Kajian Literatur *Named Entity Recognition* pada Domain Wisata

Annisa Zahra
Program Studi Informatika
Universitas Islam Indonesia
Yogyakarta, Indonesia
17523075@students.uii.ac.id

Ahmad Fathan Hidayatullah
Program Studi Informatika
Universitas Islam Indonesia
Yogyakarta, Indonesia
fathan@uii.ac.id

Septia Rani
Program Studi Informatika
Universitas Islam Indonesia
Yogyakarta, Indonesia
septia.rani@uii.ac.id

***Abstrak*—Saat merencanakan perjalanan wisata, pencarian destinasi wisata merupakan hal yang umumnya dilakukan. Proses tersebut seringkali dilakukan menggunakan bantuan mesin pencari, yaitu dengan membaca artikel yang tersedia di internet dan ditulis oleh orang lain. Pada proses pencarian informasi tersebut, terkadang dibutuhkan waktu yang tidak sedikit karena perlu membaca artikel-artikel yang tersedia untuk memperoleh informasi yang relevan. *Named Entity Recognition* (NER) dapat digunakan dalam mendeteksi entitas nama pada suatu teks sehingga dapat membantu pengguna dalam menemukan informasi yang diinginkan. Makalah ini mengkaji sebanyak 8 literatur mengenai NER pada domain wisata yang didapat dari hasil pencarian pada *Google Scholar* dengan kata kunci "*Tourism Named Entity Recognition*". Dari kajian literatur yang telah dilakukan, diperoleh informasi bahwa model NER yang paling banyak digunakan pada domain wisata adalah *Bidirectional Encoder Representations from Transformers* (BERT). Model BERT bertujuan untuk melakukan pelatihan representasi kata menggunakan konverter dua arah dengan menyesuaikan konteks pada sisi kiri dan kanan semua lapisan. Sehingga, penggunaan BERT dapat membantu mencegah terjadinya ambiguitas pada suatu kata yang mengakibatkan kesalahan pengenalan entitas. Hasil penelitian ini diharapkan dapat membantu dalam pengembangan NER pada domain wisata selanjutnya.**

Kata kunci—named entity recognition, wisata

I. LATAR BELAKANG

Pertumbuhan pariwisata terjadi setiap tahunnya. Pada tahun 2019 tercatat sebanyak 1,5 miliar kunjungan wisatawan internasional secara global. Hal tersebut mengalami peningkatan sebesar 4% dari tahun sebelumnya [1]. Salah satu studi yang dilakukan oleh *Google Travel* menemukan bahwa 74% wisatawan merencanakan perjalanan mereka melalui internet [2]. Pencarian destinasi wisata adalah salah satu tahapan yang umumnya dilakukan pada saat merencanakan perjalanan. Proses pencarian tersebut dapat dilakukan melalui internet, seperti dengan membaca artikel yang beredar. [3] berpendapat bahwa informasi pariwisata saat ini sudah tersebar di sekitar, namun untuk mencari informasi, biasanya memakan waktu jika harus menelusuri hasil dari mesin pencari, memilih, dan melihat detailnya. Sehingga untuk memudahkan calon wisatawan dalam memperoleh informasi, dapat dilakukan ekstraksi informasi pada teks dengan domain topik mengenai wisata.

Named Entity Recognition (NER) dapat membantu proses ekstraksi informasi dengan cara mengidentifikasi suatu entitas nama. NER membantu pengguna untuk menghasilkan korpus yang lebih bermakna dengan mengidentifikasi nama-nama yang tepat di korpus dan mengklasifikasikannya ke dalam kelompok-kelompok seperti orang, organisasi, lokasi, dan lainnya [4]. Pada domain wisata, entitas yang diidentifikasi dapat berupa nama tempat wisata, tempat penginapan, fasilitas, serta lokasinya. Identifikasi entitas terkait diharapkan dapat memudahkan calon wisatawan dalam menemukan destinasi wisata melalui internet.

