

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/322963586>

Named Entity Recognition Menggunakan Hidden Markov Model dan Algoritma Viterbi pada Teks Tanaman Obat

Article · December 2017

CITATIONS

2

READS

1,059

3 authors, including:



Mochammad Ali Fauzi

Brawijaya University

102 PUBLICATIONS 198 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



Twitter Sentiment Analysis [View project](#)



Sentiment Analysis [View project](#)

Named Entity Recognition Menggunakan Hidden Markov Model dan Algoritma Viterbi pada Teks Tanaman Obat

Agung Setiyoaji¹, Lailil Muflikhah², Mochammad Ali Fauzi³

Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Brawijaya

Email: ¹setiyoajiagung@gmail.com, ²lailil@ub.ac.id, ³moch.ali.fauzi@ub.ac.id

Abstrak

Media untuk menyampaikan informasi dapat melalui televisi, radio, social media, dan website. Website merupakan karya tulis dari seseorang terletak pada sebuah domain yang mengandung informasi. Perkembangan website semakin banyak dan informasi semakin tidak terbendung sehingga muncul permasalahan sulitnya mencari informasi yang sesuai dengan kebutuhan pengguna internet, sehingga dibutuhkan klasifikasi dan ekstraksi informasi untuk informasi yang ada dalam website. Named Entity Recognition (NER) yang merupakan turunan dari ekstraksi informasi, bertujuan untuk memudahkan mencari informasi dengan cara pemberian nama entitas pada setiap kata dalam sebuah teks. Pada penelitian ini akan dilakukan pengenalan empat entitas yaitu NAMA, TEMPAT, ZAT, dan KEGUNAAN dari teks tanaman obat. Algoritma pada penelitian ini menggunakan Hidden Markov Model (HMM) dan algoritma Viterbi. Secara keseluruhan pengenalan entitas menghasilkan nilai yang paling rendah dengan f-measure 0.41, dan yang paling tinggi dengan f-measure 0.72.

Kata Kunci: *Hidden Markov Model, pengenalan entitas bernama, tanaman obat*

Abstract

Media to convey information can be through television, radio, social media, and website. Website is a work of someone located in a domain that contains information. The development of websites more and more information is not unstoppable so that the problem arises difficult to find information in accordance with the needs of Internet users, so that the required classification and extraction of information for information on the website. Named Entity Recognition which derives from the extraction of information, NER aims to facilitate the search for information by naming entities on each word in a text. In this research will be done the introduction of four entities namely the NAME, PLACE, SUBSTANCE, and FUNCTION of the text on medicinal plants. The algorithm used Hidden Markov Model (HMM) and Viterbi algorithm. Overall entity recognition count the lowest value with f-measure 0.41, and the highest with f-measure 0.72.

Keywords: *Hidden Markov Model, named entity recognition, medicinal plant*

1. PENDAHULUAN

Tahun 2017 merupakan tahun era informasi, informasi yang di sampaikan dari berbagai bidang keilmuan meliputi keuangan, politik, kesehatan, dan masih banyak lagi. Media untuk menyampaikan informasi dapat melalui televisi, radio, social media, dan website. Website merupakan karya tulis dari seseorang terletak pada sebuah domain yang mengandung informasi, dengan harapan dapat di lihat oleh pengguna Internet (Winoto dan Irianto, 2012). Sejak pertama kali di temukannya World Wide Web tahun 1991, tahun inilah awal partam istilah

website ditemukan, setelah tiga tahun website menjadi populer dikarenakan didirikannya perusahaan yahoo pertama kali tahun 1994, setelah itu website semakin populer dengan adanya google tahun 1998 dan paypal tahun 1999, hingga pada tahun 2014 jumlah website mencapai satu milyar, hingga tahun 2017 mencapai 1.151.441.000 website, penghitungan sudah di setuju oleh World Wide Web Consortium dan World Wide Web (FoundationInternet Live Stats, 2017).

