Tinjauan Literatur : Named Entity Recognition pada Ulasan Wisata

by Muhammad Fakhri Despawida Aulia Putra

Submission date: 24-Nov-2020 10:13PM (UTC+0700)
Submission ID: 1455061199
File name: njauan_Literatur_Named_Entity_Recognition_pada_Ulasan_Wisata.pdf (430.52K)
Word count: 2284
Character count: 14248

Tinjauan Literatur : *Named Entity Recognition* pada Ulasan Wisata

***Abstraksi*— Indonesia memiliki banyak sekali tempat wisata, membuat orang-orang antusias untuk mengunjunginya. Beberapa orang memilih untuk melihat ulasan wisata terlebih dahulu sebagai pertimbangan. Tetapi banyaknya tempat wisata membuat orang-orang kesulitan menemukan informasi yang diinginkan dari ulasan tersebut. *Named Entity Recognition* (NER) berguna untuk mengekstraksi informasi pada sebuah teks sehingga dapat lebih memudahkan orang-orang untuk mengetahui informasi yang terdapat pada suatu teks maupun dokumen. NER dapat digunakan untuk mempermudah menemukan informasi yang diinginkan seperti nama wisata, nama lokasi dan fasilitas. Saat ini NER sudah diterapkan pada biomedis, berita, medis, *twitter* dan *tourism*. Dalam melakukan NER metode LSTM, BiLSTM, CNN maupun CRF biasa diterapkan pada NER. Hasil penelitian ini diharapkan bisa digunakan untuk membantu mengembangkan penelitian NER selanjutnya. Tinjauan literatur ini dibuat untuk mengkaji literatur sebelumnya tentang NER dengan *deep learning* dan NER pada pada bidang *tourism* sehingga dapat membantu pengembangan penerapan NER selanjutnya**

Keywords—Named Entity Recognition, deep learning, ekstraksi fitur

I. PENDAHULUAN

Indonesia adalah negara yang salah satunya dipenuhi dengan banyak sekali tempat wisata. Dengan banyaknya tempat wisata yang terdapat di Indonesia ini membuat banyak orang antusias mengunjungi tempat tersebut. Akan tetapi denganbanyaknya tempat wisata, sering kali membuat orang-orang kebingungan untuk memilih destinasi wisata. Dikutip dari halaman *website* TripAdvisor, 83% pengguna global TripAdvisor biasanya selalu melakukan *review* pada ulasan-ulasan di TripAdvisor sebelum membuat keputusan. Pengguna juga setuju TripAdvisor membantu dalam merencanakan perjalanan wisata dengan baik dan mereka memilih untuk mempertimbangkan serta membaca ulasan untuk merencanakan perjalanan wisata [1]. Akan tetapi dengan banyaknya ulasan yang ada dapat membuat orang-orang kesulitan dalam menemukan informasi yang diinginkan.

Named Entity Recognition (NER) dapat berguna untuk mengekstraksi informasi pada sebuah teks dengan mengidentifikasi dan mengenali entitas yang ada [2]. *Named Entity Recognition* dapat digunakan untuk membantu pengguna dalam mengetahui informasi penting yang dibutuhkan. Dalam masalah ini, informasi yang dibutuhkan seperti nama wisata, nama lokasi dan fasilitas, sehingga dapat membantu pengguna dalam mendapatkan informasi tersebut.

Oleh karena itu, tinjauan pustaka ini ditulis bertujuan untuk melakukan komparasi metode pada NER, seperti metode apa saja yang sudah digunakan pada penelitian sebelumnya dan bagaimana hasilnya. Selain itu, bertujuan juga untuk melihat tren penerapan dari NER, untuk bidang apa saja NER diterapkan. Tinjauan literatur ini juga dilakukan untuk mengetahui sejauh mana penelitian NER dilakukan dalam bidang tourism. Penelitian ini diharapkan dapat memiliki kontribusi membantu peneliti dalam menentukan metode yang tepat untuk NER dan mengetahui tren penelitian dalam bidang NER.

.

II. METODOLOGI PENELITIAN

A. *Pertanyaa Penelitian*

Langkah pertama adalah mengidentifikasi pertanyaan penelitian. Pertanyaan yang dilakukan harus ringkas dan jelas. Pada penelitian ini, pertanyaan penelitiannya adalah sebagai berikut :

(i) Apa penelitian terbaru yang berkaitan dengan NER?

