Model *Natural Language Processing* untuk Perumusan Keluhan Pasien

Chanifah Indah Ratnasari[1], Sri Kusumadewi[2], Linda Rosita[3]

[1,2]Teknik Informatika Universitas Islam Indonesia
[3]Kedokteran Universitas Islam Indonesia
Jl. Kaliurang km 14 Yogyakarta 55510
Telp (0274) 895287 ext 122, fax (0274) 895007 ext 148
chanifah.indah@uii.ac.id[1], cicie@fti.uii.ac.id[2]

Abstract. Anamnesis atau wawancara medis (*history taking*) merupakan tahap awal dari rangkaian pemeriksaan pasien. Pada umumnya pencatatan anamnesis keluhan pasien yang berupa teks bebas atau narasi medis pada *Electronic Medical Record*sulit untuk dilakukan pemrosesan komputasi, dikarenakan pencatatan sering terdapat kesalahan eja / salah ketik, tata bahasa yang buruk, dan singkatan yang tidak konvensional. Hal ini perlu ditangani dengan cara yang tepat mengingat informasi penting dalam teks bebas tersebut dapat dipergunakan untuk menjalankan berbagai macam pendukung keputusan klinis (*Clinical Decission Support*). Teks masukan yang berupa keluhan pasien merupakan bahasa alami (*natural language*), sehingga agar teks keluhan pasien tersebut dapat dikenali oleh komputer maka dipergunakan pengolahan bahasa alami / *Natural Language Processing* (NLP). Penelitian ini bertujuan untuk membangun sebuah model *Natural Language Processing* untuk perumusan keluhan pasien yang mampu memetakan narasi keluhan pasien ke dalam objek yang tepat, sehingga dapat menghasilkan teks dengan makna yang sama dengan teks narasi keluhan pasien, tetapi dalam bahasa yang lebih baku atau dalam bahasa medis yang tepat.

Keywords:anamnesis, keluhan pasien, *naturallanguageprocessing*

1. Pendahuluan

1.1. Latar Belakang

Dalam dunia medis, istilah anamnesis digunakan untuk pengumpulan informasi mengenai kondisi yang pernah dan sedang dirasakan oleh pasien yang dilakukan oleh dokter untuk tujuan perawatan medis[10]. Anamnesis merupakan tahap awal dari rangkaian pemeriksaan pasien yang bertujuan untuk menegakkan diagnosis[9].

Pencatatan anamnesis tersimpan dalam rekam medis (*medical record*)[11]. Saat ini rekam medis yang umum digunakan masih berbasis kertas. Rekam medis berbasis kertas memiliki kekurangan, yaitu kertas dapatdengan mudahrusak atau hilang seiring berjalannya waktu, memakan ruang, dan sering tidak terbaca[1]. Masalah yang paling serius dengan rekam medis berbasis kertas adalah menghambat dalam pemberian dukungan keputusan klinis, dikarenakan data tersimpan dalam format yang tidak dapat diakses sehingga tidak dapat terhubung atau memicu alat pendukung[2].

Electronic Medical Record (EMR) merupakan lebih dari sekedar sebuah versi elektronik dari *paper-based record* karena menawarkan banyak fungsi, seperti pandangan yang terintegrasi dari data pasien, pendukung keputusan klinis, pemasukan order dokter, dukungan komunikasi yang terintegrasi, dan akses ke sumber pengetahuan. Selain itu EMRjuga dapat diintegrasikan dengan *Computerized Decision Support Systems*(CDSS) diagnosis penyakit.

Agar dapat diolah oleh CDSS, data dalam EMR harus dalam bentuk terstruktur. Akan tetapi pada kenyataannya, pencatatan rekam medis pasien tidak terstruktur, berupa deskripsi tekstual / narasi medis yang bervariasi sesuai dengan dokter yang membuatnya. Terlebih lagi biasanya rekam medis pasien terisi dengan kesalahan tulis/ketik, terjadi duplikasi, ambiguitas, serta penggunaan singkatan yang tidak baku[5].Salah satu pendekatan yang dapat dilakukan untuk mengatasi hal tersebut adalah penggunaan pengolahan bahasa alami / *natural language processing* (NLP) pada pemasukan data teks bebas[13].

1.2. Tujuan Penelitian

Dalam penelitian ini, data teks narasi keluhan pasien dipetakan ke dalam objek riwayat penyakit sekarang yang tepat, sehingga dapat menghasilkan teks dengan makna yang sama dengan teks narasi keluhan pasien, tetapi dalam bahasa yang lebih baku atau dalam bahasa medis yang tepat. Selain itu, hasil pengubahan teks keluhan pasien menjadi bentuk yang lebih terstruktur dapat dijadikan sebagai salah satu bahan yang dapat diintegrasikan dengan sistem pendukung keputusan klinis,sehingga dapat meningkatkan prosedur diagnostik.

