Kajian Literatur Named Entity Recognition pada Domain Artikel Wisata

by John Doe

Submission date: 24-Nov-2020 08:52PM (UTC+0700)
Submission ID: 1455103567
File name: iteratur_Named_Entity_Recognition_pada_Domain_Artikel_Wisata.pdf (308.34K)
Word count: 3757
Character count: 23725

Kajian Literatur *Named Entity Recognition* pada Domain Artikel Wisata

[1]xxx, [2]yyy, [3]zzz
[1]Program Studi Informatika Program Sarjana
[2,3]Jurusan Informatika
Universitas Islam Indonesia
Yogyakarta, Indonesia
[1]xxx@students.uii.ac.id, [2]yyy@uii.ac.id, [3]zzz@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 literatur mengenai NER dengan *deep learning* dari tahun 2018 hingga 2020 dan NER pada domain artikel wisata. Hasil penelitian ini diharapkan dapat membantu dalam pengembangan NER pada domain artikel wisata selanjutnya.**

Kata kunci—named entity recognition, deep learning, wisata

I. LATAR BELAKANG

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. [1] 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 mengidentifikasi dan mengekstrak informasi yang disebut *named-entity*. 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 [2].

Kajian ini bertujuan untuk mengetahui metode NER mana yang terbaik serta populer pada dua tahun terakhir, dan juga mengetahui bagaimana penelitian NER pada domain wisata yang sebelumnya sudah dilakukan. Sehingga harapannya dengan adanya kajian pustaka ini dapat membantu peneliti di masa depan dalam menentukan metode mana yang sebaiknya dipilih agar penelitian mengenai NER pada domain wisata ke depannya dapat memperoleh hasil yang lebih baik.

II. METODOLOGI

A. Pertanyaan Penelitian

Pertanyaan yang mendasari penelitian ini sebagai berikut.

1) Bagaimana tren penelitian NER saat ini?
2) Metode apa yang memiliki performa paling baik untuk menyelesaikan kasus NER?
3) Bagaimana penelitian NER pada domain wisata yang sudah dilakukan sebelumnya?

B. Pengumpulan Data

Pengumpulan data yang berupa literatur dibagi menjadi dua tahap dan keduanya dilakukan melalui *Google Scholar*. Tahap pengumpulan data yang pertama adalah untuk mencari literatur yang memiliki kriteria berupa penelitian NER yang menggunakan pendekatan *deep learning* dan diterbitkan dari tahun 2018 hingga tahun 2020. Kata kunci yang digunakan, yaitu "*Deep learning for named entity recognition*". Sedangkan tahap pengumpulan data yang kedua adalah untuk mencari literatur yang membahas NER pada domain wisata dengan kata kunci "*Named entity recognition for tourism*".

Total literatur yang digunakan pada penelitian ini berjumlah 16 buah. Sebanyak 11 literatur diambil pada tahap pengumpulan data yang pertama seperti yang terdapat pada Tabel 1, dan sebanyak 5 literatur diambil pada tahap yang kedua dan ditunjukkan pada Tabel 2.

Tabel 1. Literatur yang diambil pada tahap pengumpulan data pertama

Referensi	Tahun
[3]	2019
[4]	2019
[5]	2019
[6]	2019
[7]	2019
[8]	2019
[9]	2018
[10]	2018
[11]	2018
[12]	2018
[13]	2018

Tabel 2. Literatur yang diambil pada tahap pengumpulan data kedua

Referensi	Tahun
[14]	2019
[15]	2019
[16]	2016
[17]	2008
[18]	2008

III. NAMED ENTITY RECOGNITION DENGAN DEEP LEARNING

Pada bagian ini akan dibahas mengenai hasil kajian dan temuan terhadap literatur-literatur yang melakukan proses NER menggunakan pendekatan *deep learning*.

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 dan Pelabelan Data

Data yang akan diolah dapat menggunakan *dataset* yang sudah ada sebelumnya dan sudah diberi label. Pemilihan data disesuaikan oleh domain penelitian yang sedang dilakukan. Tabel 3 menunjukkan *dataset* yang digunakan pada penelitian sebelumnya. *Dataset* yang digunakan cukup bervariasi dan dalam satu penelitian bisa menggunakan lebih dari satu *dataset* dengan domain yang berbeda. Domain penelitian juga cukup beragam, namun didominasi oleh domain biomedis.

