet). Ada dua jenis penggunaan bahasa Indonesia dalam sebuah tweet, yaitu bahasa Indonesia formal dan bahasa Indonesia non-formal. Bahasa Indonesia formal bersifat struktural, sedangkan bahasa Indonesia non-formal tidak memperdulikan struktur kalimat[4]. Data ini kemudian kemudian dibagi menjadi dua data, yakni data train untuk kepentingan pembuatan model dan data uji untuk menguji performa. Selanjutnya, Data train dibagi lagi menjadi dua jenis data yakni data train POS Tagger dan data train Parser. Tabel 1 menunjukkan jumlah data keseluruhan, data train maupun data uji yang digunakan.

Tabel 1. Data yang Digunakan

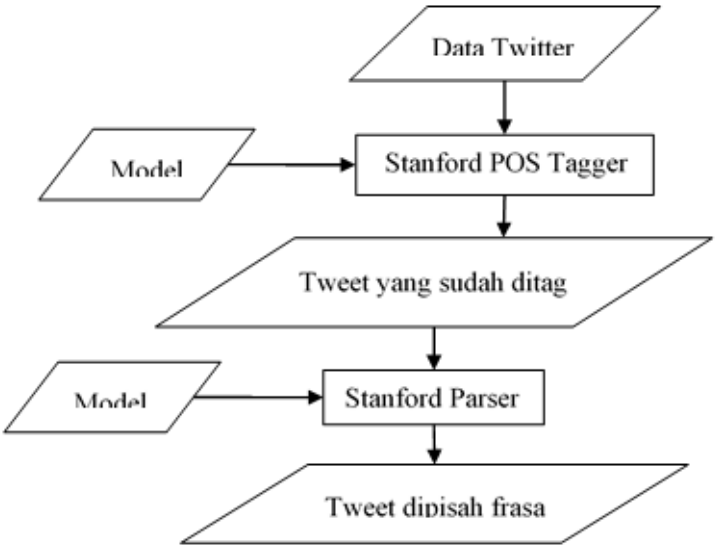
Jenis data	Jumlah tweet	
	Train	uji
Train POS Tag	500	100
Train Parser	250	100
Total	750	200

2.3 Left-corner Parsing

Left-corner parsing adalah metode gabungan top-down parsing dan bottom-up parsing. Adanya penggabungan kedua metode ini dikarenakan terdapat kekurangan pada kedua metode sebelumnya. Misalnya pada metode top-down parsing, metode ini tidak dapat menganggani grammar dengan left recursion. Sedangkan bottom-up parsing, tidak dapat menanggani empty production [3][5].

Penelitian ini mengadaptasi metode left-corner sebagai berikut:
1. Menggunakan POS Tagger sebagai proses top-down untuk mengenali jenis kata.
2. Menggunakan shift reduce sebagai proses bottom-up untuk melakukan proses parsing.

Gambar 1 berikut adalah gambaran tugas akhir yang dimodelkan dalam sebuah arsitektur sistem.



Gambar 1. Arsitektur Sistem

2.3.1 Tagging

Part-of-speech (POS) tagging, adalah sebuah proses pelabelan kelas kata secara otomatis pada sebuah kalimat [6][7]. Pada pendeketan ini digunakan proses anotasi POS Tagging sebagai proses dari top-down. Proses ini membutuhkan tagset yang digunakan untuk pemberian label pada setiap kata. Tagset yang digunakan merujuk pada tagset bahasa Indonesia yang dikembangkan oleh Universitas Indonesia. Berikut adalah tagset yang digunakan pada proses anotasI [6].

Tabel 2. Tagset Bahasa Indonesia

No	Tagset	Deskripsi	Contoh	No	Tagset	Deskripsi	Contoh
1	CC	koordinator	Dan, tetapi, atau	13	PR	Pronomina penunjuk	Ini, itu, sini
2	CD	Numeral kardinal	Dua, juta, banyak	14	PRP	Pronomina persona	Saya, kami, kalian
3	OD	Numeralia ordinal	Ketiga, pertama	15	RB	adverbia	Sangat, hanya, segera
4	DT	artikula	Para, sang, si	16	RP	partikel	Pun, lah, kah
5	FW	Kata bahasa asing	Climate, change	17	SC	Konjungtor subordinatif	Sejak, jika, tanpa

6	IN	Preposisi	Dalam, dengan. Di, ke	18	SYM	simbol	+, $,@
7	JJ	Adjektiva atau kata sifat	Bersih, marah	19	UH	interjeksi	Oh, aduh, ayo
8	MD	Verba modal	Boleh, harus,	20	VB	Verba atau kata kerja	Mengatur, pergi, bekerja
9	NEG	Kata ingkar	Tidak, belum, jangan	21	WH	Pronomina penanya	Siapa, apa, dimana
10	NN	Nomina atau kata benda	Monyet, atas, nanti	22	X	Kata ambigu	statemen
11	NNP	Proper noun	Indonesia, Januari	23	Z	Tanda baca	" . , ? !
12	NND	Nomina ukuran	Orang, ton, helai				

Contoh hasil dari proses tagging menggunakan tagset bahasa Indonesia ditampilkan pada Tabel 3.

