PENANGANAN OOV (OUT OF VOCABULARY) PADA POS TAGGING HIDDEN MARKOV MODEL

I Wayan Hendra Maha Putra[1], Imelda Atastina[2], Alfian Akbar Gozali[3]

[1]Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

Abstrak

POS Tagging merupakan salah satu teknik dalam Data Mining untuk mengenali jenis-jenis kata yang ada. POS Tagging ini sangat berguna dalam berbagai proses dalam pemrosesan bahasa Natural (NLP) seperti pengolahan teks suara, teks dan ilmu bahasa. POS Tagging dengan menggunakan metode Supervised Hidden Markov Model memiliki beberapa kekurangan,salah satunya adalah penanganan OOV atau Out-of-Vocabulary. Penanganan OOV ini adalah penanganan terhadap kata yang tidak diketahui atau kata yang sebelumnya belum pernah muncul pada saat training, sehingga hal ini dapat mengurangi akurasi dari proses tagging tersebut karena kemungkinan pemberian tag yang salah cukup tinggi. Algoritma Forward dan algoritma Backward merupakan salah satu algoritma yang memiliki karakterisktik yang dapat digunakan untuk menangani masalah OOV.

Kata Kunci : Pos tagging bahasa Indonesia, Supervised Hidden Markov Model, Algoritma

Abstract

POS Tagging is one of the techniques in Data Mining to identify the types of words that exist. POS Tagging is very useful in a variety of processes on Natural language processing (NLP) such as text processing voice, text and linguistics. POS Tagging using Supervised Hidden Markov Model method has some drawbacks, one of which is the handling of OOV or Out-of-Vocabulary. Handling OOV is handling the unknown word or words not previously appeared at the time of training, so it can reduce the accuracy of the tagging process is due to the possibility of giving the wrong tag is quite high. Forward algorithm and Backward algorithm is one of the algorithms that have the characteristics that can be used to handle the OOV problem.

Keywords : Indonesian Pos Tagging, Supervised Hidden Markov Model, Forward

1. PENDAHULUAN

1.1 Latar belakang Masalah

Bahasa Indonesia adalah alat yang mampu menjembatani penduduk Indonesia yang terdiri dari berbagai suku dan bahasa untuk dapat berkomunikasi satu sama lainnya. Dalam Bahasa Indonesia dikenal beberapa jenis kata dalam sebuah kalimat seperti kata kerja, kata sifat, kata keterangan, subjek dan lain sebagainya. Setiap jenis kata tersebut tentunya memiliki fungsi yang berbeda-beda. Dalam menentukan jenis kata tersebut secara manual tentunya diperlukan waktu dan biaya yang tidak sedikit. Oleh karena itu, diperlukan suatu cara untuk menentukan jenis kata secara otomatis dengan teknik yang dinamakan *Part-Of-Speech Tagging.*

Part-Of-Speech Tagging atau yang biasa disingkat dengan POS Tagging merupakan salah satu teknik dalam *Data Mining* untuk mengenali jenis-jenis kata yang ada. POS Tagging ini sangat berguna dalam berbagai proses dalam pemrosesan bahasa Natural (NLP) seperti pengolahan teks suara, teks dan ilmu bahasa. Ada beberapa metode yang dapat digunakan untuk melakukan Tagging ini, salah satu diantaranya adalah *Hidden Markov Model*, Ruled based, Conditional Random Field dan lainnya.

POS Tagging dengan menggunakan metode *Hidden Markov Model* memiliki beberapa kekurangan,salah satunya adalah penanganan OOV atau Out-of-Vocabulary. Penanganan OOV ini adalah penanganan terhadap kata yang tidak diketahui atau kata yang sebelumnya belum pernah muncul pada saat training, sehingga hal ini dapat mengurangi akurasi dari proses tagging tersebut karena kemungkinan pemberian tag yang salah cukup tinggi. Untuk itu perlu dilakukan sebuah proses yang dapat digunakan untuk menangani masalah seperti ini.

Dalam Bahasa Indonesia sendiri OOV dapat disebabkan oleh beberapa faktor salah satunya adalah perubahan jenis kata akibat adanya penambahan imbuhan terhadap sebuah kata, misalnya kata “makan” merupakan kata kerja mendapatkan imbuhan berupa akhiran “an” menjadi “makanan” yang termasuk kata benda. Akan tetapi penanganan tersebut masih terdapat kekurangan yaitu bagaimana jika OOV tersebut merupakan sebuah kata dasar dan bukan kata berimbuhan tentu saja akan mengalami kesulitan. Tugas akhir ini akan menangani masalah OOV pada POS Tagging Hidden Markov Model dengan memanfaatkan karakterristik algoritma *Forward* dan *Backward* yang ada pada metode tersebut.

1.2 Rumusan Masalah

Penelitian ini dilakukan untuk menganalisis performansi metode Supervised Hidden Markov Model Bigram dalam menentukan jenis kata berbahasa Indonesia dengan penanganan oov menggunakan algoritma *Forward* dan *Backward* dan menarik kesimpulan dari hasil analisis.

