RULE-BASED LEXICON-BASED POS TAGGER UNTUK TEKS BERBAHASA JAWA

Skripsi

oleh
DOMINICUS ARIEL CHRISTIANTO
71140082

PROGRAM STUDI INFORMATIKA FAKULTAS TEKNOLOGI INFORMASI
UNIVERSITAS KRISTEN DUTA WACANA
2018

RULE-BASED LEXICON-BASED POS TAGGER UNTUK TEKS BERBAHASA JAWA

Skripsi

Diajukan kepada Program Studi Informatika Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana
Sebagai Salah Satu Syarat dalam Memperoleh Gelar
Sarjana Komputer

Disusun oleh

DOMINICUS ARIEL CHRISTIANTO
71140082

PROGRAM STUDI INFORMATIKA FAKULTAS TEKNOLOGI INFORMASI
UNIVERSITAS KRISTEN DUTA WACANA
2018

PERNYATAAN KEASLIAN SKRIPSI

Saya menyatakan dengan sesungguhnya bahwa skripsi dengan judul:

RULE-BASED LEXICON-BASED POS TAGGER UNTUK TEKS BERBAHASA JAWA

yang saya kerjakan untuk melengkapi sebagian persyaratan menjadi Sarjana Komputer pada pendidikan Sarjana Program Studi Informatika Fakultas Teknologi Informasi Universitas Kristen Duta Wacana, bukan merupakan tiruan atau duplikasi dari skripsi kesarjanaan di lingkungan Universitas Kristen Duta Wacana maupun di Perguruan Tinggi atau instansi manapun, kecuali bagian yang sumber informasinya dicantumkan sebagaimana mestinya.

Jika dikemudian hari didapati bahwa hasil skripsi ini adalah hasil plagiasi atau tiruan dari skripsi lain, saya bersedia dikenai sanksi yakni pencabutan gelar kesarjanaan saya.

Yogyakarta, 21 Juni 2018

DOMINICUS ARIEL CHRISTIANTO
71140082

HALAMAN PERSETUJUAN

Judul Skripsi	:	RULE-BASED LEXICON-BASED POS TAGGER UNTUK TEKS BERBAHASA JAWA
Nama Mahasiswa	:	DOMINICUS ARIEL CHRISTIANTO
N I M	:	71140082
Matakuliah	:	Skripsi (Tugas Akhir)
Kode	:	TIW276
Semester	:	Genap
Tahun Akademik	:	2017/2018

Telah diperiksa dan disetujui di
Yogyakarta,
Pada tanggal 21 Juni 2018

Dosen Pembimbing I	Dosen Pembimbing II
Lucia Dwi Krisnawati, Dr. Phil.	Danny Sebastian, S.Kom., M.M., M.T.

HALAMAN PENGESAHAN

RULE-BASED LEXICON-BASED POS TAGGER UNTUK TEKS BERBAHASA JAWA

Oleh: DOMINICUS ARIEL CHRISTIANTO / 71140082

Dipertahankan di depan Dewan Penguji Skripsi
Program Studi Informatika Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana - Yogyakarta
Dan dinyatakan diterima untuk memenuhi salah satu syarat memperoleh gelar
Sarjana Komputer
pada tanggal 30 Mei 2018

Yogyakarta, 21 Juni 2018
Mengesahkan,

Dewan Penguji:

1. Lucia Dwi Krisnawati, Dr. Phil.
2. Danny Sebastian, S.Kom., M.M., M.T.
3. Budi Susanto, SKom.,M.T.
4. Yuan Lukito, S.Kom., M.Cs.

Dekan	Ketua Program Studi
(Budi Susanto, S.Kom., M.T.)	(Gloria Virginia, Ph.D.)

UCAPAN TERIMA KASIH

Pertama-tama Penulis ingin mengucapkan syukur dan terima kasih kepada Tuhan yang Maha Kuasa atas berkat dan rahmat-Nya, penulis mampu menyelesaikan skripsi berjudul "Rule-Based Lexicon-Based Pos Tagger untuk Teks Berbahasa Jawa" dengan baik.