Dengan perkembangan website semakin banyak dan informasi semakin tidak terbendung sehingga muncul permasalahan sulitnya mencari

informasi yang sesuai dengan kebutuhan pengguna internet, sehingga dibutuhkan klasifikasi dan ekstraksi informasi untuk informasi yang ada dalam website (Priyanta, Hartati, dan Harjoko, 2016). Ekstraksi informasi dalam teks memiliki struktur bahasa dan entitas kata, untuk mengidentifikasi entitas kata pada kalimat muncul istilah Named Entity Recognition NER bertujuan untuk memudahkan mencari informasi dengan cara pemberian nama entitas pada setiap kata dalam sebuah teks.

Pada penelitian yang dilakukan oleh Todorovic, dkk, (2008) mengatakan dalam penelitian Named Entity Recognition (NER) menggunakan hidden markov model pada bahasa inggris dengan mengkategorikan teks pada data latih tanpa menggunakan grammer menemukan akurasi sistem tertinggi 91.71% (Todorovic, dkk, 2008). Penelitian Alfred, dkk, 2014 mengatakan mengenai named entity recognition menggunakan pendekatan ketetanggaan rule-based pada teks berita terbaru pada website mencapai akurasi sistem tertinggi 89.47%. Penelitian named entity recognition pada berita kesehatan menggunakan Support Vector Machine (SVM) mendapati akurasi tertinggi 90% (Suwarningsih, dkk, 2014). Hidden Markov Model (HMM) merupakan metode machine learning yang menggunakan model probabilitas berurutan untuk menyelesaikan masalah, dan membutuhkan data latih, dimana setiap kata data latih membutuhkan penamaan entitas kata, proses penamaan entitas kata bisa disebut POS-Tagging (Todorovic, dkk, 2008).

Agusta (2015) mengatakan setidaknya ada 50 ribu jenis tumbuhan yang ada di Indonesia, diketahui ada 7500 tanaman yang sudah diteliti dan dapat digunakan sebagai obat. Dari data tanaman 7500 yang dapat dijadikan sebagai obat mempunyai metode penyampaian informasi melalui teks paragraf pada halaman website, teks pada halaman website mengakibatkan kesulitan untuk menemukan data tanaman yang sesuai dengan kebutuhan pengguna internet. Untuk mengurangi kesulitan pencarian informasi mengenai tanaman obat pada teks paragraf maka dapat dilakukan pembuatan named entity recognition pada teks paragraf tanaman obat, untuk metode yang digunakan dalam named entity recognition ada tiga pilihan yaitu Hidden Markov Model (HMM), rule-based, dan Support Vector Machine (SVM).

Penelitian ini akan menggunakan Hidden Markov Model (HMM) dan algoritma Viterbi dalam mempermudah pencarian informasi pada

teks paragraf tanaman obat. Pemilihan hidden markov model sebagai algoritma pada penelitian ini didasari karena pada penelitian yang dilakukan Todorovic, dkk, (2008) menggunakan hidden markov model mendapatkan akurasi tinggi yaitu 91.71%, dibandingkan dengan penelitian yang dilakukan oleh Suwarningsih, dkk, (2014) dengan menggunakan metode Support Vector Machine (SVM) mencapai akurasi 90% dan hidden markov model juga unggul oleh penelitian Alfred, dkk, (2014) dengan rule-based yang memiliki akurasi sistem paling tinggi 89.47%. Untuk mengetahui nama entitas dari data uji pada sistem yang menggunakan metode hidden markov model maka menggunakan algoritma Viterbi, oleh karena itulah algoritma Viterbi digunakan pada penelitian ini (Todorovic, dkk, 2008).