1.3 Tujuan

Tujuan dari penelitian ini adalah :

1. Membuat aplikasi penentuan jabatan kata dalam kalimat berbahasa Indonesia.
2. Menganalisis performansi dari Pos Tagging menggunakan Hidden Markov Model dengan penanganan OOV dalam menentukan jabatan kata dalam kalimat berbahasa Indonesia.

1.4 Manfaat

Manfaat dari tugas akhir ini adalah:

Bagi penyusun :

1. Dapat memahami penanganan OOV pada *Hidden Markov Model Pos Tagging*

Bagi dunia pendidikan :

1. Dapat menjadi perbandingan atau referensi dalam pembuatan kamus data berbahasa Indonesia untuk pengolahan data.
2. Dapat menambah pengetahuan mengenai pemrosesan bahasa natural.

1.5 Batasan Masalah

Berikut ini merupakan batasan masalah dari penelitian:

1. Penelitian ini dilakukan pada *Pos Tagging Hidden Markov Model* bigram.
2. *Input* berupa kalimat dalam berita berbahasa Indonesia. Jumlah kalimat yang diinputkan adalah sebuah kalimat dan sebuah paragraf.
3. *Output* berupa kata-kata beserta jabatan katanya dalam kalimat yang diinputkan.

1.6 Hipotesa

Algoritma *Forward* dan *Backward* dapat menangani masalah Out-Of-Vocabulary yang terjadi pada metode HMM.

1.7 Metodologi Penyelesaian Masalah

1. Studi literatur

Pada tahap ini penulis mencari referensi-referensi yang berkaitan dengan *POS Tagging*, metode *Hidden Markov Model* dan algoritma *Backward* dan sumber lain yang masih relefan dengan judul Tugas Akhir

2. Pengumpulan data

Mencari data berupa daftar tag dan kata dalam Bahasa Indonesia sebagai bahan dalam pembuatan aplikasi.

3. Perancangan

Dalam tahap perancangan ini, akan dilakukan beberapa kegiatan seperti :

a. Perancangan desain antarmuka dari aplikasi yang dibangun.
b. Merancang penerapan Metode Hidden Markov Model dalam skema yang telah dibuat.

4. Implementasi

Tahap implementasi dari sistem ini adalah dengan membangun sebuah sistem yang telah dirancang untuk penentuan jenis kata Bahasa Indonesia menggunakan metode Supervised Hidden Markov Model dan algoritma *Backward* untuk penanganan OOV. Dalam mengimplementasikannya dibutuhkan perangkat keras dan perangkat lunak yang memadai. Berikut adalah daftar perangkat keras dan perangkat lunak yang diperlukan, yaitu :

a. Spesifikasi perangkat keras
    1. Processor : Intel(R) Core(TM) i5-2450M CPU @2.50GHz (4 CPUs)
    2. Memory : 4,00 GB
    3. Harddisk : 500 GB

b. Spesifikasi perangkat lunak
    1. Sistem Operasi : Microsoft Windows Seven
    2. Bahasa Pemograman : Java

5. Pengujian

Pada tahap pengujian akan dilakukan 2 (dua) buah proses pengujian secara berulang kali terhadap aplikasi yang dibangun yaitu :

a. Pengujian apakah aplikasi yang dibangun sudah siap dan tidak terjadi kesalahan.
b. Pengujian dengan menggunakan data sehingga menghasilkan keluaran berupa dokumen dengan kalimat ber-tag.

6. Analisis hasil

Hasil yang didapat kemudian dari pengujian dengan menggunakan aplikasi akan dianalisis, dilakukan serangkaian pengamatan sehingga diperoleh suatu kesimpulan. Adapun hal yang akan di analisis adalah bagaimana kesesuaian output yang dihasilkan, serta mengukur performansi dari algoritma yang dipakai.

7. Pembuatan laporan

Keseluruhan data, aplikasi, analisis hasil dll, disertakan pada pembuatan laporan tugas akhir. Penyusunan laporan dilakukan secara bertahap sampai proses pembuatan aplikasi selesai.

1.8 Sistematika Penulisan

Penelitian ini diuraikan dengan sistematika sebagai berikut :

Bab I Pendahuluan

Pada bab ini berisi uraian mengenai latar belakang, rumusan masalah, tujuan, batasan masalah, metodologi penelitian serta sistematika penulisan.

Bab II Landasan Teori

Bab ini berisi literatur yang relevan dengan permasalahan yang diangkat yang diperoleh dari berbagai sumber mengenai Tagging, Hidden Markov Model, dan algoritma *Backward* dan semua literatur yang berkaitan dengan Tugas Akhir ini.

Bab III Analisis dan Perancangan Sistem

Bab ini berisi uraian mengenai perangkat lunak seperti spesifikasi kebutuhan perangkat, perancangan struktur data dan spesifikasi proses dari perangkat lunak yang dibuat.