2. Tinjauan Pustaka

2.1. Anamnesis

Data anamnesis terdiri atas beberapa kelompok data penting, yaitu identitas pasien, riwayat penyakit sekarang (didahului keluhan utama), riwayat penyakit dahulu, anamnesis sistem, riwayat kesehatan keluarga, dan riwayat pribadi, sosial-ekonomi-budaya[3,9]. Pada kelompok data riwayat penyakit sekarang, berdasarkan jawaban pasien terhadap pertanyaan mengenai keluhan utama, maka jawaban tersebut dikembangkan, digali riwayat penyakit sekarang, mulai dari onset, frekuensi serangan, sifat serangan, durasi, sifat sakit, lokasi, perjalanan penyakitnya, riwayat pengobatan sebelumnya, dan akibat gangguan yang timbul[8].

Menurut (Cesarani et al., 1996)[6], "Penting bahwa pasien menggambarkan gejala yang dialami dengan kata-katanya sendiri dengan cara yang paling sederhana." Pencatatan anamnesis oleh dokter dilakukan pada dokumen rekam medis berupa narasi medis yang tidak terstruktur dan beragam, bervariasi untuk setiap pasien dan dari waktu ke waktu[4].

2.2. *Natural Language Processing* (NLP)

Bahasa alami (*natural language*) adalah bahasa yang biasa digunakan oleh manusia untuk berkomunikasi[12]. Istilah '*Natural Language Processing*' (NLP) biasanya digunakan untuk mendeskripsikan fungsi dari komponen perangkat lunak atau perangkat keras pada sistem komputer yang dapat menganalisis atau menyintesis bahasa alami, baik lisan ataupun tulisan (teks)[7].

3. Pengambilan Data

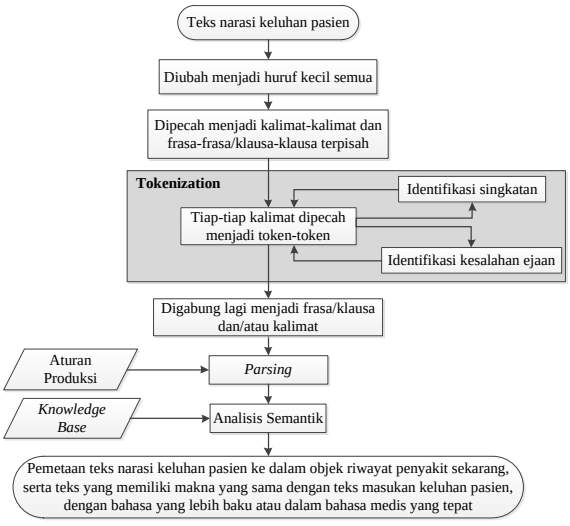
Data yang digunakan dalam penelitian ini adalah data yang diperoleh dari dokter. Data tersebut berupa:

a. Teks narasi keluhan pasien seperti yang biasanya dokter catat pada berkas rekam medis anamnesis keluhan pasien.
b. Pemetaan teks narasi keluhan pasien ke dalam objek-objek riwayat penyakit sekarang dan teks dengan makna yang samadengan bahasa yang lebih baku atau dalam bahasa medis yang tepat.

Data pada *point* (a) akan digunakan sebagai masukan dari prototipe dan data pada *point* (b) akan digunakan sebagai acuan keluaran yang diharapkan.

4. Gambaran Model

Gambaran model dalam penelitian ini ditunjukkan pada Gambar 1.



Gambar 1. Gambaran model

Adapun penjelasan dari Gambar 1 adalah sebagai berikut:

a. Teks Narasi Keluhan Pasien

Teks narasi keluhan pasien dapat berupa satu kalimat atau lebih. Kalimat tersebut dapat terdiri dari satu frasa/klausa atau lebih. Apabila dalam satu kalimat terdapat lebih dari satu frasa/klausa, maka dipisahkan dengan koma. Frasa/klausa ke-1, ke-2, …, ke-*n* merupakan penjelasan dari frasa/klausa ke-0 dalam kalimat yang sama.

b. Pengubahan menjadi Huruf Kecil Semua

Masukan berupa teks narasi keluhan pasien diubah menjadi huruf kecil semua untuk nantinya mempermudah dalam proses-proses selanjutnya.