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. Proses *pre-processing* yang dilakukan pada [5], yaitu mengubah semua data yang berbentuk angka dengan nilai "NUM", mengubah semua huruf menjadi huruf kecil, dan untuk kata yang tidak ada dalam *pre-trained word embedding* akan ditandai dengan UNK (*unknown*).

3) Ekstraksi Fitur

Algoritma *deep learning* tidak dapat menerima masukan berupa teks secara langsung, 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 *word embedding* yang mengubah sebuah kata menjadi kumpulan angka.

Word embedding dapat dihasilkan dengan menggunakan berbagai macam teknik, seperti *Word2Vec* yang memiliki dua algoritma (*Skip-gram* dan *Continuous Bag of Words*), *GloVe, FastText,* dan *Hellinger* PCA (H-PCA) seperti yang dilakukan oleh [6] untuk melihat mana representasi yang terbaik. *Word embedding* yang sudah pernah dilatih sebelumnya (*pre-trained word embedding)* juga dapat digunakan pada penelitian lain, seperti [10] yang menggunakan *pre-trained word embedding* pada MEDLINE *abstracts* untuk digunakan pada penelitiannya.

Tabel 3. *Dataset* pada penelitian sebelumnya

Referensi	*Dataset*	Domain	Ukuran *Dataset*	Bahasa *Dataset*
[3]	EMR *from the Second Affiliated Hospital of Harbin Medical University*	Biomedis	55.485 kalimat	Cina
[4]	CCKS-2017 Task 2	Biomedis	10.024 kalimat	Cina
[5]	NCBI *disease* & BC5CDR	Biomedis	NCBI: 7.621 kalimat BC5CDR: 1.500 artikel	Inggris
[6]	ANERcorp	Berita	5.887 kalimat	Arab
[7]	CCKS-2017 Task 2	Biomedis	10.024 kalimat	Cina
[8]	*Chinese* EMR NER *task in* CCKS-2018	Biomedis	1.000 catatan untuk pasien	Cina
[9]	NCBI-*disease*, BC5CDR, BC2GM, JNLPBA	Biomedis	NCBI-disease: 7.295 kalimat JBLPBA: 24.806 kalimat BC5CDR: 13.907 kalimat BC2GM: 20.000 kalimat	Inggris
[10]	GENIA, JNLPBA	Biomedis	GENIA: 18.546 kalimat JNLPBA: Tidak disebutkan	Inggris
[11]	CoNLL-2003 *English*, OntoNotes 5.0 *English* dan *Chinese*	Berita	Tidak disebutkan	Inggris, Cina
[12]	WeiboNER, SighanNER, MSR	WeiboNER: Media sosial SighanNER: Berita MSR: Berita	Tidak disebutkan	Cina
[13]	Menggunakan korpus yang digunakan pada [19]	Tidak disebutkan	Tidak disebutkan	Indonesia

Selain itu, terdapat juga *character embedding* yang dapat menangani masalah yang timbul jika suatu kata tidak terdapat pada kamus yang dihasilkan oleh *word embedding*. *Character embedding* ini dapat dihasilkan dengan menggunakan *Convolutional Neural Networks* (CNN) [9]. Pada tahap ekstraksi fitur, [8] menggunakan *Part of Speech Tagging*

(POS *Tagging*) untuk mengelompokkan kata menjadi kata benda, kata kerja, kata sifat, dan sebagainya.

4) Pemodelan Named Entity Recognition

Langkah selanjutnya adalah membangun model NER yang terdiri dari satu atau lebih metode di dalamnya. Metode yang digunakan peneliti sebelumnya dapat dilihat pada Tabel 4. Gabungan metode *bidirectional long short-term memory* (BLSTM) dan *conditional random fields* (CRF) adalah yang paling banyak digunakan. Lapisan CRF umumnya direpresentasikan oleh garis yang menghubungkan lapisan keluaran yang berurutan dan memiliki *state transition matrix* sebagai parameter. Maka dengan adanya lapisan seperti itu, label dari suatu kata dapat diprediksi dengan menggunakan label dari kata sebelum dan sesudahnya secara efisien. Hal ini serupa dengan penggunaan fitur masukan dari masa lalu dan masa depan yang ada pada BLSTM [4]. Pengambilan informasi dari masa lalu dan masa depan sangat berguna untuk tugas penandaan urutan seperti NER [6].