2. DASAR TEORI

2.1 Ekstraksi Informasi

Ekstraksi informasi merupakan proses pengambilan informasi dari teks tidak terstruktur yang menghasilkan teks terstruktur dan rapi sekaligus mempermudah untuk menemukan informasi dari teks terstruktur (Ilyas & Khodra, 2015). Proses Ekstraksi Informasi bisa melakukan dua teknik pendekatan yaitu teknik berdasarkan aturan dan berdasar statistik, teknik berdasarkan aturan dibuat dan dirancang dengan keahlian seorang pakar, dimana pakar dengan keahlian sesuai dengan kumpulan teks yang digunakan, teknik berdasar statistik dibangun dan dibuat dengan menggunakan data latih, data latih sesuai dengan teks yang dijadikan sebagai object Ekstraksi Informasi (Ilyas & Khodra, 2015).

2.2 Preprocessing

Indurkha dan Damerau (2010) mengatakan ada dua jenis preprocessing yaitu *document triage* dan *text segmentation*, *document triage* merupakan pengambilan data dari digital yang kemudian di simpan ke dalam sebuah file teks, pengambilan data dalam *document triage* dapat dilakukan secara manual atau menggunakan sebuah fungsi. Proses *document triage* ada tiga langkah, pertama identifikasi karakter *encoding* untuk menentukan karakter dalam file tertentu, kedua deteksi bahasa yang digunakan dalam teks, ketiga melakukan penghapusan elemen, gambar, tabel, *header*, *link*, dan kode HTML (Indurkha

dan Damerau, 2010).

2.2.1 Tokenisasi

Tokenization atau tokenisasi merupakan proses pada preprosesing yang bertujuan untuk menghapus tanda penghubung pada sebuah teks, angka, tanda baca berupa titik (.), petik("), tanda seru (!) dan masih banyak lainnya, pada tokenization juga merubah semua huruf besar menjadi huruf kecil (Sekarwati & Kurniawati, 2012). Proses tokenisasi pada saat penghapusan tidak memperhatikan susunan dari kalimat, tokenisasi hanya menghapus dan tidak perlu memperhatikan susunan kalimat. Tabel 1. merupakan contoh hasil dari tokenisasi.

Tabel 1. Contoh tokenisasi

Sebelum tokenisasi	Sesudah tokenisasi
Banyak tanaman yang mati karena zat-zat "glifosat". Tanaman-tanaman yang mati meliputi bunga mawar, rumput, dan tanaman berduri.	banyak tanaman yang mati karena zat zat glifosat tanaman tanaman yang mati meliputi bunga mawar rumput dan tanaman berduri

2.2.2 Filtering

Filtering atau *stopword removal* bertujuan untuk menghilangkan kata-kata yang tidak memiliki arti, seperti kata hubung (yang, di, dan, oleh), kata yang tidak memiliki arti sudah dikelompokkan pada masing-masing bahasa yang ada di dunia. Kata yang tidak memiliki arti disebut dengan *stopword* (Sekarwati & Kurniawati, 2012). Tabel 2. merupakan contoh hasil proses dari *filtering*.

Tabel 2. Contoh hasil *filtering*

Teks sebelum <i>filtering</i>	Teks sesudah <i>filtering</i>
banyak tanaman yang mati karena zat zat glifosat tanaman tanaman yang mati meliputi bunga mawar rumput dan tanaman berduri	tanaman mati zat zat glifosat tanaman tanaman mati meliputi bunga mawar rumput tanaman berduri

2.2.3 Stemming

Stemming merupakan proses preprosesing yang berguna untuk menghapus awalan dan juga akhiran pada setiap kata yang ada dalam teks yang menyisakan kata dasar dari kata tersebut dan masih memiliki arti yang sama seperti sebelumnya, sehingga *stemming* masuk ke dalam sistem temu kembali informasi (Maarif, 2015). *Stemming* menjadi cara yang efisien dalam meningkatkan performa pemrosesan teks

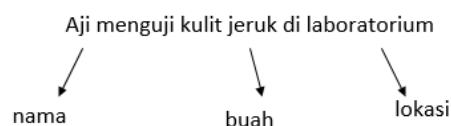
dikarenakan dapat mengubah kata menjadi bentuk kata dasarnya. Tabel 3. adalah contoh dari hasil stemming dari beberapa kata dasar yang memiliki awalan dan akhiran.