Meskipun banyak terdapat halangan dan hambatan selama mengerjakan skripsi ini, Penulis mendapatkan bantuan, dukungan dan kerjasama dari berbagai pihak sehingga Penulis mampu menyelesaikan skripsi ini. Maka dari itu, Penulis ingin mengucapkan terima kasih terkhusus pada:

1. S. Liem Kiem Hok dan Tuti Cahyaningsih selaku orang tua Penulis yang selalu memberikan bantuan baik berupa material maupun moral bagi Penulis untuk bisa menyelesaikan skripsi ini.
2. Bapak Budi Susanto, S.Kom., M.T. selaku Dekan Fakultas Teknologi Informasi Universitas Kristen Duta Wacana yang memberikan banyak inspirasi bagi Penulis.
3. Ibu Dr. Lucia Dwi Krisnawati dan Bapak Danny Sebastian S. Kom., M.M. M. T. selaku dosen pembimbing I dan dosen pembimbing II yang senantiasa memberikan arahan bagi Penulis dalam menyelesaikan tugas akhir ini.
4. Teman-teman dari kelompok A.H.I.E yang selalu senantiasa memberikan semangat dan menemani pengerjaan skripsi Penulis hingga subuh selama mengerjakan tugas akhir.
5. Teman-teman Blibli Future Program Batch 1 dan keluarga besar PT. Global Digital Niaga yang menjadi sumber motivasi bagi Penulis untuk segera menyelesaikan skripsi ini.
6. Teman-teman kelompok AVA di Semarang yang senantiasa memberikan bantuan moral pada Penulis.

7. Keluarga besar Cerita Kopi Yogyakarta yang selalu memberikan tempat dan menyeduhkan *marocchino* bagi Penulis selama menuliskan skripsi di malam hari.
8. Seluruh pihak yang tidak dapat dituliskan satu-persatu yang telah secara langsung maupun tidak langsung memberikan bantuan bagi Penulis.

Peneliti menyadari bahwa masih banyak kekurangan dalam penelitian ini, baik dalam penulisan dan pembahasan. Akhir kata peneliti mengucapkan terima kasih kepada semua pihak yang telah berkontribusi dalam penelitian tugas akhir ini. Peneliti juga berharap semoga tugas akhir ini dapat bermanfaat bagi para pembaca.

INTISARI

RULE-BASED LEXICON-BASED POS TAGGER UNTUK TEKS BERBAHASA JAWA

Bahasa Jawa merupakan salah satu bahasa daerah yang ada dan dipakai oleh mayoritas orang-orang di pulau Jawa di Indonesia. Sayangnya, hingga saat ini masih sedikit sekali produk-produk teknologi informasi yang menggunakan dan melakukan pengolahan pada bahasa Jawa. Hal ini disebabkan karena kurangnya data-data yang dibutuhkan untuk memproses teks berbahasa Jawa, dimana salah satu data yang cukup penting adalah kelas kata.

Part-of-Speech tagging merupakan salah satu cara dari bidang pemrosesan bahasa Natural yang dapat menghasilkan data berupa kelas kata. Penggunaan aturan dan leksikon untuk memberikan label kelas kata adalah salah satu cara yang dapat digunakan untuk mengatasi ketidaktersediaan data latih berupa kata yang memiliki label kelas kata. Penelitian ini akan berpusat pada 3 macam kelas kata, yaitu nama entitas, kata kerja dan kata benda.

Penelitian ini berhasil memberikan nilai akurasi sebesar 80.22% dan *F-1 measure* sebesar 79.62%, dengan catatan bahwa eilai akhir dapat mencapai hasil yang tinggi dikarenakan banyaknya jumlah tag 0 yang bukan menjadi inti dari penelitian.. Kelemahan dari sistem dalam penilitian ini adalah ketidakmampuan untuk memilih kelas kata dari kata yang memiliki kelas kata ganda dan hasil dari sistem semacam ini sangat bergantung pada kelengkapan dari kelas kata yang dijadikan sebagai leksikon.