Kajian ini bertujuan untuk mengetahui tren pemodelan NER yang pernah digunakan pada domain wisata dan juga mengetahui bagaimana penelitian NER pada domain wisata yang sudah dilakukan sebelumnya. Sehingga harapannya dengan adanya kajian pustaka ini dapat membantu peneliti di masa depan dalam melakukan penelitian NER pada domain wisata agar mendapatkan hasil yang lebih baik.

Berdasarkan tujuan kajian di atas, pertanyaan yang mendasari penelitian ini sebagai berikut:

1) Bagaimana tren pemodelan NER yang pernah digunakan pada domain wisata?
2) Bagaimana penelitian NER pada domain wisata yang sudah dilakukan sebelumnya?

II. METODOLOGI

Pengumpulan data yang berupa literatur dilakukan melalui *Google Scholar*. Kriteria inklusi dan eksklusi yang diterapkan dapat dilihat pada Tabel 1. Kata kunci yang digunakan, yaitu "*Tourism named entity recognition*". Total literatur yang digunakan pada penelitian ini berjumlah 8 buah seperti yang ditunjukkan pada Tabel 2.

Tabel 1. Kriteria inklusi dan eksklusi yang diterapkan

Kriteria Inklusi	Kriteria Eksklusi
Makalah akademis	Tidak termasuk ke dalam makalah akademis
Literatur membahas NER pada domain wisata	Literatur mengenai NER, namun bukan pada domain wisata
Literatur ditulis dalam bahasa Indonesia atau bahasa Inggris	Literatur mengenai NER pada domain wisata yang ditulis selain dalam bahasa Indonesia dan Inggris.

Tabel 2. Literatur yang diambil

Referensi	Tahun
[5]	2020
[6]	2019
[7]	2019
[8]	2019
[9]	2019
[10]	2016
[11]	2008
[12]	2008

III. PEMBAHASAN

Pada bagian ini akan dibahas mengenai hasil kajian dan temuan terhadap literatur-literatur yang dikaji.

A. Hasil Kajian

Berdasarkan studi yang dilakukan, implementasi dari NER membutuhkan beberapa tahapan untuk menghasilkan keluaran yang diinginkan. Berikut ini merupakan tahapan yang dilakukan pada proses NER.

1) Pengumpulan Data

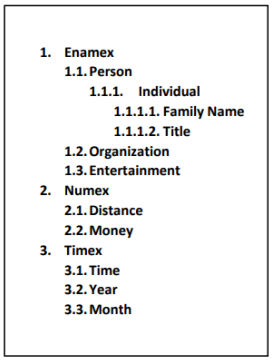
Data yang akan diolah dapat menggunakan *dataset* yang sudah ada sebelumnya dan sudah diberi label, ataupun membuat *dataset* sendiri. Rincian *dataset* yang digunakan pada setiap literatur dapat dilihat pada Tabel 2. Sebagian besar literatur menggunakan *dataset* yang dibuat sendiri oleh para peneliti, namun terdapat juga literatur yang menggunakan *dataset* yang sudah ada sebelumnya, yaitu [7] yang menggunakan *Microsoft Research Asia* (MSRA) *dataset*.

2) Pre-processing

Tahapan ini digunakan untuk mempersiapkan data sebelum masuk ke tahap selanjutnya agar mengurangi kemungkinan terjadinya kegagalan saat evaluasi model. Proses *pre-processing* yang dilakukan pada data tergantung dengan kebutuhan masing-masing penelitian. [10] melakukan *pre-processing* berupa tokenisasi, *part-of-speech tagging*, dan *noun phrase chunking*. [8] juga melakukan *noun phrase chunking* dengan menggunakan *spaCy noun chunk*.