Tabel 3. Hasil Proses Tagging

No	Data Uji	Hasil Tagging
1.	@BAA_UMM: listrik padam berdampak pada server krs, malam ini proses data di server sedang berjalan, tunggu stabil di hari sabtu pagi.	@BAA_UMM/@ :/Z listrik/NN padam/JJ berdampak/NN pada/IN server/NN krs/NN ,/Z malam/NN ini/PR proses/NN data/NN di/IN server/NN sedang/RB berjalan/VB ,/Z tunggu/VB stabil/JJ di/IN hari/NN sabtu/NN pagi/NN ./Z
2.	@BAA_UMM: Selamat kepada calon mahasiswa UMM, sebelum herregistrasi anda bisa mengisi data secara online melalui http://akademik.umm.ac.id.	@BAA_UMM/@ :/Z Selamat/JJ kepada/IN calon/NN mahasiswa/NN UMM/NN,/Z sebelum/SC herregistrasi/NN anda/NN bisa/MD mengisi/VB data/NN secara/JJ online/NN melalui/VB http://akademik.umm.ac.id/NN ./Z
3.	@BAA_UMM: Bagi Mahasiswa UMM Berstatus Non Aktif dan Butuh Perpanjangan masa Studi agar segera ke Biro Administrasi Akademik Sebelum 15 Januari 2018.	@BAA_UMM/@ :/Z Bagi/IN Mahasiswa/NN UMM/NN Berstatus/NN Non/NEG Aktif/NN dan/CC Butuh/NN Perpanjangan/NN masa/NN Studi/NN agar/SC segera/VB ke/IN Biro/NNP Administrasi/NN Akademik/NN Sebelum/NN 15/CD Januari/NN 2018/CD./Z

2.3.2 Parsing

Parsing adalah proses penguraian sebuah inputan dengan memecah – mecah rangkaian masukan hingga menghasilkan suatu pohon uraian (parse tree) [3][5]. Analisa sintaksis Pada proses pembuatan model parsing membutuhkan file treebank dalam format Penn treebank. Dengan Context-Free Grammar (CFG) sebagai acuan ntuk mengolah masukan agar mendapatkan hasil pola kalimat.

Grammar yang digunakan untuk membentuk pola kalimat dapat dilihat pada Tabel 4 Context-Free Grammar Bahasa Indonesia.

Tabel 4. Grammar Bahasa Indonesia

NP	VP	PP	ADJP	ADVP
@ + Z	VP + VP	PP + PP	ADJP + ADJP	ADVP + ADVP
NP + NP	WH + VP	IN + NP	JJ + IN	WH + ADVP
WH + NP	CC + VP	CC + PP	WH + ADJP	CC + ADVP
CC + NP	DT + VP	MD + PP	CC + ADJP	ADVP + MD
DT + NP	MD + VP	SC + PP	MD + ADJP	PR + ADVP
NP + MD	VP + MD	IN + VP	ADJP + MD	SC + ADVP

WH + PR	VP + PR	NN + PP	SC + ADJP	NEG + ADVP
NP + PR	PR + VP	NNP + PP	NEG + ADJP	PRP + ADVP
SC + MD	SC + VP	IN + CD	PRP + JJ	RB + NP
PRP + MD	NEG + VP	PRP + PP	CD + ADJP	ADVP + NN
SC + NP	JJ + VP	JJ + PP	ADJP + PR	ADVP + CD
WH + SC	VP + JJ	IN + PRP	JJ + ADVP	ADVP + PRP
NEG + NP	VP + CD	IN + ADVP	ADJP + NN	RB + VP
JJ + NP	CP + PRP		PR + ADJP	RB + PP
NP + JJ	PRP + VP			RB + ADJP
CD + NP	VB + ADVP			
NP + CD	VB + NP			
PRP + NP	VP + NN			
NN + ADJP	VB + ADJP			
NN + VP	DT + VP			
NN + ADVP	VB + PP			
NN + VB				
SC + WH				
NN + PP				
OD + NP				
NP + OD				
DT + NP				

Contoh hasil dari proses parsing menggunakan grammar diatas ditampilkan pada Tabel 5.

Tabel 5. Hasil Proses Parsing

No	Data Uji	Hasil Parsing
1.	@BAA_UMM : listrik padam berdampak pada server krs, malam ini proses data di server sedang berjalan, tunggu stabil di hari sabtu pagi.	(ROOT (S (Z :) (NP (NN Listrik) (JJ padam)) (VP (VB berdampak) (PP (IN pada) (NP (NN server) (NN KRS)))) (Z ,) (NP (NN malam) (NN ini) (NP (NN server) (NN data))) (PP (IN di) (NP (NN server) (VP (MD sedang) (VB berjalan)))) (Z ,) (VP (VB tunggu) (JJ stabil)) (PP (IN di) (NP (NN hari) (NP (NN sabtu) (NN pagi)))) (Z .)))
2.	@BAA_UMM : Selamat kepada calon mahasiswa UMM, sebelum herregistrasi anda bisa mengisi data secara online melalui http://akademik.umm.ac.id .	(ROOT (S (Z :) (S (NP (NP (JJ Selamat) (PP (IN kepada) (NP (NN calon) (NP (NN mahasiswa) (NNP UMM))))) (Z ,) (SC sebelum) (NP (NN herregistrasi) (NN anda)))