c. Pemecahan menjadi Kalimat-kalimat dan Frasa-frasa/Klausa-klausa Terpisah

Pemecahan kalimat dilakukan berdasarkan karakter titik ataupun *new line*. Sedangkan pemecahan frasa/klausa dilakukan berdasarkan karakter koma.

d. Tokenization

Kalimat-kalimat dan frasa-frasa/klausa-klausayang telah diperoleh dari tahap sebelumnya, dilakukan pemecahan menjadi token-token (*tokenization*), dengan menggunakan karakter spasi atau tanda baca sebagai pemisah. Dalam tahap ini dilakukan *exception*apabila terdapat karakter titik, koma, atau garis miring yang diapit oleh karakter angka, dalam hal ini tidak dilakukan pemecahan token. Dari token-token yang diperoleh, dilakukan identifikasi singkatan, mengingat pencatatan anamnesis keluhan pasien tidak terlepas dari penggunaan singkatan yang tidak baku. Sebagai contoh, kata "sll" untuk "selalu". Pendekatan yang digunakan untuk identifikasi singkatan tersebut adalah menggunakan leksikon / kamus. Setelah itu dilakukan identifikasi kesalahan ejaan. Setiap token akan dicari pada leksikon, jika tidak ditemukan akan dianggap sebagai kata yang salah eja, sehingga akan dilakukan *spelling correction*. Pendekatan yang digunakan adalah *edit distance*, yaitu dengan melakukan perhitungan kemiripan antara token yang tidak ditemukan dalam leksikon tadi dengan kata-kata yang terdapat dalam leksikon.

e. Penggabungan Kembali menjadi Frase dan/atau Kalimat

Token yang sudah dilakukan identifikasi singkatan dan bebas dari kesalahan eja akan digabung kembali menjadi frasa/klausa dan/atau kalimat terpisah.

f. Parsing

Berdasarkan data yang diperoleh dari dokter, dibuat aturan produksi untuk memetakan kalimat masukan keluhan pasien ke dalam objek-objek riwayat penyakit sekarang, yang nantinya dapat membentuk bahasa baku/bahasa medis yang tepat dari serangkaian pemetaan objek yang terbentuk. Aturan produksi dalam penelitian ini adalah sebagai berikut:

S → Kalimat
Kalimat → Frasa/Klausa
Kalimat → Keluhan Utama | <kolom_lain>

Keluhan Utama → Keluhan Utama +<kolom_lain>
<kolom_lain> → <kolom_lain> + <kolom_lain>
Frasa/Klausa → <kolom_lain>
<kolom_lain> → Onset | Keluhan Lain | Keterangan | Frekuensi Serangan | Sifat Serangan | Durasi | Sifat Sakit | Lokasi | Perjalanan Penyakit | Riwayat Pengobatan Sebelumnya | Akibat Gangguan yang Timbul

g. Analisis Semantik

Berdasarkan pemetaan yang diperoleh dari *parsing*, dilakukan pencarian makna dari kombinasi pemetaan objek dari masing-masing kalimat teks masukanpada *knowledge base*. Pendekatan yang digunakan pada tahap ini adalah dengan menggunakan lekskon.

5. Algoritma Pencarian Kandidat Kecocokan

Masing-masing frasa/klausa dalam masing-masing kalimat dilakukan pencarian kecocokan ke dalam leksikon. Dalam penelitian ini, digunakan *treshold* sebagai nilai minimal kecocokan. Nilai kecocokan leksikon yang melebihi *treshold* dan tertinggi (maksimum) terpilih sebagai kandidat leksikon terpilih. Adapun algoritma pencarian kandidat kecocokan dalam penelitian ini adalah sebagai berikut.

a. Frasa/klausa ke-0 dalam kalimat dilakukan perhitungan kecocokan ke dalam leksikon keluhan utama.
b. Apabila tidak terdapat nilai yang memenuhi *treshold*, maka dilakukan perhitungan kecocokan pada leksikon kolom lain (selain keluhan utama).
c. Apabila dari frasa/klausa ke-0 kalimat yang dihitung kecocokannya tadi terdapat token sisa (token lebihan yang tidak cocok dengan kandidat leksikon terpilih), maka token sisa tersebut dilakukan perhitungan kecocokan dengan kolom lain (selain keluhan utama) dari kandidat leksikon terpilih.
d. Apabila dalam kalimatyang dihitung kecocokannya terdapat lebih dari satu frasa/klausa, maka frasa/klausa selanjutnya dilakukan perhitungan kecocokan pada kolom lain (selain keluhan utama) dari kandidat leksikon terpilih.
e. Apabila nilai kecocokan kandidat leksikon terpilih dari frasa/klausa ke-0 lebih dari satu (terdapat nilai maksimumyang samalebih dari *treshold*lebih dari satu), maka akan dilakukan perhitungan rata-rata kecocokan dari masing-masing kandidat leksikon terpilih. Nilai rata-rata yang tertinggi yang menjadi kandidat pemenang.