(ii) Apa metode yang digunakan dalam NER?

(iii) Bagaimana penelitian NER di bidang *tourism* pada penelitian sebelumnya?

B. *Pengumpulan Data*

Pengumpulan data mencari literatur penelitian sebelumnya dilakukan menggunakan *Google Scholar*, *Medwell, Elsevier*, *Research Gate* dan *arXiv*. Dalam tahap ini literatur yang dikumpulkan merupakan penelitian NER dalam bidang umum yang menggunakan *deep learning* dan juga penelitian NER pada bidang *tourism*. Tinjauan literatur ini menganalisis literatur mulai dari tahun 2016 hingga 2020. Pengumpulan data dalam bidang umum ini menggunakan kata kunci pencarian "*Named Entity Recognition with Deep Learning*". Sedangkan untuk penelitian NER bidang *tourism* menggunakan kata kunci pencarian "*Named Entity Recognition Tourism*". Literatur yang berhasil dikumpulkan yaitu sebanyak 10 jurnal dengan 8 literatur membahas tentang umum dan 2 literatur membahas tentang *tourism*. Seperti pada tabel 1, berisi daftar mengenai referensi dan tahunnya. Kemudian dari 10 literatur yang sudah terkumpul akan dipetakan berdasar literatur, konferensi dan dari manakah literatur tersebut diperoleh, seperti yang ada dalam tabel 2.

TABLE I. TABEL REFERENSI DAN TAHUN PENELITIAN

No.	Tahun	Referensi
1.	2018	[3]
2.	2017	[4]
3.	2017	[5]
4.	2018	[2]
5.	2018	[6]

6.	2019	[7]
7.	2020	[8]
8.	2016	[9]
9.	2016	[10]
10.	2019	[11]

III. PEMBAHASAN

Dalam bab ini akan membahas bagaimana tahapan-tahapan dalam melakukan NER. Tahapan tersebut terdiri dari beberapa langkah seperti pengumpulan data NER, *preprocessing*, ekstraksi fitur dan juga penerapan model.

1) Pengumpulan Data

Data yang digunakan dalam NER berupa teks datau dokumen yang didapatkan dari berbagai macam sumber. Seperti pada tabel 3 yang menampilkan *dataset* dari penelitian sebelumnya yang mencakup jumlah datanya dan juga jenis bahasanya yang diambil dalam penelitian. Jika dilihat dari tabel 3 tersebut maka dapat dianalisis bahwa dari 10 literatur penelitian, bahasa yang paling banyak digunakan adalah bahasa Indonesia dan bahasa Inggris dengan masing-masing berjumlah 4.

TABLE II. TABEL REFERENSI *DATASET*, JUMLAH DAN BAHASA PADA PENELITIAN

Referensi	***Dataset***	**Jumlah**	**Bahasa**
[3]	China EMR	55.485 kalimat	Cina
[4]	*Twitter*	480 tweet	Indonesia
[5]	*GM BioCreative II* dan korpus *JNLPBA*	20.000 dan 22.402 kalimat	Inggris
[2]	Wikipedia	700.000 artikel	Indonesia
[6]	*Twitter* dan *Gazetteer OpenStreetMap* (OSM)	1.152 tweet	Indonesia
[7]	Artikel Web	-	Indonesia
[8]	*China Judgements Online*	1.000 dokumen	Cina
[9]	ConLL 2003 dan OntoNotes 5.0	5.000 kalimat	Inggris dan Jerman
[10]	TripAdvisor dan Wikipedia	6.996 kalimat	Inggris
[11]	TripAdvisor, Traveloka, Hotel.com	-	Inggris

2) Preprocessing

Preprocessing merupakan tahapan untuk memproses dan mempersiapkan data agar lebih mudah untuk diolah. Dalam tahap ini ada beberapa jenis *preprocessing* yang dapat digunakan seperti *case folding*, normalisasi, tokenisasi dan *stopword*, bergantung sesuai dengan kebutuhannya. Pada penelitian [8], mereka menghapus seluruh spasi dan memberi keterangan pada data dengan label BIO. Sedangkan [6], melakukan beberapa *preprocessing* dengan menghapus semua tanda baca kecuali tanda tanya (?), titik (.) dan strip (-) kemudian menghapus kata "RT" di awal kalimat, menghapus URL, *case folding* dan mengganti kata singkatan dengan nama aslinya.