3) Pelabelan Data

Tahapan selanjutnya adalah memberikan label/*tag* pada data. Walaupun semua literatur termasuk pada domain wisata, namun pelabelan entitasnya berbeda, seperti yang ditunjukkan pada Tabel 2. Terdapat dua jenis NER jika dilihat dari sisi pelabelan datanya, yaitu *nested* NER dan *flat* NER. *Flat* NER memiliki variasi label yang tidak terlalu banyak, sehingga kurang dapat digunakan untuk menangkap informasi semantik dalam teks. Sedangkan *nested* NER memiliki label yang banyak, namun arsitektur pemodelannya menjadi kompleks. Hampir semua literatur yang dikaji menggunakan *flat* NER, kecuali [12] yang menggunakan *nested* NER pada penelitiannya. Pelabelan pada penelitian tersebut berfokus pada domain wisata dan wisata kesehatan, seperti nama tempat, alamat, museum, tempat religius, taman, monumen, bandar udara, stasiun kereta api, acara, pengobatan untuk penyakit, jarak, dan tanggal. Contoh dari *nested* NER dapat dilihat pada Gambar 1.



Gambar 1. Contoh pelabelan bertingkat

4) Ekstraksi Fitur

Algoritma tertentu tidak dapat menerima masukan berupa teks secara langsung, seperti pada algoritma *deep learning*. Sehingga dibutuhkan semacam representasi numerik dari masukan yang berupa teks agar masukan tersebut dapat diproses. Ekstraksi fitur digunakan untuk merepresentasikan seluruh data dan juga untuk menggali informasi berguna di dalamnya yang akan digunakan pada tahap selanjutnya.

Representasi numerik dapat dilakukan dengan menggunakan teknik *character embedding* yang akan mengubah karakter menjadi kumpulan angka seperti yang dilakukan [7]. Penelitian tersebut mentransformasi sebuah karakter dalam bahasa Cina ke dalam sebuah vektor berdimensi 768. Selain itu, mereka juga menghitung representasi fitur lokal dengan ResCNN, hasilnya kemudian akan digabungkan dengan hasil dari *character embedding* sebelum dimasukkan ke dalam jaringan *Bidirectional Long Short-Term Memory* (BLSTM) dan *Conditional Random Fields* (CRF).

5) Pemodelan Named Entity Recognition

Langkah berikutnya adalah membangun model NER yang terdiri dari satu atau lebih metode di dalamnya. Model yang digunakan pada setiap literatur dapat dilihat pada Tabel 2. Penggunaan model BERT (*Bidirectional Encoder Representations from Transformers*) adalah yang paling banyak digunakan. Beberapa literatur mengombinasikannya dengan metode lain seperti BLSTM dan CRF. BERT merupakan sebuah model pra-pelatihan bahasa dalam sekala besar yang dirilis oleh *Google* dengan berdasarkan *bidirectional transformer*. Model BERT bertujuan untuk melakukan pelatihan representasi kata menggunakan konverter dua arah dengan menyesuaikan konteks pada sisi kiri dan kanan semua lapisan. Sehingga, penggunaan BERT dapat membantu mencegah terjadinya ambiguitas pada suatu kata yang mengakibatkan kesalahan pengenalan entitas [9].

6) Evaluasi

Evaluasi dilakukan terhadap data pengujian untuk mengukur kinerja dari model NER yang sudah dibuat. Pengukuran dilakukan menggunakan F1-*Score* atau F-*Measure*. Tabel 2 menunjukkan F1-*Score* terbaik yang didapatkan pada setiap penelitian sebelumnya. Nilai tertinggi diantara semua literatur yang dikaji pada penelitian ini diraih oleh [6].

