LITERATURE REVIEW NAMED ENTITY RECOGNITION IN BIOMEDICAL DALAM APLIKASI BASIS DATA

Guntur Eka Saputra

Teknik Informatika
Jl. Margonda Raya No. 100 Pondok Cina Depok
Email: guntur@staff.gunadarma.ac.id

ABSTRAK

Data merupakan kumpulan dari datum (data umum) yang terdiri dari material mentah dari informasi yang ingin dihasilkan. Pengguna membutuhkan data yang sudah menjadi sebuah informasi yang berguna atau bermakna untuk dapat dijadikan suatu referensi atau acuan atau dasar tertentu untuk mengambil suatu keputusan. Informasi yang dibutuhkan untuk menjadi sesuatu yang memiliki makna harus dapat dikenali oleh komputer dengan sebuah prosedur atau metode. Named Entity Recognition (NER) dikenal sebagai identifikasi entitas, chunking entity, dan ekstraksi entitas) merupakan sub-tugas ekstraksi informasi yang berupaya untuk mencari dan mengklasifikasikan entitas yang disebutkan dalam teks ke dalam kategori yang telah ditentukan sebelumnya, seperti nama orang, organisasi, lokasi, ekspresi waktu, jumlah nilai moneter, presentase. Salah satu informasi yang dibutuhkan untuk dapat dikenali makna tersbeut adalah biomedis. Misalnya NER dapat mengenali bahwa "kanker pancreas" adalah penyakit. Teks biomedis bukan termasuk homogen. Catatan atau record medis ditulis berbeda dari artikel ilmiah, anotasi urutan, atau pedoman kesehatan masyarakat lainnya. Berbagai metode dalam biomedis untuk dapat dikenali maknanya sudah banyak dikembangkan dengan menggunakan machine learning. Penelitian ini membahas penelitian yang sudah dilakukan sebelumnya mengenai NER dalam bidang biomedis.

Kata Kunci: Data, Informasi, NER, Makna, Biomedis, Machine Learning.

ABSTRACT

Data is a collection of datum (general data) which consists of raw material from the information to be generated. Users need data that has become useful or meaningful information to be used as a reference or reference or a certain basis for making a decision. The information required to become meaningful must be recognized by the computer by a procedure or method. Named Entity Recognition (NER) is known as entity identification, entity chunking, and entity extraction) is an information extraction sub-task that seeks to locate and classify entities mentioned in the text into predetermined categories, such as names of people, organizations, locations, time expression, total monetary value, percentage. One of the information needed to recognize the meaning is biomedicine. For example NER can recognize that "pancreatic cancer" is a disease. Biomedical texts are not homogeneous. Medical records or records are written differently from scientific articles, sequence annotations, or other public health guidelines. Various methods in biomedicine to identify its meaning have been developed using machine learning. This study discusses previous research on NER in the biomedical sector.

Keywords: Data, Information, NER, Meaning, Biomedical, Machine Learning.

PENDAHULUAN

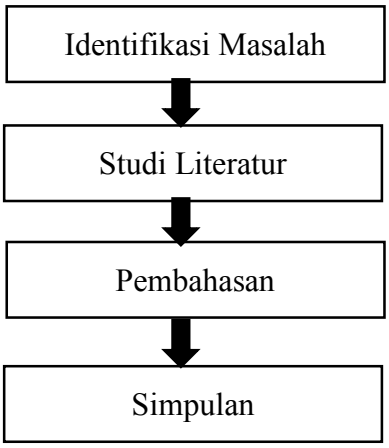
Informasi diperoleh dari kumpulan atau koleksi data. Data merupakan kumpulan dari datum (data umum) yang terdiri dari material mentah dari informasi yang ingin dihasilkan. Data juga dapat berupa bahasa, matematis, dan pengganti simbolik lainnya yang secara umum disepakati untuk menyatakan orang, objek, periwtiwa, kejadian, dan atau konsep yang belum memiliki makna tertentu untuk menjadi sebuah informasi. Manusia atau *user* atau pengguna membutuhkan data yang sudah menjadi sebuah informasi yang berguna atau bermakna untuk dapat dijadikan suatu referensi atau acuan atau dasar tertentu untuk mengambil suatu keputusan, sehingga informasi dapat dikatakan sebagai data yang ditempatkan ke dalam suatu konteks bermakna bagi pengguna atau penerimanya (Mulyati, 2005).