3) Ekstraksi Fitur

Tahapan ekstraksi fitur digunakan untuk mengubah data mentah yang berupa teks menjadi vektor, ini dilakukan karena *deep learning* tidak dapat bekerja secara langsung pada data mentah. Melalui ekstraksi fitur ini, algoritma pembelajaran dapat mengetahui karakter dari sebuah data. *Word vector representation* atau yang biasa disebut *word embedding* dapat melakukan konversi sebuah teks menjadi angka. Bererapa macam *word embedding* yang dapat digunakan seperti *Word2Vec*, *GloVe* dan *Skip-gram*.

Pada penelitian [2], *word embedding* dilatih dengan model ruang vektor menggunakan pendekatan *Skip-gram*. Representasi kata yang terdistribusi pada ruang vektor berguna untuk membantu algoritma pembelajaran dalam mendapatkan performa yang lebih baik dengan mengelompokkan kata yang mirip. Kemudian [4] menggunakan *word embedding, neighbor word embedding dan POS Tag*. Hasil dari *word embedding* akan menjadi input pada *neighbor word embedding*. *Neighbor word embedding* ini terdiri dari satu kata di sebelah kiri dan juga satu kata di sebelah kanan kata saat ini, jika kata ada di awal kalimat maka kata tidak memiliki kata di sebelah kiri sehingga vektor di sebelah kiri adalah 0, jika kata ada di akhir kalimat maka sebaliknya bahwa vektor 0 di sebelah kanan. Untuk *POS Tag*, ini dilakukan untuk melabeli tiap kata dengan kata kerja, kata benda, kata sifat dan sebagainya. Selain itu [4] menerapkan *propose the continuous bag-of-words* (CBOW) dari kumpulan data yang tidak dianotasi karena teks biomedis berbeda dengan korpora domain umum.

4) Model NER

Dalam penelitian NER sebelumnya, terdapat beberapa model metode yang digunakan dalam membangun NER. Metode-metode yang telah digunakan sebelumnya dapat dilihat pada tabel 4.

TABLE III. TABEL METODE DAN REFERENSI

Referensi	Metode
[4]	LSTM, BiLSTM
[5]	LSTM, BiLSTM
[2]	BiLSTM, CNN
[6]	*Recurrent* CNN
[7]	BiLSTM, CRF
[8]	BiLSTM, CRF
[9]	BiLSTM, CNN
[3]	*Multitask* BiRNN
[11]	BERT, spaCy
[10]	CRF

Setelah dilakukan analisis dari seluruh literatur penelitian yang ada, ternyata ditemukan bahwa tren dalam penerapan NER menggunakan *deep learning* lebih banyak menggunakan metode BiLSTM yang dikombinasikan dengan metode lain seperti LSTM, CNN dan CRF. Kemudian dalam penelitian NER sebelumnya juga diketahui bahwa penelitian telah dilakukan di berbagai bidang. Bidang ataupun domain dalam NER yaitu pada media sosial *Twitter* yang menganalisis *tweet* bahasa Indonesia, seperti yang dilakukan oleh [4]. Selain itu [6] juga melakukan NER yang sama tetapi untuk mengenali sebuah kejadian daei sebuah *tweet*.

Pada bidang biomedis terdapat penelitian yang dilakukan [5], dengan mengenali entitas dalam domain biomedis seperti *DNA, RNA*, protein, *cell line*, dan *cell type*. Penelitian yang dilakukan [2] menerapkan NER dalam domain berita dengan mengambil beria dan artikel tentang sejarah Indonesia. Domain yang sama juga dilakukan oleh [7] tetapi berita yang diambil adalah berita tentang politik. Penelitian [9] juga melakukan NER pada artikel berita, akan tetapi tidak memiliki kriteria khusus tentang data berita apa yang akan diambil. Kemudian penelitian [8] melakukan NER dalam bidang hukum dengan mengidentifikasi entitas *name, organization, judicial organization, docket number,* dan *crime type*. Dalam bidang medis, penelitian dilakukan oleh [3] dengan menerapkan NER pada *electronic medical records* yang berbahasa Cina. Selain itu untuk bidang *tourism* sendiri, baru dua penelitian yang ditemukan dan tidak menggunakan *deep learning* karena minimnya penelitian. Penelitian NER bidang *tourism* telah dilakukan oleh [11] dan [10]. Penelitian [11] mengambil data dari artikel web yaitu TripAdvisor, Traveloka, dan *Hotels.com*. dengan bahasa Inggris, untuk entitas yang diidentifikasi adalah *location, organization* dan *facility*. Sedangkan [10] mengambil data dari TripAdvisor dan Wikipedia dengan mengidentifikasi entitas *location, person, organization, money, percent, date* dan *time*. Setelah dilihat diketahui bahwa dalam penerapan NER pada bidang *tourism* lebih banyak dalam mengenali entitas lokasi dan organisai dan data yang diambil hanya dari *website*.