Tabel 2. *Overview* data yang diekstrak dari setiap literatur

Referensi	*Dataset*	Ukuran *Dataset*	Bahasa *Dataset*	Model	Label Entitas	F1-*score* (%)
[5]	*Mongolian Tourism Corpus*	16.000 kalimat	Mongol	BERT	*Mongolian, foreigner, administration place, natural sight, public place, marker building, business, religion, culture, education, sports, music, department, company, politics, charity, military, car*.	82,09
[6]	Data diperoleh dari percakapan antara pengguna dan agen *customer support.*	21.000 pesan	Inggris	*Library SpaCy*	Hotel, *location,* hotel + *location.*	96
[7]	MSRA	55.280 kalimat	Cina	BERT – ResCNNs – BLSTM – CRF	*Location, organization, person.*	95,41
	CTFAE	15.845 kalimat			*Chinese name, area, construction time, internal attraction, location, nickname*.	92,17
[8]	Data berasal dari Tripadvisor, Traveloka, dan Hotels.com. Delapan provinsi yang dipertimbangkan dalam pengumpulan data ini, yaitu Bangkok, Phuket, Chaingmai, Phang-nga, Chonburi, Suratani, Krabi, Prachuap Khiri Khan. Pada setiap provinsinya terdapat 6 fitur yang dikumpulkan, yaitu nama, deskripsi, alamat, fasilitas, *nearby*, dan ulasan.	Tidak disebutkan	Tidak disebutkan	*Library SpaCy*, BERT	LOC (lokasi atau nama tempat), ORG (hotel atau nama akomodasi), FACILITY (fasilitas)	Dengan menggunakan BERT: - Pada data latih: 28,7 - Pada data uji: 46,4
[9]	Data teks perjalanan yang diambil dari *Ctrip* dan *Mafengwo, Raiders.*	13.464 kalimat	Cina	BERT – BLSTM – CRF (BBLC)	*Person, location, organization, time, thing, other*.	- Entitas *person*: 84,79 - Entitas *location:* 91,46 - Entitas *organization:* 71,13 - Entitas *time*: 85,22 - Entitas *thing*: 91,39
[10]	Data diperoleh dari hasil pencarian teratas pada *Google Search* dengan menggunakan kata kunci: "*Top tourism place in %*". Simbol % menunjukkan berbagai benua seperti Asia, Eropa, Amerika, Afrika, dan Australia.	2.686 entitas unik	Inggris	YATSI (*Yet Another Two Stage Idea*) dengan NBC (*Naïve Bayes Classifer*) dan KNN (*K-Nearest Neighbor*)	*Nature, city, region, negative*.	69,1
[11]	Korpus paralel yang dikumpulkan dengan menggunakan *web crawler*, serta berasal dari domain *travel, tourism,* dan *culture*.	Tidak disebutkan	Inggris dan Hindi	*Phonetic Matching*	PE (nama orang), OE (nama organisasi), LE (nama lokasi)	Tidak disebutkan
[12]	Korpus yang berisi kata-kata yang dikumpulkan dari domain wisata.	94.000 kata	Tamil	CRF	Terdapat 106 label yang digunakan pada penelitian ini dan saling terkait satu sama lain secara hierarki.	80,44%

Sorotan pada setiap literatur yang dikaji dapat dilihan pada Tabel 3. Sorotan yang dibahas bervariasi mulai dari data hingga metode yang digunakan.