Informasi yang dibutuhkan untuk menjadi sesuatu yang memiliki makna harus dapat dikenali oleh komputer dengan sebuah prosedur atau metode. Informasi dapat diekstraksi yang dikenal dengan *Named Entity Recognition* (NER). *Named Entity Recognition* (NER) dikenal sebagai identifikasi entitas, *chunking entity*, dan ekstraksi entitas) merupakan sub-tugas ekstraksi informasi yang berupaya untuk mencari dan mengklasifikasikan entitas yang disebutkan dalam teks ke dalam kategori yang telah ditentukan sebelumnya, seperti nama orang, organisasi, lokasi, ekspresi waktu, jumlah nilai moneter, presentase, dll. Sistem NER telah dibuat menggunakan teknik berbasis tata bahasa lingusitik, serta model statistic seperti pembelajaran mesin *(machine learning*) (Gupta, 2018).

Named Entity Recognition (NER) merupakan langkah pertama menuju ekstraksi pertama menuju ekstraksi informasi yang berupaya mencari dan mengklasifikasikan entitas yang disebutkan dalam teks ke dalam kategori yang telah ditentukan sebelumnya, sepetti nama orang, organisasi, lokasi, ekspresi, waktu, jumlah, nilai moneter, presentase, dll. NER juga digunakan dalam banyak bidang dalam *Natural Language Processing* (NLP) (Li, 2018). NLP menempatkan entitas dalam teks yang tidak terstruktur atau semi-terstruktur. Entitas-entitas ini dapat berupa berbagai hal dari seseorang hingga sesuatu yang sangat spesifik, seperti istilah biomedis. Misalnya NER dapat mengenali bahwa “kanker pancreas” adalah penyakit (Innoplexus, 2019). Teks biomedis bukan termasuk homogen. Catatan atau *record* medis ditulis berbeda dari artikel ilmiah, anotasi urutan, atau pedoman kesehatan masyarakat lainnya. Selain itu, terdapat dialek local yang tidak jarang diketahui (Rodriguez-Esteban, 2009). Berdasarkan hal inilah, dibutuhkan ekstraksi informasi untuk dapat diketahui makna dari suatu kata atau kalimat yang dapat digunakan oleh penerimanya.

METODE PENELITIAN

Dalam penulisan artikel ilmiah ini diperlukan kerangka kerja dalam melaksanakan penelitian untuk melakukan *literature review* mengenai ekstraksi informasi dalam bidang biomdedial, seperti pada gambar 1 metode penelitian.



Gambar 1. Metode Penelitian

PEMBAHASAN

Pada penulisan ini dilakukan juga kajian ilmu dengan cara *literature review* terhadap enam penelitian yang telah dilakukan mengenai *Named Entity Recognition* (NER) dalam Biomedical.

(Collier et al, 2000) menggunakan *hidden markov model* (HMM) telah dibuktikan bahwa nilai untuk berbagai tugas dalam ekstraksi awal dan hasilnya menunjukkan bahwa kinerja yang baik ini dapat dicapai diseluruh domain, yaitu dalam bidang molekuler biologi serta meningkatnya laporan-laporan surat kabar. Pada penelitian ini sejumlah besar istilah yang bukan kata benda yang tepat, seperti yang terdapat pada sub-kelas sumber serta adanya tumpeng tindih secara leksikal yang besar antara kelas-kelas seperti, PROTEIN dan DNA. Model; ini masih memiliki keterbatasan, seperti ketika dibutuhkan identitas batas istilah untuk frase yang mengandung struktur local yang berpotensi adanya makna ambigu, dalam kasus ini dibutuhkan dengan menambahkan aturan *postprocessing*.

(Lee et al, 2004) mengatakan bahwa penelitian ini telah mengusulkan metode baru bernama biomedis dua fase (*two-phase biomedical*) berdasarkan SVM. Pada fase pertama, didentifikasi batas-batas setiap entitas dengan satu klasifikasi SVM dan pasca-proses dengan pencarian kamus sederhana untuk memperbaiki kesalahan SVM. Pada fase kedua, diklasifikasikan entitas yang diidentifikasi ke dalam kelas semantiknya dengan menggunakan SVM hirarkis. Penelitian ini dengan membagi tugas menjadi dua sub-tugas, pengenalan, dan

klasifikasi semantic, dapat memilih fitur yang lebih relevan untuk setiap tugas dan mengadopsi metode klasifikasi yang sesuai dengan tugas tersebut. Proses dua fase ini menghasilkan pengurangan biaya pelatihan SVM dan pengurangan masalah distribusi kelas yang tidak seimbang.