Tabel 3. Contoh Stemming

Sebelum stemming	Sesudah stemming
perhitungan	hitung
berduri	duri
menggali	gali
searah	arah
menjepit	jepit
digunakan	guna

2.3 Named entity recognition(NER)

Named entity recognition merupakan turunan dari ekstraksi informasi, yang berguna untuk mengidentifikasi jenis kata (entitas kata) dari setiap kata yang ada dalam sebuah teks. Dalam bahasa Indonesia memiliki susunan kata atau tipe kata, ada banyak tipe kata dalam bahasa Indonesia meliputi kata kerja, kata penghubung, objek, predikat, keterangan tempat dan lainnya, dari banyaknya tipe kata yang paling sederhana untuk menjadi idetentifikasi subjek, predikat objek, dan keterangan (Indurkhyia & Damerau, 2010).



Gambar 1. Contoh entitas

Pada Gambar 1. kata aji yang dulunya subjek menjadi nama, jeruk menjadi buah, dan laboratorium menjadi tempat, oleh karena POS-Tagging inilah digunakan dalam proses named entity recognition untuk memudahkan mencari kategori atau jenis dari kata sehingga memudahkan untuk mencari informasi tertentu.

2.4 Hidden Markov Model (HMM)

Hidden Markov Model (HMM) merupakan metode machine learning yang menggunakan model probabilitas berurutan untuk menyelesaikan masalah dan data latih yang digunakan membutuhkan POS-Tagging untuk setiap kata, dan data latih di modelkan kedalam bentuk model hidden markov model (Todorovic, dkk, 2008).

Dalam *hidden markov model* ada

pemodelan umum yang dipakai dalam pemodelan permasalahan, Persamaan 1. merupakan permodelan umum yang digunakan dalam *hidden markov model*.

$$\lambda = (A, B, \pi) \quad (1)$$

Pada Persamaan 1. terdapat symbol lamda (λ) sebagai *model markov*, A sebagai probabilitas transisi, B adalah probabilitas emisi, dan symbol phi (π) merupakan probabilitas keadaan awal. Dalam penulisan pemodelan sebuah permasalahan dengan Hidden Markov Model menggunakan lima tuple yaitu : (Lin dan Dyer, 2010)

1. *Observed state (O)*

Pada *observed state* di buat dengan simbol $O = O_1, O_2, O_3, \dots O_n$ *observed state* yaitu pemodelan permasalahan yang dapat diamati.

2. *Hidden state (Q)*

Hidden State merupakan state yang tersembunyi dan tidak dapat diamati, di simbolkan dengan $Q = Q_1, Q_2, Q_3, \dots Q_n$.

3. Matrik Peluang Transisi (A)

Peluang transisi merupakan peluang perpindahan dari state i menuju ke state j. Di simbolkan dengan $A = a_{01}, a_{02}, a_{n1}, \dots a_{nm}$; a_{ij} , banyaknya jumlah matrik peluang transisi yaitu $Q \times Q$.

4. Matrik Peluang Emisi (B)

Peluang emisi merupakan peluang perpindahan state i dengan syarat waktu O_t (Observed State). Di simbolkan dengan $B = b_i(O_t)$ banyaknya jumlah matrik peluang emisi yaitu $Q \times O$.

5. Peluang Keadaan Awal (π)

Peluang awal di simbolkan dengan π . Pada *named entity recognition* yaitu jumlah nama entitas kata, jika peluang tiap nama entitas kata dijumlahkan maka akan bernilai satu.