		(VP (VP (MD bisa) (VB mengisi)) (NP (NN data) (JJ secara) (NP (NN online) (VP (VB melalui) (NN http://akademik.umm.ac.id/))))) (Z .))))
3.	@BAA_UMM : Bagi Mahasiswa UMM Berstatus Non Aktif dan Butuh Perpanjangan masa Studi agar segera ke Biro Administrasi Akademik Sebelum 15 Januari 2018.	ROOT (S (Z :) (PP (IN bagi) (NP (NN Mahasiswa) (NP (NNP UMM) (NNP bestatus) (ADJP (NEG non) (ADJP (JJ aktif) (NP (CC dan) (NP (JJ butuh) (NP (NN perpanjangan) (NP (NN masa) (NN studi))))))))) (VP (SC agar) (VP (VB segera) (PP (IN ke) (NP (NNP Biro) (ADVP (NN akademik) (RB sebelum)))))) (NP (CD 15) (NP (NNP Januari) (CD 2018))) (Z .)))

3. Hasil Penelitian dan Pembahasan

Dari pengujian yang dilakukan pada dua model parser bahasa Indonesia yakni menggunakan pendekatan Left-corner dan model parser bahasa Indonesia menggunakan pendekatan Bottom-up. Nilai yang di peroleh berdasarkan hasil pengujian pada data uji, nilai-nilai yang didapat memiliki perbandingan yang cukup besar. Berikut adalah Tabel 6 perbandingan hasil pengujian. Berikut adalah tabel perbandingan hasil pengujian.

Tabel 6. Hasil Pengujian

Hasil Pengujian	Left-corner	Bottom-up
Level Precision	88,29%	68,7%
Level Recall	68,3%	47,12%
F1 Measure	77,02%	55,9%

4. Kesimpulan

Berdasarkan hasil implementasi dan pengujian *Parsing* Twitter Menggunakan Metode Left-Corner Parsing Dengan Memanfaatkan POS Tagger, penulis menyimpulkan bahwa pembuatan model parser menggunakan pendekatan Left-corner yakni dengan menggabungkan metode top-down dan bottom-up memiliki nilai lebih besar dibandingkan pembuatan model parser menggunakan pendekatan bottom-up saja. Nilai precision pada pendekatan left-corner sebesar 88,29%, nilai recall sebesar 68,3% dan F1 measure sebesar 77,02%. Sedangkan nilai precision dari bottom-up sebesar 68,7%, recall sebesar 47,12% dan F1 measure sebesar 55,9%.

Nilai yang didapat dari perbandingan kedua model cukup besar, hal ini disebabkan karena pada proses top-down pengenalan kelas kata pada tiap kata sangat berpengaruh pada pengenalan struktur kalimat untuk proses selanjutnya atau bottom-up.

Referensi

[1] T. E. Damayanti, “Pemanfaatan Twitter sebagai Media Information Sharing di Perpustakaan (Studi Kasus Tentang Pemanfaatan Media Sosial Twitter Sebagai Media Information Sharing di Perpustakaan Wilayah Kota Surabaya),” *J. Airlangga Univ.*, vol. 3, no. 2, 2014.

[2] N. K. Wangsanegara and B. Subaeki, “Implementasi Natural Language Processing Dalam Pengukuran Ketepatan Ejaan Yang Disempurnakan (EYD) Pada Abstrak Skripsi

Menggunakan Algoritma Fuzzy Logic Jurusan Teknik Informatika , Fakultas Sains dan Teknologi UIN Sunan Gunung Djati Bandung Ejaan yang," vol. 8, no. 2, 2015.

[3] J. Suciadi, "Studi Analisis Metode-Metode Parsing Dan Interpretasi Semantik Pada Natural Language Processing," *J. Inform. Fak. Teknol. Ind. Univ. Kristen Petra*, vol. 2, no. 1, pp. 13–22, 2014.

[4] D. Tata and B. Baku, "Kesantunan Berbahasa Indonesia Sebagai Upaya," vol. XVII, pp. 17–25, 2015.

[5] F. Khusaini and F. Kurniawan, "Implementasi Left Corner Parsing Untuk Pembelajaran Grammar Bahasa Inggris Pada Game 3D Adventure 'Go To London,'" *Matics*, vol. 5, no. 3, 2013.

[6] A. Dinakaramani, F. Rashel, A. Luthfi, and R. Manurung, "Designing an Indonesian part of speech tagset and manually tagged Indonesian corpus," *Proc. Int. Conf. Asian Lang. Process. 2014, IALP 2014*, pp. 66–69, 2014.

[7] N. Sabloak, "Part-of-Speech (POS) Tagging Bahasa Indonesia Menggunakan Algoritma Viterbi," no. x, pp. 1–11, 2016.

ISSN: 2714-7975; E-ISSN: 2716-1382