Kata Kunci: Bahasa Jawa, *Pos tagging*, leksikon, *rule-based*

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR LAMPIRAN

INTISARI

RULE-BASED LEXICON-BASED POS TAGGER UNTUK TEKS BERBAHASA JAWA

Bahasa Jawa merupakan salah satu bahasa daerah yang ada dan dipakai oleh mayoritas orang-orang di pulau Jawa di Indonesia. Sayangnya, hingga saat ini masih sedikit sekali produk-produk teknologi informasi yang menggunakan dan melakukan pengolahan pada bahasa Jawa. Hal ini disebabkan karena kurangnya data-data yang dibutuhkan untuk memproses teks berbahasa Jawa, dimana salah satu data yang cukup penting adalah kelas kata.

Part-of-Speech tagging merupakan salah satu cara dari bidang pemrosesan bahasa Natural yang dapat menghasilkan data berupa kelas kata. Penggunaan aturan dan leksikon untuk memberikan label kelas kata adalah salah satu cara yang dapat digunakan untuk mengatasi ketidaktersediaan data latih berupa kata yang memiliki label kelas kata. Penelitian ini akan berpusat pada 3 macam kelas kata, yaitu nama entitas, kata kerja dan kata benda.

Penelitian ini berhasil memberikan nilai akurasi sebesar 80.22% dan *F-1 measure* sebesar 79.62%, dengan catatan bahwa eilai akhir dapat mencapai hasil yang tinggi dikarenakan banyaknya jumlah tag 0 yang bukan menjadi inti dari penelitian.. Kelemahan dari sistem dalam penilitian ini adalah ketidakmampuan untuk memilih kelas kata dari kata yang memiliki kelas kata ganda dan hasil dari sistem semacam ini sangat bergantung pada kelengkapan dari kelas kata yang dijadikan sebagai leksikon.

Kata Kunci: Bahasa Jawa, *Pos tagging*, leksikon, *rule-based*

BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Bahasa adalah salah satu metode yang digunakan oleh manusia untuk berkomunikasi satu sama lain. Metode komunikasi ini memiliki banyak sekali variasi dan akan selalu berkembang seiring berjalannya waktu. Di Indonesia, salah satu bahasa yang menjadi warisan budaya adalah bahasa Jawa yang mayoritas digunakan di Puau Jawa. Selain itu, menurut sensus pada tahun 200 dari Ethnologue (Simons & Fennig, 2018), jumlah penutur Bahasa Jawa di Indonesia mencapai 84.000.000 manusia. Meskipun demikian, tidak banyak penelitian yang mengeksplorasi Bahasa Jawa dalam bentuk digital.

POS Tagging adalah sebuah proses pemberian label kelas kata pada suatu kata dalam sebuah kalimat. Proses ini merupakan dasar untuk sistem pengolahan bahasa natural yang lebih lanjut. Dengan mengetahui kelas kata, dapat dilakukan pengambangan sistem berbasis NLP berikutnya seperti *speech recognition* dan *machine translation*. Meskipun label kelas kata ini penting, namun jumlah korpus bahasa Jawa dengan label kelas kata hampir tidak ada, dan untuk melabeli dokumen-dokumen tersebut secara manual membutuhkan waktu yang dapat dikatakan tidak sedikit dan adanya seorang ahli untuk memberikan label secara tepat.

Melihat hal-hal tersebut, diperlukan sebuah sistem untuk memberi label pada dokumen bahasa Jawa secara otomatis. Dengan pemberian label secara otomatis, dimungkinkan untuk memberi label dengan waktu yang lebih singkat, sehingga dapat membantu penelitian yang akan dibuat selanjutnya.

1.2 Rumusan Masalah

Berdasarkan latar belakang tersebut, penelitian ini dibuat untuk menjawab pertanyaan berikut:

1. Bagaimana cara pengkonstruksian aturan-aturan untuk memberi label kelas kata pada kata dalam teks berbahasa Jawa?
2. Bagimana hasil evaluasi dari aturan yang dikonstruksi untuk memberi label kelas kata dalam teks berbahasa Jawa secara otomatis?