Untuk mengetahui performa dari sistem *named entity recognition* menggunakan *precision (P)*, *recall (R)*, dan *F-measure*. Penghitungan *recall (R)* menggunakan Persamaan 2, *precision (P)* menggunakan Persamaan 3, dan *F-measure* menggunakan Persamaan 4.

$$recall(R) = \frac{TP}{(TP+FN)} \quad (2)$$

$$precision(P) = \frac{TP}{(TP+FP)} \quad (3)$$

$$F - measure(f) = \frac{2PR}{(P+R)} \quad (4)$$

Pada Persamaan 2. untuk persamaan *recall* digunakan untuk menghitung seberapa besar pengenalan entitas yang dapat dilakukan oleh sistem dengan cara melakukan penghitungan antara jumlah pengenalan entitas yang benar dibagi dengan total jumlah entitas dalam teks yang digunakan untuk testing (Todorovic, dkk, 2008). Pada Persamaan 3. *precision* di hitung dengan jumlah pengenalan entitas yang benar oleh sistem NER dibagi dengan semua pengenalan entitas yang dilakukan oleh sistem NER (Todorovic, dkk, 2008).

Pada Persamaan 4. *f-measure* merupakan performa keseluruhan dari sistem *named entity recognition*. Penghitungan *f-measure* dilakukan sesuai Persamaan 4. dengan syarat (P) adalah *precision* dan (R) adalah *recall* (Indurkha, dan Damerau, 2010).

2.5 Algoritma Viterbi

Algoritma Viterbi merupakan algoritma yang digunakan untuk menemukan urutan Hidden State yang paling optimal dari permasalahan yang bisa diamati (Wibisono, 2008). Dalam menemukan urutan Hidden State paling optimal Algoritma Viterbi menggunakan *viterbi trellis*. *Viterbi trellis* yaitu penghitungan secara mundur atau rekrusif. Misalkan *viterbi trellis* pada state j melalui O_t dan Q_t , maka *viterbi trellis* akan di hitung dengan Persamaan 5. *viterbi trellis* pada kasus ini di simbolkan dengan $V_t(i)$.

$$V_t(i) = \max_{1 \leq i \leq N-1} V_{t-1}(i) A_{ij}(O_t) \quad (5)$$

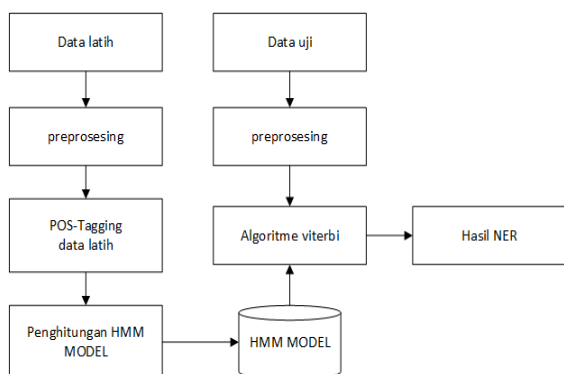
Keterangan dari Persamaan 5. yaitu :

1. $V_{t-1}(i)$, merupakan *viterbi trellis* saat kondisi $t-1$
2. A_{ij} , yaitu peluang transisi dari O_t melalui Q_t
3. (O_t) , merupakan peluang emisi saat O_t pada state t

3. PERANCANGAN SISTEM

Penelitian ini akan menghasilkan program yang dapat mengenali nama entitas pada teks tanaman obat. Nama entitas yang dapat di kenali diantaranya nama tanaman, tempat tanaman, kandungan zat, dan kegunaan tanaman. POS-Tagging pada data latihan untuk masing-masing entitas yaitu NAMA, TMP, ZAT, dan GUNA. Gambar 3.1 merupakan gambaran proses

pengenalan entitas secara umum.