6. Perhitungan Kecocokan

Rumus yang digunakan untuk pencocokan data yang terdapat pada leksikon dengan kalimat masukan keluhan pasien ditunjukkan pada Persamaan (1).

$$T_i = \frac{nX_i}{\max(nL_i, nt)} \qquad (1)$$

dengan:

T_i = nilai kecocokan dengan leksikon ke-i.
nX_i = banyaknya kesamaan token leksikon ke-i dengan token kalimat atau frasa/klausa masukan.
nL_i = banyaknya token pada leksikon ke-i.
nt = banyaknya token pada kalimat atau frasa/klausa masukan.

Apabila pada suatu kalimat masukan terdapat lebih dari satu frasa/klausa atau frasa/klausa ke-0nya memiliki token sisa, dan frasa/klausa ke-0nya memiliki nilai kecocokan lebih dari satu yang melebihi *treshold* dan maksimum, maka dilakukan perhitungan rata-rata kecocokan seperti yang ditunjukkan pada Persamaan (2).

$$R_i = \frac{pF_0 + pS_0 + pF_1 + ... + pF_m}{\max(nC_1, nC_2, ..., nC_m)} \quad (2)$$

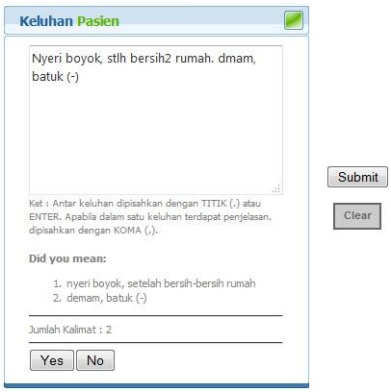
dengan:

R_i = nilai rata-rata kecocokanleksikon ke-i.
pF_0 = persentase kecocokan frasa/klausa ke-0.
pS_0 = persentase kecocokan token sisa dari frasa/klausa ke-0.
pF_1 = persentase kecocokan frasa/klausa ke-1.
pF_m = persentase kecocokan frasa/klausa ke-m.
nC_1 = banyaknya persentase kecocokan pada leksikon yang memiliki nilai kecocokan sama yang pertama.
nC_2 = banyaknya persentase kecocokan pada leksikon yang memiliki nilai kecocokan sama yang kedua.
nC_m = banyaknya persentase kecocokan pada leksikon yang memiliki nilai kecocokan sama yang ke-m.

Nilai rata-rata kecocokan yang tertinggi dipilih sebagai kandidat terpilih.

7. Implementasi dan Pengujian

Antarmuka (*interface*) dari prototipe penelitian ini berupa antarmuka berbasis *web*. Pengujian prototipe akan dilakukan oleh dokter selaku pihak yang akan menjadi pengguna prototipe ini nantinya. Pengujian dilakukan dengan memasukkan teks narasi keluhan pasien. Sebagai contoh apabila dokter memasukkan keluhan pasien "Nyeri boyok, stlh bersih2 rumah. dmam, batuk (-)", maka apabila terdapat kesalahan eja dalam teks masukan, prototipe akan memberikan usulan kandidat pembenaran kata (Gambar 2). Pada kondisi ini, dokter dapat memilih usulan yang dihasilkan oleh prototipe, atau tetap menggunakan masukan awal sebagai masukan yang valid untuk pemrosesan selanjutnya. Selanjutnya akan tampil keluaran berupa bahasa medis / bahasa yang lebih baku dari teks masukan, yaitu "Low back pain (+)" untuk kalimat masukan pertama dan "Febris" untuk kalimat masukan kedua, serta pemetaannya ke dalam objek-objek riwayat penyakit sekarang (Gambar 3).