1.3 Batasan Masalah

Dalam penelitian ini diberikan batasan masalah sebagai berikut:

1. Kelas kata yang akan diuji terbatas pada kata benda dan kata kerja.
2. Aturan yang digunakan terbatas pada aturan-aturan yang berupa imbuhan pada kata
3. Sistem tidak menangani kata yang berlabel ganda.

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah pada bagian sebelumnya, penilitian ini diajukan untuk membangun sebuah sistem untuk pemberian label kelas kata pada suatu kata dalam teks berbahasa Jawa. Khususnya pada kata kerja dan kata benda, serta untuk menghitung evaluasi dengan metode penghitungan akurasi dan *F-1 measure* dari sistem pelabelan kelas kata bahasa Jawa yang dibangun.

1.5 Manfaat Penelitian

Penelitian ini dapat memberikan manfaat sebagai berikut:

1. Memberikan sistem untuk pemberian kelas kata berbasi aturan dan leksikon pada teks berbahasa Jawa.
2. Memberikan label kelas kata pada beberapa dokumen dalam korpus TRAWACA

1.6 Metodologi Penelitian

Metode yang akan digunakan dalam penelitian ini adalah sebagai berikut

1. Studi Pustaka

Studi pustaka akan dilakukan pada literatur terkhusus mengenai linguistik bahasa Jawa. Metode ini perlu dilakukan untuk menentukan aturan untuk pemberian label dengan tepat.

2. Konsultasi

Konsultasi akan dilakukan kepada dosen pembimbing secara teratur dengan tujuan supaya penelitian selalu terarah.

3. Pengumpulan Data

Data dokumen yang akan digunakan dalam penelitian ini berasal dari korpus TRAWACA (Mahastama & Krisnawati, 2017) yang berupa teks berbahasa Jawa, ditulis menggunakan aksara latin tanpa label kelas kata sejumlah 30 dokumen.

4. Pembangunan Sistem

Sistem akan dibangun dengan mengikuti rancangan yang dituliskan dalam Bab 3.

5. Evaluasi

Evaluasi akan dilakukan dengan menghitung nilai akurasi dan *F-1 measure* dari hasil keluaran sistem.

1.7 Sistematika Penulisan

Laporan ini dibagi menjadi 5 bab dalam penulisannya. Bab-bab tersebut antara lain adalah Bab 1 Pendahuluan, Bab 2 Tinjauan Pustaka, Bab 3 Analisis dan Perancangan Sistem, Bab 4 Implementasi dan Analisis Sistem, dan Bab 5 Kesimpulan dan Saran.

Bab 1 Pendahulan mencakup latar belakang masalah, perumusan masalah, batasan masalah tujuan penelitian, metode penelitian dan sistematika penulisan.

Bab 2 Tinjauan Pustaka berisi landasan teori yang diperlukan dan mendukung penelitian dan pembuatan sistem.

Bab 3 Analisis dan perancangan sistem meliputi analisis mengenai sistem yang akan dibangun dan juga perancangan atas sistem yang akan dibangun, meliputi masukan, keluaran dan evaluasi atas sistem.

Bab 4 Implementasi dan Analisis Sistem berisi tentang implementasi atas rancangan sistem yang sudah dipaparkan pada bab sebelumnya.

Bab 5 Kesimpulan dan Saran berisi kesimpulan atas penelitian yang telah dilakukan serta saran untuk penelitian selanjutnya.

BAB 5
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan analisis yang sudah dipaparkan dalam bab 4, penelitian ini memberikan kesimpulan sebagai berikut:

1. Sistem pelabelan untuk kelas kata nama entitas, kata benda dan kata kerja pada teks berbahasa Jawa berdasarkan aturan mampu dibangun menggunakan aturan imbuhan dan kamus kata dasar.
2. Sistem pelabelan untuk kelas kata berbahasa Jawa berbasiskan aturan imbuhan dan kamus kata dasar dalam penelitian ini memberikan nilai akurasi sebesar 80.22% dan *F-1 measure* sebesar 79.62%. Nilai akhir dapat mencapai hasil yang tinggi dikarenakan banyaknya jumlah tag 0 yang bukan menjadi inti dari penelitian.