Gambar 2. Deskripsi umum sistem

Gambar 2. adalah gambaran umum tahapan untuk melakukan pengenalan nama entitas. Tahap preprocessing akan dilakukan proses preprocessing sesuai kebutuhan, jadi tidak semua proses preprocessing dilakukan pada tahapan preprocessing. Tahap penghitungan HMM model akan melakukan penghitungan probabilitas awal, probabilitas transisi, dan probabilitas emisi.

3.1 Data penelitian

Data pada penelitian ini di ambil dari tiga website yaitu mitalom.com dan alwib.net sebagai data latih, dan tanamtanaman.com sebagai data uji. Setiap satu teks tanaman obat, untuk 60 data adalah 60 teks tanaman obat, terdiri antara 4 sampai 16 kalimat, setiap kalimat minimal ada 9 kata.

4. PENGUJIAN DAN ANALISIS

4.1 Pengujian data uji sesuai dengan data latih

Tabel 4. Hasil pengujian 1

Entitas	<i>precision</i>	<i>recall</i>	<i>f-measure</i>
NAMA	0.3480	0.85	0.4452
TMP	0.4383	0.8222	0.5361
ZAT	0.5679	0.5555	0.5357
GUNA	0.7446	0.7333	0.7255
rata-rata	0.52474	0.74027	0.56064

Pengujian dilakukan dengan menggunakan data uji sesuai dengan data latih, dimana 30 data uji di random dari 120 data latih yang digunakan dalam sistem.

4.2 Pengujian data uji tidak sesuai dengan data latih

Tabel 5. Hasil pengujian 2

Entitas	<i>precision</i>	<i>recall</i>	<i>f-measure</i>
NAMA	0.303838	0.769444	0.410008
TMP	0.428961	0.636667	0.487662
ZAT	0.459974	0.5	0.456459
GUNA	0.560429	0.755556	0.609421
rata-rata	0.438301	0.665417	0.490887

Pengujian dilakukan dengan menggunakan data uji yang tidak dipakai sebagai data latih, dimana 30 data uji di random dari 36 data yang tersedia.

4.3 Pengujian penambahan pos-tagging

Tabel 6. Hasil pengujian 3

Entitas	<i>precision</i>	<i>recall</i>	<i>f-measure</i>
NAMA	0.367424	0.630556	0.410683
TMP	0.408776	0.677778	0.480964
ZAT	0.447556	0.555556	0.464348
GUNA	0.670459	0.752222	0.687939
rata-rata	0.473554	0.654028	0.510984

Pengujian dilakukan dengan data penambahan tiga pos-tagging pada data latih untuk membuat model HMM. Tiga pos-tagging yang di tambahkan yaitu preposisi, noun, dan verb. Setelah di lakukan pengujian dengan penambahan pos-tagging yang di harapkan dapat menambah akurasi ternyata setelah dilakukan pengujian tidak dapat menambahkan nilai akurasi.

4.4 Pengujian validasi

Pengujian validasi dilakukan untuk mengetahui apakah sistem name entity recognition menghasilkan pengenalan entitas yang valid dengan cara melakukan pengenalan entitas menggunakan 10 data latih dan 10 data uji. Data uji harus sama dengan data latih karena digunakan untuk validasi sistem name entity recognition.

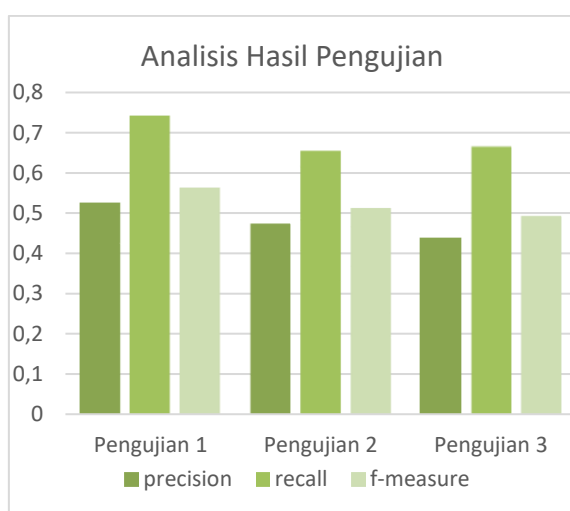
Tabel 6. Hasil pengujian validasi

Entitas	<i>precision</i>	<i>recall</i>	<i>f-masure</i>
NAMA	1.0000	0.4524	0.5800
TMP	0.4607	1.0000	0.5810