Temuan selanjutnya yaitu tentang hasil dari penelitian yang sudah ada tentang NER. Dapat diketahui bahwa seperti yang ada pada tabel 5, diidentifikasi dengan referensi, hasil penelitian dan entitas yang digunakan. Kemudian kesulitan-kesulitan ataupun kekurangan yang terdapat pada penelitian NER dapat diketahui. Dapat dilihat seperti pada penelitian [4] bahwa jika hanya menerapkan *word embedding* tanpa dikombinasikan *POS Tag*, kinerjanya masih kurang tetapi jika dikombinasikan maka akan meningkatkan akurasi sebanyak 13.12%. Kemudian pada [2] diketahui jika model yang lebih kompleks yaitu BiLSTM-CNN-LSTM justru mendapatkan skor terkecil. Menurut peneliti masalah tersebut terjadi karena *dataset*-nya yang masih sedikit sehingga tidak dapat mengakomodasi model kompleks untuk dipelajari. Pada penelitian [6] terdapat entitas nama lokasi yang unik dan tidak ada dalam data latih serta *Gazetteer* ternyata dapat membuat sistem tidak mengenali kelas entitas dengan tepat.

TABLE IV. TABEL HASIL DAN ENTITAS

Referensi	Metode	Hasil	Entitas
[3]	*Multitask* BiRNN	Memperoleh akurasi skor F sebesar 93.31%	*Disease, symptom, treatment, test,* dan *disease group*.
[4]	LSTM, BiLSTM	Mendapatkan skor F1 sebesar 77.08%, dengan kombinasi POS Tag dan Word Embedding dapat meningkatkan skor F1sebesar 13.12%	*Organization, person, location*.
[5]	LSTM, BiLSTM	Dengan BiLSTM-RNN dipadukan dengan CRF mendapat skor F1 86.55% pada data GM dan 73.99% pada data JNLPBA.	*DNA, RNA, protein, cell line, cell type*.
[2]	BiLSTM, CNN	Pada kombinasi BiLSTM-CNN mendapatkan skor F1 79.43%	*Person, organization, location, event*.
[6]	*Recurrent* CNN	Memperoleh skor F1 93.53%	*Loc, Gpe, Bld, Npl, Hwymse, Obj, Mse, Time, Date, Other*.
[7]	BiLSTM, CRF	BiLSTM dengan CRF mampu memperoleh akurasi sebesar 87,77%	*person, organization, time, quantity location* dan *other*.
[8]	BiLSTM, CRF	Skor F1 yang didapat sebesar 0.855.	*Name, Location, Judicial Organization, Docket Number* dan *Crime Type*
[9]	BiLSTM, CNN	Mendapatkan akurasi skor F1 91.62% pada CoNLL 2003 dan 86.28% pada OntoNotes.	*Location, organization, person,* dan *miscellaneous*
[10]	CRF	F1 skor yang didapat sebesar 83%.	*Location, person, organization, money,*

			percent, date dan *time*
[11]	BERT, spaCy	Diketahui untuk dengan spaCy untuk entitas LOC/ORG mendapat akurasi 150.97, FAC sebesar 77.42, dan LOC/ORG/FAC 91.75 pada data *test*. Sedangkan dengan BERT untuk entitas LOC/ORG mendapat akurasi F1 0.258, FAC 0.245 dan LOC/ORG/FAC 0.464. Model ini masih menghasilkan error 8%-25%	*Location, organization, facility*