5.2 Saran

Dari hasil penelitian ini, terdapat beberapa hal yang dapat menjadi masukan untuk penelitian untuk pemberian label kelas kata untuk bahasa Jawa berikutnya. Saran yang dihasilkan dalam penelitian ini antara lain adalah sebagai berikut:

1. Penggunaan aturan imbuhan sebaiknya dicoba dengan mengkombinasikan urutan dari aturan imbuhan yang ada.
2. Penambahan entri untuk kamus kelas kata dasar dan penambahan variasi untuk kelas kata dasar.
3. Pengecekan hubungan antar kelas kata untuk mengetahui kelas kata dari kata yang memiliki lebih dari 1 kemungkinan kelas kata.

DAFTAR PUSTAKA

Elnatan, Y. K. (2017). *POS Tagging Bahasa Indonesia Berbasis Aturan dan Lexicon (Kata Kerja + Kata Keterangan).* (Undergraduate thesis, Duta Wacana Christian University, 2017). Retrieved from http://sinta.ukdw.ac.id.

Garrette, D., Mielens, J., & Baldridge, J. (2013). Real-World Semi-Supervised Learning of POS-Taggers for Low-Resource Languages. *Proceedings of The 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 583-592.

Jurafsky, D., & Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Preocessing Computational Lingusistics, and Speech Recognition.* New Jersey: Prentice Hall.

Krisnawati, L. D., & Schulz, K. U. (2013). Plagiarism Detection for Indonesian texts. *Proceedings of Internation Conference on Information Integration and Web-based Application & Services*, 595.

Mahastama, A., & Krisnawati, L. D. (2017). Histogram Peak-Based Binarization for Historical Documents. *International Conference on Smart Cities, Automation & Intelligent Computing Systems (ICON-SONICS) 2017)* (hal. 93-98). Yogyakarta: IEEE.

Oakes, M. (2009). Javanese. Dalam B. Comrie, *The world's major languages* (hal. 819-832). Routledge.

Paroubek, P. (2007). Evaluating Part-of-Speech Tagging and Parsing. Dalam L. Dybkjær, H. Hemsen, & W. Minker, *Evaluation of Text and Speech Systems. Text, Speech and Language Techonology* (Vol. 37, hal. 99-124). Dordrecht: Springer.

Pisceldo, F., Manurung, R., & Adriani, M. (2009). Probabilistic part-of-speech tagging for bahasa indonesia. *Third International MALINDO Workshop*.

Pramudita, H. R., Utami, E., & Amborowati, A. (2016). Pengaruh Part of Speech Tagging berbasis Aturan dan Distribusi Probabilitas Maksimum Entropy untuk Bahasa Jawa. *Jurnal Buana Informatika, 7*(4).

Simons, G., & Fennig, C. (2018). *Ethnologue: Languages of The World, Twenty-first Edition.* Dallas, Texas: SIL International.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management, 45*(4), 427-437.

Sudaryanto. (1991). *Tata bahasa baku bahasa jawa.* Duta Wacana University Press.

Wedhawati. (2006). *Tata bahasa jawa mutakhir.* Yogyakarta: Kanisius.

Widhiyanti, K., & Harjoko, A. (2013). POS Tagging Bahasa Indonesia dengan HMM dan Rule Based. *Informatika: Jurnal Teknologi Komputer dan Informatika, 8*(2).

Wiktionary:Kamus bahasa Jawa – bahasa Indonesia. (t.thn.). Dipetik Mei 3, 2018, dari Wiktionary: https://id.wiktionary.org/wiki/Wiktionary:Kamus_bahasa_Jawa_%E2%80%93_bahasa_Indonesia