(Saha et al, 2009) mengatakan bahwa NER biomedical teks adalah tugas yang kompleks. Penelitian ini mempelajari pendekatan pembelajaran mesin berbasis MaxEnt. Kinerja pada pendeketan ini tergantung pada kesesuaian fitur. Penelitian ini juga ditunjukkan bahwa penggunaan teknik pengurangan dimensionalitas dapat meningkatkan kinerja secara substansial. Dua pendekatan untuk pengurangan dimensi, yaitu pemilihan kata/afiks informative, dan pengelompokkan kata, afiks. Penelitian ini tekah memberikan kinerja yang lebih baik daripada NER biomedis yang ada yang tidak menggunakan pengetahuan domain yang mendalam.

(Goulart et all, 2011) mengemukakan bahwa kebanyakan studi ilmiah dalam NER Biomedical (92.1% diantara referensi yang dipelajari) sedang dikembangkan melalui karya eksperimental yang didasarkan pada teknik korpora dan pembelajaran mesin *(machine learning*). Pendekatan lain hanya mewakili 7,89% dari studi ini pada peiode 2007-2009. Ini tidak berbeda dari penelitian sebelumnya yang telah disajikan.

(Alshaikhdeeb & Ahmad, 2016) mengemukakan bahwa penelitian ini telah memberikan tinjauan luas mengenai fitur-fitur BNER di mana taksonomi telah diidentifikasi berdasarkan pada representasi (yaitu nominal, numerik, dan Boolean). Selain itu, diskusi telah dilakukan untuk memberikan analisis kritikal untuk setiap fitur. Fitur Boolean morfologis menunjukkan keunggulan dibandingkan dengan fitur lainnya. Membangun studi banding menggunakan fitur-fitur ini akan menjadi peluang besar untuk masa depan penelitian dalam hal mengidentifikasi kinerja untuk mengekstraksi entitas biomedis. Penelitian ini bertujuan untuk mengakomodasi studi tinjauan pada fitur yang dapat digunakan untuk BNER dimana berbagai jenis fitur akan diperiksa, termasuk fitur morfologi, fitur berbasis kamus, fitur leksikal, dan fitur berbasis jarak.

Tabel 1. Rangkuman Hasil Kajian Ilmu (*Literature Review*)

Authors	**Judul**	**Metode**	**Hasil**
Collier, N., Nobata, C. Tsujii, Jun-Ichi	*Extracting the names of genes and gene products with a hidden Markov model*	*Linear interpolating Hidden Markov Model (HMM)*	Penelitian ini dihasilkan untuk ekstraksi terminology teknikal dari abstrak MEDLINE dan teks domain dari molecular-biologi. Tahap pertama dilakukan memperbaharui basis dat abologi, kemudian ditraning dengan HMM berdasarkan leksikal dan karakter fitur-fitur dalam corpus kecil yaitu 100

			MEDLINE subtracts. Penelitain ini dihasilkan dengan mencapai nilai F-Score sebesar 0.73.
Lee, Ki-Joong, et al	*Biomedical named entity recognition using two-phase model based on SVMs*	*Two-phase SVM, identification phase, and classification phase*	NER telah menjadi salah satu tugas paling mendasar dalam akuisisi pengetahuan biomedis. Dalam penelitian ini, menyajikan dua fase SVM. Hasil percobaan pada GENIA corpus menunjukkan bahwa metode yang diusulkan efektif tidak hanya dalam mengurangi biaya komputasi tetapi juga dalam meningktakan kinerja. Skor-F (β = 1) untuk identifikasi batas adalah 74.8 dan skor-F untuk klasifikasi semantic adalah 66,7.
Saha, S. K., Sarkar, S. Mitra, P.	*Feature selection techniques for maximum entropy based biomedical named entity recognition*	*MaxEnt based machine learning*	Penelitian ini dihasilkan studi mengenai pengelompokkan kata dan pendekatan pengurangan fitur berbasis seleksi untuk pengenalan entitas yang diberi nama dengan menggunakan penggolong entropi maksimum. Identifikasi dan pemilihan fitur sebagian besar dilakukan secara otomatis tanpa menggunakan pengetahuan domain. Kinerja sistem ditemukan lebih unggul dari sistem yang ada, yang tidak menggunakan pengetahuan domain.
Goulart, R. R. V. Lima, V. L. S. d., Xavier, C.	*A systematic review of named entity recognition in biomedical texts*	*Kopora, NER, Machine Learning*	Hasil penelitian ini mengidentifikasi metode utama dalam NER biomedit, fitur dan metodologi untuk implementasi sistem NER. Selain dari kecenderungan yang teridentifikasi, beberapa kesenjangan terdeteksi yang mungkin merupakan peluang studi baru di area tersebut.
Alshaikhdeeb, B. Ahmad, K.	*Biomedical Named Entity Recognition: A Review*	*BNERR Features (Morphological, Lexical, Distance-based, Dictionary-based)*	*Biomedical Named Entity Recognition* (BNER) adalah tugas untuk mengidentifikasi contoh biomedis seperti senyawa kimia, gen, protein, virus, gangguan, DNA, dan RNA. Tantangan utama dibali BNER terletak pada metode yang akan digunakan mengekstraksi entitas tersebut. Sebagian besar metode yang digunakan untuk BNER mengandalkan *Supervised machine Learning* (SML).