IV. KESIMPULAN

Dalam tinjauan literatur ini, terdapat 10 literatur yang didapatkan melalui *Google Scholar, Elsevire, arXiv* dan *Medwell* dalam rentang tahun 2016 hingga 2020. Literatur telah dikaji dan dapat diketahui bahwa dalam penelitian NER sebelumnya, penelitian terbaru dilakukan menggunakan *deep learning* dengan beberapa metode seperti LSTM, BiLSTM, CNN dan dapat dikombinasikan satu sama lain. Penggunaan metode BiLSTM sangat populer dan paling banyak digunakan oleh peneliti. Kemudian penggunaan ekstraksi fitur sangat mempengaruhi kinerja metode dan dapat meningkatkan hasil dari metode. Selain itu, saat ini tren penerapan NER beberapa tahun terakhir diterapkan untuk biomedis, medis, berita, media sosial dan *tourism*. Akan tetapi ternyata penelitian NER dalam bidang *tourism* masih sedikit dilakukan dan sedikit yang menggunakan *deep learning*. Melihat dari tinjauan literatur ini maka penerapan NER pada bidang *tourism* masih harus dikembangkan lagi penelitiannya. Pengembangan penelitian pada bidang *tourism* bisa dilakukan dengan memperbanyak penelitiannya di bidang tersebut menggunakan *deep learning* yang lebih bervariasi dan dapat mengenali kelas entitas yang lebih luas.

REFERENCES

[1] "TripAdvisor Network Effect and the Benefits of Total Engagement | TripAdvisor Insights." [Online]. Available: https://www.tripadvisor.com/TripAdvisorInsights/w828. [Accessed: 24-Jun-2020].

[2] W. Gunawan, D. Suhartono, F. Purnomo, and A. Ongko, "Named-Entity Recognition for Indonesian Language using Bidirectional LSTM-CNNs," *Procedia Comput. Sci.*, vol. 135, pp. 425–432, 2018.

[3] S. Chowdhury *et al.*, "A multitask bi-directional RNN model for named entity recognition on Chinese electronic medical records," *BMC Bioinformatics*, vol. 19, no. Suppl 17, 2018.

[4] V. Rachman, S. Savitri, F. Augustianti, and R. Mahendra, "Named entity recognition on Indonesian Twitter posts using long short-term memory networks," *2017 Int. Conf. Adv. Comput. Sci. Inf. Syst. ICACSIS 2017*, vol. 2018-Janua, pp. 228–232, 2017.

[5] C. Lyu, B. Chen, Y. Ren, and D. Ji, "Long short-term memory RNN for biomedical named entity recognition," *BMC Bioinformatics*, vol. 18, no. 1, pp. 1–11, 2017.

[6] F. N. Putra and C. Fatichah, "Klasifikasi jenis kejadian menggunakan kombinasi NeuroNER dan Recurrent Convolutional Neural Network pada data Twitter," *Regist. J. Ilm. Teknol. Sist. Inf.*, vol. 4, no. 2, p. 81, 2018.

[7] H. Permana and K. K. Purnamasari, "Named Entity Recognition Using Bidirectional Lstm-Crf Methods in Indonesian Text," *Procedia Comput. Sci.*, no. 112, 2019.

[8] P. Tang, P. Yang, Y. Shi, Y. Zhou, F. Lin, and Y. Wang, "Recognizing Chinese judicial named entity using BiLSTM-CRF," *J. Phys. Conf. Ser.*, vol. 1592, no. 1, 2020.

[9] J. P. C. Chiu and E. Nichols, "Named Entity Recognition with Bidirectional LSTM-CNNs," *Trans. Assoc. Comput. Linguist.*, vol. 4, no. 2003, pp. 357–370, 2016.

[10] J. Vijay and R. Sridhar, "A Machine Learning Approach to Named Entity Recognition for the Travel and tourism Domain," *Asian J. Inf. Technol.*, vol. 15, no. 21, pp. 4309–4317, 2016.

[11] C. Chantrapornchai and A. Tunsakul, "Information Extraction based on Named Entity for Tourism Corpus," *JCSSE 2019 - 16th Int. Jt. Conf. Comput. Sci. Softw. Eng. Knowl. Evol. Towar. Singul. Man-Machine Intell.*, pp. 187–192, 2019.

Tinjauan Literatur : Named Entity Recognition pada Ulasan Wisata

ORIGINALITY REPORT

2%	2%	0%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	journal.unipdu.ac.id:8080 Internet Source	1%
2	www.vemale.com Internet Source	1%
3	textly.net Internet Source	<1%
4	www.slideshare.net Internet Source	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		