Tabel 4. Metode NER pada penelitian sebelumnya

Referensi	Metode NER
[3]	*Transfer multitask bidirectional* RNN (TMBRNN), BLSTM
[4]	BLSTM, CRF
[5]	BLSTM, CRF
[6]	BLSTM, CRF
[7]	CNN, *Lattice* LSTM, AT-Lattice LSTM, CRF
[8]	BLSTM, CRF
[9]	BLSTM, CRF
[10]	BLSTM
[11]	CNN, LSTM, CRF
[12]	BLSTM, CRF
[13]	*Bidirectional Gated Recurrent Unit*, CRF

5) Evaluasi

Evaluasi dilakukan terhadap data pengujian untuk mengukur kinerja dari model NER yang sudah dibuat. Pengukuran dilakukan menggunakan F1-*Score* atau F-*Measure*. Tabel 5 menunjukkan F1-*Score* terbaik yang didapatkan pada setiap penelitian sebelumnya. Nilai tertinggi diantara semua literatur yang dikaji pada penelitian ini diraih oleh [4].

Tabel 5. Nilai evaluasi terbaik pada penelitian sebelumnya

Referensi	Nilai F1-*Score* (%)	Model	*Dataset*
[3]	84,25	*Proposed model*	EMR *from the Second Affiliated Hospital of Harbin Medical University*
[4]	91,24	*Position-Dependent Entity Type* (PDET) + BLSTM + CRF	CCKS-2017 Task 2
[5]	85,40	*Dictionary information* + *pre-trained word embedding* + *character embedding* + *global score* + "IOBES" *segment representation scheme* + BLSTM + CRF	NCBI
[6]	90,60	Nadam + *variational dropout* + BLSTM + CRF	ANERcorp
[7]	89.64	AT-*Lattice* LSTM-CRF	CCKS-2017 Task 2
[8]	80,07	*Self-matching* + LSTM + CRF	*Chinese* EMR NER *task in* CCKS-2018
[9]	89.28	*Proposed model* + *bidirectional language model pretrained*	BC5CDR
[10]	78,4	*Proposed model*	JNLPBA
[11]	90,89	LSTM (*character*) + LSTM (*word*) + LSTM (*tag*)	CoNLL-2003 English
[12]	90,64	BLSTM + CRF + *adversarial* + *self-attention*	SighanNER
[13]	72,9	*Proposed model*	Menggunakan korpus yang digunakan pada [19]

Literatur yang dikaji membahas masalah yang berbeda dan menghasilkan beberapa hal yang dapat menjadi sorotan seperti yang ditunjukkan pada Tabel 6. Sorotan yang dibahas bervariasi mulai dari data hingga metode yang digunakan.