Tabel 3. Sorotan pada penelitian sebelumnya

Referensi	Sorotan
[5]	Berangkat dari kurangnya data yang tersedia untuk mengidentifikasi *travel-related named entities*, terutama dalam bahasa Mongol, penelitian tersebut memperkenalkan korpus baru untuk *Mongolian Tourism Named Entity Recognition* (MTNER) yang dapat digunakan pada berbagai penugasan *Natural Language Processing* (NLP) dalam bahasa Mongol.
[6]	Pada penelitian ini, penggabungan hotel dan lokasi sebagai satu entitas menghasilkan akurasi yang lebih baik dibandingkan dengan hotel dan lokasi yang dianggap sebagai entitas terpisah.
[7]	Penelitian ini adalah penelitian pertama pada penugasan *sequence labeling* yang menggabungkan BERT, ResCNNs, BLSTM, dan CRF. Gabungan model tersebut berhasil memperoleh F1-*score* yang cukup tinggi, yaitu sebesar 95,41%.
[8]	Penelitian tersebut dimulai dari proses *scraping* informasi dari web yang dipilih berdasarkan *tag* HTML untuk membangun korpus yang akan digunakan pada pemodelan NER. Masukan untuk model yang dibangun adalah berupa kalimat beserta label entitasnya. Selain itu dibuat juga *word embedding* untuk domain wisata. Pendekatan dengan *library spaCy* menghasilkan akurasi sekitar 91% untuk data latih dan data uji. Sedangkan untuk BERT, akurasi yang didapatkan adalah sekitar 70%.
[9]	Penelitian ini mengombinasikan BERT dengan metode BLSTM dan CRF yang disebut dengan model BBLC (BERT-BLSTM-CRF). Model tersebut kemudian diuji coba dengan menggunakan *dataset* yang berisikan data berupa teks pada domain wisata. *Dataset* yang sama juga diuji coba pada model BLSTM-CRF dan model CRF. BBLC memperoleh nilai F1-*Score* yang lebih tinggi daripada model lainnya pada entitas *location*, *organization*, dan *thing*. Sementara F1-*Score* tertinggi pada entitas *person* diraih oleh model CRF. Model BLSTM-CRF meraih nilai F1-*Score* tertinggi pada entitas *time*.
[10]	Penelitian tersebut menggunakan algoritma *Yet Another Two Stage Idea* (YATSI) yang memiliki dua tahap klasifikasi. Tahap klasifikasi pertama menggunakan *Naïve Bayes Classifier* (NBC) yang akan dilatih menggunakan data yang sudah berlabel. Alasan pemilihan NBC adalah karena NBC bersifat probabilistik dan mudah diimplementasikan. Tahap klasifikasi kedua menggunakan pendekatan *K-Nearest Neighbor* (KNN) untuk memprediksi data yang belum berlabel.
[11]	Peningkatan akurasi pada penelitian tersebut dapat dilakukan dengan menambahkan lebih banyak aturan umum. Sistem yang dibangun juga masih harus diperbaiki agar dapat mengenali frasa nama entitas dan juga singkatan.
[12]	Penelitian ini tergolong menggunakan *nested tagging* pada pelabelan datanya. Penanganan *nested tagging* dan pencegahan ambiguitas dapat dilakukan dengan mengisolasi *tagset* ke dalam beberapa *subset* yang sesuai dengan jumlah level pada hierarkinya. Masing-masing *subset* berisi *tag* dari satu level yang sama. *Tagset* pada penelitian tersebut terdiri dari 3 level, sehingga terdapat tiga model CRF yang masing-masing menangani level yang berbeda. Hasil dari ketiga model tersebut akan digabungkan.

B. Temuan

Saat ini, penelitian NER pada domain wisata masih tergolong sedikit. Terdapat banyak pilihan jenis entitas untuk dikenali pada domain wisata, mulai dari nama tempat penginapan, atraksi wisata, hingga jarak dan tanggal. Pengenalan entitas yang diinginkan disesuaikan dengan kebutuhan masing-masing peneliti. Kebutuhan yang berbeda tersebut menyebabkan banyak peneliti membuat *dataset* sendiri untuk memenuhi kebutuhan penelitiannya. Kendala dalam pembuatan *dataset* dapat berupa keterbatasan biaya dan ukuran data seperti pada [9] yang akhirnya berdampak pada performa model yang dihasilkan.

IV. KESIMPULAN

Penelitian ini mengkaji 8 literatur mengenai NER pada domain wisata yang didapatkan melalui pencarian pada *Google Scholar*. Dari hasil kajian didapatkan informasi bahwa saat ini penelitian NER pada domain wisata masih tergolong sedikit dan model yang paling banyak digunakan adalah BERT. Model BERT bertujuan untuk melakukan pelatihan representasi kata menggunakan konverter dua arah dengan menyesuaikan konteks pada sisi kiri dan kanan semua lapisan. Sehingga, penggunaan BERT dapat membantu mencegah terjadinya ambiguitas pada suatu kata yang mengakibatkan kesalahan pengenalan entitas.