ZAT	1.0000	1.0000	1.0000
GUNA	1.0000	1.0000	1.0000
rata-rata	0.8652	0.8631	0.7902

Pengujian validasi menghasilkan nilai *f-measure* paling tinggi 0.9875 pada entitas guna, dan paling rendah 0.49 pada entitas nama. Paling tinggi mencapai 0.9875 ini membuktikan bahwa sistem termasuk pada kategori cukup bagus untuk melakukan pengenalan entitas pada teks tanaman obat.

4.5 Analisis hasil pengujian



Gambar 3. Analisis pengujian

Nilai *f-measure* menunjukkan performance dari keseluruhan sistem. Pada pengujian dengan skenario data latih juga sebagai data uji menghasilkan nilai rata-rata *f-measure* tertinggi dari semua skenario pengujian dengan nilai 0.56064. Tingginya nilai itu dikarenakan data latih juga digunakan sebagai data uji. Pengujian kedua dengan skenario data latih tidak sebagai data uji menghasilkan nilai rata-rata *f-measure* 0.5109.

Pengenalan pada skenario dua sudah sangat bagus di lihat data uji bukan dari data latih, di lihat dari Tabel 6.2 yang menyumbang rata-rata paling banyak yaitu entitas guna, dikarenakan data pada teks tanaman obat entitas kegunaan bisa sama dengan teks tanaman obat lainnya. Pada skenario tiga dengan penambahan tiga pos-tagging yaitu precision, noun, dan verb, yang menghasilkan nilai rata-rata *f-measure* 0.4908.

Pada skenario tiga juga tidak dapat meningkatkan performance sistem dikarenakan semakin banyaknya entitas kata yang digunakan maka pada perhitungan viterbi peluang hasil

pengenalan pada entitas yang tepat semakin berkurang dan penyebab lainnya adalah terlalu banyaknya kata noun dan verb di data latih menyebabkan probabilitas transisi terlalu besar nilainya sehingga berdampak pada pengenalan entitas lainnya.

5. KESIMPULAN

Untuk mengimplementasikan algoritma hidden markov model dan algoritma viterbi dibutuhkan data latih. Data latih yang digunakan harus di beri pos-tagging pada setiap kata yang termasuk dalam entitas yang akan digunakan dalam penelitian. Sebelum pemberian pos-tagging, data latih bisa dilakukan preprosesing (*tokenisasi*, *filtering*, dan *stemming*). Data latih yang sudah dilakukan pemberian pos-tagging akan di proses dan di jadikan model HMM. Model HMM ada tiga yaitu probabilitas awal, probabilitas emisi, dan probabilitas transisi. Model HMM akan digunakan untuk penghitungan pada algoritma viterbi. Algoritma viterbi menghitung probabilitas setiap kata pada data uji, probabilitas tertinggi akan dijadikan sebagai output dari sistem NER pada penelitian ini.

Performance dari sistem NER dapat di hitung dengan menggunakan *precision*, *recall*, dan *f-measure*. Pada penelitian ini nilai tertinggi untuk rata-rata nilai *precision* 0.5447 yang didapat dari skenario pertama dengan data latih sebagai data uji. Nilai rata-rata *recall* tertinggi di dapat pada skenario pertama dengan nilai 0.7402. Nilai *f-measure* tertinggi pada skenario pertama dengan nilai 0.5606. Secara keseluruhan pengenalan entitas nilai yang paling rendah pada skenario ke tiga pada entitas nama dengan *f-measure* 0.4100, dan yang paling tinggi pada skenario pertama pada entitas guna dengan *f-measure* 0.7255.