Gambar 2. Usulan kandidat pembenaran kata dari masukan yang terdapat kesalahan eja

Riwayat Penyakit Sekarang	Kalimat ke 0	Kalimat ke 1
Keluhan Utama	nyeri boyok	demam
Onset		
Keluhan Lain		batuk (-)
Keterangan	setelah bersih-bersih rumah	
Frekuensi Serangan		
Sifat Serangan		
Durasi		
Sifat Sakit		
Lokasi		
Perjalanan Penyakit		
Riwayat Pengobatan Sebelumnya		
Akibat Gangguan yang Timbul		
Bahasa Medis	Low back pain (+)	Febris

Gambar 3. Keluaran berupa bahasa medis / bahasa yang lebih baku dari teks masukan dan pemetaannya ke dalam objek-objek riwayat penyakit sekarang

8. Kesimpulan

Kesimpulan yang dapat diambil dari penelitian yang dilakukan adalah sebagai berkut:

a. *Natural Language Processing*dapat diterapkan untuk menangani teks masukan keluhan pasien yang berupa teks bebas / narasi medis yang tidak terstruktur.
b. Hasil dari pengubahan teks keluhan pasien menjadi bentuk yang lebih terstruktur dapat dijadikan sebagai salah satu bahan yang dapat diintegrasikan dengan sistem pendukung keputusan klinis.

Pustaka

1. Ajami, S., & Bagheri-Tadi, T. (2013). Barriers for Adopting Electronic Health Records (EHRs) by Physicians. *Acta Informatica Medica*, *21*(2), 129–134.
2. Bates, D. W., Ebell, M., Gotlieb, E., Zapp, J., & Mullins, H. C. (2003). A Proposal for Electronic Medical Records in U.S. Primary Care. *Journal of the American Medical Informatics Association*, *10*(1), 1–10. BMJ Publishing Group Ltd.

3. Bickley, L., & Szilagyi, P. G. (2013). *Bates' Guide to Physical Examination and History-Taking -Eleventh Edition* (Vol. 2012). Lippincott Williams & Wilkins.
4. Bleeker, S. E., Derksen-Lubsen, G., van Ginneken, A. M., van der Lei, J., & Moll, H. A. (2006). Structured data entry for narrative data in a broad specialty: patient history and physical examination in pediatrics. *BMC medical informatics and decision making*, *6*(1), 29.
5. Bursa, M., Lhotska, L., Chudacek, V., Spilka, J., Janku, P., & Huser, M. (2013). Effective Free-Text Medical Record Processing and Information Retrieval. In M. Long (Ed.), *International Federation for Medical and Biological Engineering (IFMBE) Proceedings: World Congress on Medical Physics and Biomedical Engineering May 26-31, 2012, Beijing, China*, IFMBE Proceedings (Vol. 39, pp. 1305–1308). Berlin, Heidelberg: Springer Berlin Heidelberg.
6. Cesarani, A., Alpini, D., & Brambilla, D. (1996). Anamnesis and Clinical Evaluation. In A. Cesarani, D. Alpini, R. Boniver, C. F. Claussen, P. M. Gagey, L. Magnusson, & L. M. Ödkvist (Eds.), *Whiplash Injuries* (pp 75–81). Milano: Springer Milan.
7. Jackson, P., & Moulinier, I. (2007). *Natural Language Processing for Online Applications : Text Retrieval, Extraction and Categorization - Second Revised Edition*. John Benjamins Publishing Company.
8. Kedokteran UII. (2014). *Panduan Keterampilan Medik Blok Sistem Pertahanan Tubuh dan Penyakit Infeksi (KBK 2005)* (p. 10). Yogyakarta.
9. Markum, H. M. S., & Widodo, D. (2000). *Penuntun Anamnesis dan Pemeriksaan Fisis*. (H. M. S. Markum, Ed.) (pp. 11–24). Pusat Penerbitan Departemen Ilmu Penyakit Dalam Fakultas Kedokteran Universitas Indonesia.
10. Moehr, J. R., & Hannover. (1977). Computer Assisted Medical History. In P. L. Reichertz & G. Goos (Eds.), *Informatics and Medicine*, Medizinische Informatik und Statistik (3rd ed., Vol. 3, pp. 460–578). Berlin, Heidelberg: Springer Berlin Heidelberg.
11. PERMENKES. (2008). *Peraturan Menteri Kesehatan Republik Indonesia Nomor 269/MENKES/PER/III/2008 Tentang Rekam Medis Pasal 3*. Indonesia.
12. Purnama, I. K. E., & Zaini, A. (2009). *Pengembangan Agent Antarmuka Cerdas Berbasis Bahasa Alami untuk Bahasa Indonesia yang Diterapkan Pada Game Edukasi Kecakapan Hidup (Life Skill)*. ITS Digital Repository.Diakses pada 10Oktober 2013 dari http://digilib.its.ac.id/public/ITS-Research-11468-132137894-Conclusion.pdf
13. Tange, H. J., Hasman, A., Robbé, P. F. de V., & Schouten, H. C. (1997). Medical narratives in electronic medical records. *International journal of medical informatics*, *46*(1), 7–29. Elsevier.