Tabel 6. Sorotan pada penelitian sebelumnya

Referensi	Sorotan
[3]	Pengetahuan dari domain umum yang sebelumnya sudah dilatih akan ditransfer ke *multitask deep learning* melalui *transferred layer*. Keluaran dari *transferred layer* akan masuk ke *shared layer* untuk mengekstrak relasi antar kata yang lebih akurat. Keluaran *shared layer* akan masuk ke dalam dua *layer*, yaitu *part-of-speech tagging* (POS *Tagging*) *layer* dan NER *layer*. Kedua *layer* tersebut akan dilatih secara bergantian agar pengetahuan yang dihasilkan POS *Tagging task* dapat digunakan untuk meningkatkan kinerja pada NER *task*.
[4]	Penggabungan sebuah kamus dengan sebuah model BLSTM-CRF memperoleh hasil yang lebih baik dibandingkan jika menggunakan model BLSTM-CRF saja pada saat diuji coba dengan *dataset* berbahasa Cina.
[5]	Setiap kata pada masukan yang menggunakan *dataset* berbahasa inggris akan direpresentasikan sebagai vektor. Hasil evaluasi terbaik didapat saat vektor tersebut berisi gabungan informasi dari kamus, *character embedding*, dan *word embedding*.
[6]	Pada bahasa Arab dibutuhkan representasi berbasis karakter untuk secara efektif menangkap informasi ortografik dan morfologis dari sebuah kata dan mengkodekannya ke dalam *neural representations* sebelum diproses ke dalam model. *Character representations* menggunakan CNN memperoleh hasil yang lebih baik jika dibandingkan dengan BLSTM.
[7]	*Lattice* LSTM berguna untuk mendapatkan informasi yang lebih banyak dari *electronic health record* dalam bahasa Cina. Sedangkan *adversarial training* (AT) yang merupakan sebuah metode regularisasi dapat digunakan untuk meningkatkan ketahanan sebuah model dengan menambahkan *perturbations* pada data latih.
[8]	Informasi terkait *part-of-speech* digabungkan ke dalam model pembelajaran untuk meningkatkan keakuratan saat mendeteksi batasan entitas dalam bahasa Cina. *Reduced* POS *Tagging* diusulkan sebagai metode untuk menghindari kesalahan dalam penandaan *part-of-speech*. *Self-matching attention layer* digunakan untuk menghitung relevansi dari setiap karakter terhadap keseluruhan kalimat
[9]	Pelatihan menggunakan *bidirectional language model* (BiLM) pada data yang belum diberi label dan mentransfer bobotnya ke model NER dengan arsitektur yang sama menghasilkan inisialisasi parameter yang lebih baik pada model NER.

[10]	Model yang diusulkan ditujukan untuk NER yang bersarang (*nested* NER) dengan mempertimbangkan semua wilayah yang mungkin untuk *nested* NER secara mendalam. Model tersebut dapat juga digunakan pada NER yang tidak bersarang.
[11]	Penggabungan *deep learning* dengan *active learning* dapat mengurangi jumlah data pelatihan yang diperlukan secara efisien. Jika dilihat dari sisi sampel, *active learning* tergolong efisien, namun secara komputasional dapat tergolong mahal karena memerlukan pelatihan berulang. Masalah tersebut dapat dihindari dengan memilih arsitektur NER yang ringan sehingga dapat mengurangi kompleksitas komputasi.
[12]	*Adversarial transfer learning* menggabungkan *adversarial training* ke dalam *transfer learning*. *Self-attention mechanism* digunakan untuk menangkap dependensi jarak jauh dengan lebih baik dan mensintesis informasi pada sebuah kalimat.
[13]	Kebutuhan akan pelatihan data yang besar dapat diselesaikan dengan *transfer learning*. *Transfer learning* akan membantu menyelesaikan suatu tugas yang hanya memiliki sedikit data pelatihan dengan mentransfer informasi yang didapat dari sumber lain.

B. Temuan

Beberapa temuan yang diperoleh dari hasil kajian terhadap penelitian-penelitian sebelumnya mengenai NER menggunakan *deep learning* di antaranya berupa tantangan yang dihadapi serta pengembangan/penggunaan arsitektur *deep learning* yang bervariasi.

1) Tantangan Penelitian Named Entity Recognition

Tantangan timbul pada penelitian yang menggunakan *dataset* dengan bahasa Cina. Bahasa tersebut memiliki kompleksitas yang lebih tinggi jika dibandingkan dengan bahasa yang termasuk ke dalam rumpun bahasa Roman [4]. Sama halnya dengan bahasa Cina, penerapan NER pada bahasa Arab juga memiliki tantangan karena karakteristiknya yang unik. Satu kata dalam bahasa arab dapat mempunyai bentuk morfologi yang berbeda, sehingga hal tersebut menghasilkan banyak *data sparseness*. Bahasa Arab juga tidak memiliki kapitalisasi, sehingga tidak memungkinkan untuk menggunakan kapitalisasi sebagai indikator fitur saat mendeteksi suatu entitas nama [6].