DAFTAR PUSTAKA

- Agusta, A., 2015. Indonesia Miliki 7500 Tanaman Obat. Lembaga Ilmu Pengetahuan Indonesia (LPII). [Online] Tersedia di : <
<http://lipi.go.id/berita/single/Indonesia-Miliki-7500-Tanaman-Obat/> 11540 >
 [Diakses 13 Febuari 2017]
- Albar, 2016. 100 Jenis Tanaman Obat Keluarga (Toga) dan Manfaatnya Bagi Kesehatan. [Online] Tersedia di : <
<http://alwib.net/tanaman-obat-keluarga/>

- > [Diakses 13 Febuari 2017]
- Alfred, R., Leong, L.C., Chin, K.O., & Anthony, P., 2014. Malay Named Entity Recognition Based on Rule-Based Approach. *International Journal of Machine Learning and Computing*. Vol. 4, No. 3, pp.301-309.
- Azami, 2016. 250 Jenis Tanaman Obat dan Khasiatnya (Page 1). [Online] Tersedia di : < <http://mitalom.com/250-jenis-tanaman-obat-lengkap-dengan-manfaat-dan-khasiatnya-page-1/> > [Diakses 13 Febuari 2017]
- Ilyas, R., & Khodra, M.L., 2015. Ekstraksi Informasi 5W1H pada Berita Online Bahasa Indonesia. *Jurnal Cybermatika*. Tersedia di : < <http://cybermatika.stei.itb.ac.id/ojs/index.php/cybermatika/article/view/98/61> > [Diakses 1 Maret 2017]
- Indurkha, N. dan Damerau, F.J., 2010. *Handbook of Natural Language Prosesing Second Edition*. [e-book] London : CRC Press. Tersedia di : < <https://karczmarczyk.users.greyc.fr/TEACH/TAL/Doc/> > [Diakses 2 Maret 2017]
- Internet Live Stats, 2017. Total number of Websites. [Online] Tersedia di : <<http://www.internetlivestats.com/total-number-of-websites/>> [Diakses 13 Febuari 2017]
- Maarif, A.A., 2015. Penerapan Algoritma TF-IDF Untuk Pencarian Karya Ilmiah.S1. Universitas Dian Nuswantoro. Tersedia di : < <http://eprints.dinus.ac.id/15283/> > [Diakses 1 Maret 2017]
- Rodriguez, 2016. 36 Macam Tanaman Obat Keluarga Beserta Khasiatnya. [Online] Tersedia di : < <http://tanamtanaman.com/tanaman-obat-keluarga/> > [Diakses 13 Febuari 2017]
- Suwarningsih, W., Supriana, I., &Purwarianti, A., 2014. ImNER Indonesian Medical Named Entity Recognition. 2nd International Conference on Technology, Informatics, Management, Engineering & Environment.
- Todorovic, B.T., Rancic, S.R., Markovic, L.M., Mulalic, E.H., dan Ilic, V.M., 2008. Named Entity Recognition and Classification using Context Hidden Markov Model . *Symposium on Neural Network Applications in Electrical Engineering*. NEUREL-2008.
- Winoto, P., & Irianto T.T.J., 2012. Pembuatan Website Profil Sekolah Dasar Negeri 03 Kalisoro. *Journal Speed. Sentra Penelitian Engineering dan Edukasi*. Volume 4. No1. pp.50-55.
- Wibisono, Y., 2008. Penggunaan Hidden Markov Model untuk Kompresi Kalimat. S2. Institut Teknologi Bandung. Tersedia di : < <http://digilib.itb.ac.id/files/disk1/627/jbptitbpp-gdl-yudiwibiso-31314-1-2008ts-r.pdf> > [Diakses 22 Febuari 2017]