Sebagian besar penelitian menggunakan *dataset* yang dibuat sendiri oleh para peneliti. Pelabelan yang digunakan terbagi menjadi dua jenis, yaitu *flat* NER dan *nested* NER. Kendala dalam pembuatan *dataset* berupa keterbatasan biaya dan juga ukuran data yang digunakan, sehingga berdampak pada performa model yang dihasilkan. Penelitian mengenai NER pada domain wisata selanjutnya diharapkan dapat mengembangkan *dataset* yang sudah dibuat dan menambah jumlah entitas nama yang akan dikenali, sehingga dapat menghasilkan informasi yang lebih rinci.

REFERENSI

[1] "World Tourism Barometer Nº18 January 2020 | UNWTO." https://www.unwto.org/world-tourism-barometer-n18-january-2020 (accessed Dec. 07, 2020).

[2] "The 2014 Traveler's Road to Decision." https://www.thinkwithgoogle.com/consumer-insights/2014-travelers-road-to-decision/ (accessed Jun. 26, 2020).

[3] C. Chantrapornchai and A. Tunsakul, *Information Extraction based on Named Entity for Tourism Corpus*. .

[4] R. Alfred, L. C. Leong, C. K. On, and P. Anthony, "Malay Named Entity Recognition Based on Rule-Based Approach," *Int. J. Mach. Learn. Comput.*, vol. 4, no. 3, pp. 300–306, Jun. 2014, doi: 10.7763/ijmlc.2014.v4.428.

[5] X. Cheng, W. Wang, F. Bao, and G. Gao, "MTNER: A Corpus for Mongolian Tourism Named Entity Recognition," 2020.

[6] B. Li, N. Jiang, J. Sham, H. Shi, and H. Fazal, "Real-world Conversational AI for Hotel Bookings," 2019. Accessed: Dec. 08, 2020. [Online]. Available: https://pybossa.com.

[7] Y. Hu, M. Nuo, and C. Tang, "A Deep Learning Approach for Chinese Tourism Field Attribute Extraction," in *Proceedings - 2019 15th International Conference on Computational Intelligence and Security, CIS 2019*, Dec. 2019, pp. 108–112, doi: 10.1109/CIS.2019.00031.

[8] 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, doi: 10.1109/JCSSE.2019.8864166.

[9] L. Xue, H. Cao, F. Ye, and Y. Qin, "A method of chinese tourism named entity recognition based on bblc model," in *Proceedings - 2019 IEEE SmartWorld, Ubiquitous Intelligence and Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Internet of People and Smart City Innovation, SmartWorld/UIC/ATC/SCALCOM/IOP/SCI 2019*, Aug. 2019, pp. 1722–1727, doi: 10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00307.

[10] K. E. Saputro, S. S. Kusumawardani, and S. Fauziati, "Development of semi-supervised named entity recognition to discover new tourism places," in *Proceedings - 2016 2nd International Conference on Science and Technology-Computer, ICST 2016*, Mar. 2016, pp. 124–128, doi: 10.1109/ICSTC.2016.7877360.

[11] A. Nayan, B. R. K. Rao, P. Singh, S. Sanyal, and R. Sanyal, "Named Entity Recognition for Indian Languages," *Proc. IJCNLP-08 Work. NER South South East Asian Lang.*, vol. 3, no. 11, pp. 97–104, 2008.

[12] Vijayakrishna R and Sobha L, "Domain Focused Named Entity Recognizer for Tamil Using Conditional Random Fields," *Proc. ofthe IJCNLP-08 Work. NER South South East Asian Lang.*, no. January, pp. 59–66, 2008, [Online]. Available: http://www.aclweb.org/anthology/I08-5009.