Bab IV Implementasi dan Analisis Pengujian Hasil

Bab ini berisi uraian mengenai peengolahan data klasifikasi yang digunakan serta analisis hasil pengujian perangkat lunak.

Bab V Kesimpulan dan Saran

Bab ini berisi kesimpulan dari seluruh sistem yang dibuat serta saran untuk pengembangan perangkat lunak.

5. Penutup

5.1 Kesimpulan

Berdasarkan hasil implementasi, pengujian dan analisis yang telah dilakukan, maka dapat diambil kesimpulan, yaitu :

1. Algoritma *Forward* dan juga algoritma *Backward* dapat digunakan sebagai salah satu solusi untuk menangani masalah OOV yang terjadi pada Pos Tagging HMM bigram.
2. Posisi dari OOV dalam sebuah kalimat cukup berpengaruh terhadap keberhasilan kesuksesan pemberian label tagset terhadap OOV tersebut. Posisi dari OOV juga cukup berpengaruh terhadap akurasi keseluruhan sistem.
3. Akurasi dari hasil dari pemberian label tagset oleh sistem sangat dipengaruhi oleh persentase jumlah OOV yang terkandung tetapi dengan jumlah OOV sebanyak 35% sistem masih menghasilkan akurasi diatas 50%.
4. Jumlah kalimat yang ada pada data training tidak memberikan pengaruh yang signifikan terhadap akurasi yang dihasilkan, namun variasi keberagaman kalimat yang lebih memiliki pengaruh terhadap akurasi.

5.2 Saran

1. Perlu dilakukan kajian lebih lanjut terhadap algoritma *Forward* dan algoritma *Backward* ketika digunakan untuk menangani masalah OOV pada HMM agar dapat lebih menghasilkan akurasi yang lebih tepat.
2. Pada perhitungan probabilitas transisi dapat ditambahkan proses smoothing sehingga dapat menghasilkan hasil yang lebih bagus. Smoothing dapat membantu dapat pencarian probabilitas transisi dengan menggabungkan antara probabilitas *bigram* dan probabilitas *unigram* dari tiap *state* yang berkemungkinan akan muncul.

DAFTAR PUSTAKA

[1] Akbar Gozali, Alfian. (2010). "Analisis Penggunaan Metode Hidden Markov Model dalam Ekstraksi Kalimat Utama Suatu Dokumen pada Information Retrieval", Tugas Akhir Institut Teknologi Telkom.

[2] Alfan Farizki Wicaksono,Ayu Purwarianti.(2010). "HMM Based Part-of-Speech Tagger for Bahasa Indonesia", School of Electrical Engineering and Informatics-Insitut Teknologi Bandung.

[3] Atika Sari, Susiana. (2008). "Kelas Kata Dalam Bahasa Indonesia Sebuah Tinjauan Stereotip Jender". Fakultas Sastra Universitas Diponegoro.

[4] Fahim Muhammad Hasan. (2006). "Comparison Of Different Pos Tagging Techniques For Some South Asian Languages", Department of Computer Science and Engineering of BRAC University.

[5] Femphy Pisceldo, Manurung, R., Adriani, Mirna. "Probabilistic Part-of-Speech Tagging for bahasa Indonesia". Third International MALINDO Workshop, colocated event ACL-IJCNLP 2009, Singapore, August 1, 2009.

[6] Hidden Markov Model. [Online]. Tersedia : http://www.comp.leeds.ac.uk/roger/HiddenMarkovModels/html_dev/main.html [23 Oktober 2012]

[7] Jurafsky, D., Martin, J.H. (2006). "Speech and Language Processing: An Introduction to Natural Language Processing" , Computational Linguistics, and Speech Recognition.

[8] Ketut Gde Manik karvana. (2012). "Analisis dan Implementasi Unsupervised Hidden Markov Model Untuk Penentuan Jenis Kata Bahasa Indonesia", Tugas Akhir Institut Teknologi Telkom.

[9] Kamus Besar Bahasa Indonesia. http://www.pusatbahasa.diknas.go.id , tanggal akses : 22 September 2012.

[10] L.Rabiner. *A tutorial on Hidden Markov Models and selected applications in speech recognition*. Proc. of IEEE, 77(2):257-286, 1989.

[11] Sri Jayadi. "Pengenalan Angka Terisolasi dengan Menggunakan Pemodelan HMM Melalui Ekstraksi Feature Mel Cepstrum Filter Bank", Fakultas Teknik Universitas Diponegoro.

[12] Thede, Scott M., Marry P. Harper, “A Second-Order Hidden Markov Model for Part-of-Speech Tagging”, School of Electrical and Computer Engineering of Purdue University, Hal 175-181.

[13] Wibisono, Yudi. (2008). “Penggunaan Hidden Markov Model untuk Kompresi Kalimat”, Tesis Institut Teknologi Bandung.