PENUTUP

Named Entity Recognition (NER) dan klasifikasi adalah tugas mengidentifikasi teks makna khusus dan mengklasifikasikan ke dalam beberapa kategori yang telah ditentukan. Beberapa kategori-kategori seperti orang, lokasi, orgnasisasi, jumlah, dll. NER entitas yang dinamai ini telah menjadi area penelitian yang penting sejak tahun 1996. Salah satu penelitian yang dikembangkan adalah seperti yang dibahas pada kajian ilmu yaitu Biomedical Texts. Dalam metode NER yang menggunakan *Natural Language Processing* (NLP) juga dapat menggunakan metode *Conditional Random Field, Hidden Markov Model, Maximum Entropy,*

Support Vector Machine, Machine Learning, dll. Hal ini menunjukkan bahwa studi dilakukan untuk mencari akurasi, presisi, nilai F1-Score untuk mencapai nilai optimal dalam ekstraksi informasi, sehingga dapat digunakan oleh pengguna dan penerimanya.

DAFTAR PUSTAKA

Alshaikhdeeb, B. Ahmad, K. 2016. *Biomedical Named Entity Recognition: A Review*. International Journal on Advanced Science Engineering and Information Technology. Vol. 6, No. 6. Pp. 889-895. ISSN. 2088-5334

Collier, N., Nobata, C. Tsujii, Jun-Ichi. *Extracting the names of genes and gene products with a hidden Markov model*. COLING '00: Proceedings of the 18th conference on Computational linguistics – Volume 1, July, pp. 201-2017, https://doi.org/10.3115/990820.990850

Goulart, R. R. V. Lima, V. L. S. d., Xavier, C. 2011. *A systematic review of named entity recognition in biomedical texts.* Journal of the Brazilian Computer Society 17 (2) pp. 103-116. DOI: 10.1007/s13173-011-0031-9

Gupta, M. 2018. *A Review of Named Entity Recognition (NER) Using Automatic Summarization of Resumes*. Available at: https://towardsdatascience.com/a-review-of-named-entity-recognition-ner-using-automatic-summarization-of-resumes-5248a75de175

Innoplexus. 2019. *What is Named Entity Recognition?*. Available at: https://medium.com/@Innoplexus/what-is-named-entity-recognition-7ed05beb7171

Lee, Ki-Joong, et al. 2004. *Biomedical named entity recognition using two-phase model based on SVMs.* Journal of Biomedical Informatics Volume 37, Issue 6, Dec, pp. 436-447, https://doi.org/10.1016/j.jbi.2004.08.012.

Li, S. 2018. *Named Entity Recognition with NLTK and SpaCy.* Available at: https://towardsdatascience.com/named-entity-recognition-with-nltk-and-spacy-8c4a7d88e7da

Mulyati, Y.S. 2005. *Konsep Sistem Informasi.* Jurnal Administrasi Pendidikan Vol 3. No. 1. ISSN: p.1412-8152 e.2580-1007.

Rodriguez-Esteban, R. 2009. *Biomedical Text Mining and Its Applications.* PLoS Comput Biol 5(12): e1000597. https://doi.org/10.1371/journal/pcbi.1000597.

Saha, S. K., Sarkar, S. Mitra, P. 2009. *Feature selection techniques for maximum entropy based biomedical named entity recognition*. Journal of Biomedical Informatics Volume 42, Issue 5, Oct, pp. 905-911. https://doi.org/10.1016/j.jbi.2008.12.012