2) Pengembangan Arsitektur Deep Learning

Penggunaan *deep learning* dalam menyelesaikan masalah pada setiap penelitian memperoleh hasil yang cukup baik jika dilihat dari nilai F-*Score* yang dihasilkan. Proses *word embedding* yang dilakukan pada beberapa penelitian menggunakan *pre-trained word embedding* dari sumber lain. [9] menggunakan sebuah korpus besar yang berisi PubMed *abstracts*, [10] menggunakan *pre-trained word embedding* yang sudah dilatih pada MEDLINE *abstract* yang berisi sejumlah 2.231.686 kosa kata.

Semua literatur yang dikaji menerapkan metode LSTM pada penelitiannya. Namun terdapat pengembangan yang berbeda pada arsitektur *deep learning* yang digunakan, seperti yang ditunjukkan pada Tabel 7. Penggunaan *transfer learning* adalah yang paling banyak ditemukan. *Transfer learning* menggunakan pengetahuan yang diperoleh dari sebuah *source task* untuk membantu menyelesaikan sebuah *target task* yang memiliki sedikit data untuk melatih model [13]. Sehingga dengan menggunakan *transfer learning*, data latih yang dibutuhkan tidak terlalu banyak.

Tabel 7. Pengembangan Arsitektur *Deep Learning*

Pengembangan Arsitektur *Deep Learning*	**Referensi**
Adversarial training	[7][12]
Active learning	[11]
Transfer learning	[3][9][12][13]
Adversarial transfer learning	[12]
Self-matching attention	[8]
Self-attention mechanism	[8][12]
Multitask learning	[3]

IV. NAMED ENTITY RECOGNITION PADA DOMAIN WISATA

Pada bagian ini akan dibahas mengenai hasil kajian terhadap literatur-literatur yang melakukan proses NER pada domain wisata. Literatur NER pada domain wisata yang dikaji pada penelitian ini menggunakan *dataset*, pendekatan, dan pelabelan entitas yang berbeda antara satu sama lain seperti yang tertera pada Tabel 8. Semua penelitian ini menggunakan *dataset* yang dibuat sendiri dengan mengumpulkan data yang sebagian besar berasal dari *website*. Entitas pada [14] dan [15] masih sangat umum dan sedikit jika dibandingkan dengan [18].

Terdapat dua langkah utama dalam proses ekstraksi informasi pada [14], yang pertama adalah membuat data pelatihan, dan yang kedua adalah membangun model pengenalan. 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 melalui *spaCy* menghasilkan akurasi sekitar 91% untuk data latih dan data uji. Sedangkan untuk BERT, akurasi yang didapatkan adalah sekitar 70%.

Penelitian dengan metode BERT juga dilakukan pada [15]. Namun pada penelitian tersebut, BERT dikombinasikan dengan metode BLSTM dan CRF yang selanjutnya 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*.

Metode *semi-supervised* NER digunakan pada [16]. 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. *Dataset* yang digunakan terdiri dari lima kelas entitas. Kelas *NATURAL* terdiri dari tempat wisata alam seperti pegunungan, pulau, sungai, dan pantai. Kelas *PLACE* digunakan untuk mengindikasikan tempat yang spesifik

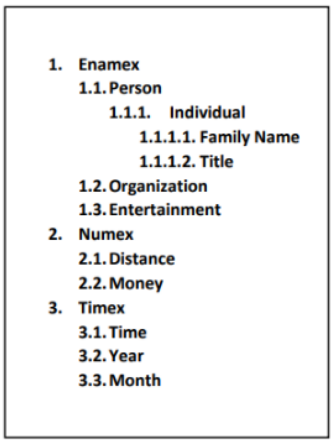
seperti candi, bandar udara, tempat ibadah, museum, dan lainnya. Kelas *CITY* terdiri dari lokasi administratif seperi kota dan distrik. Kelas *REGION* ditujukan untuk wilayah yang terpisah secara geografis, seperti sekumpulan kota atau negara, contohnya adalah ASEAN. Sedangkan kata yang tidak termasuk pada empat kelas yang sudah disebutkan sebelumnya akan masuk ke dalam kelas *NEGATIVE*. Secara keseluruhan sistem yang diusulkan memperoleh F1-*Score* sebesar 69,1%.

Tabel 8. Penelitian sebelumnya terkait *named entity recognition* pada domain wisata

Referensi	Data yang digunakan	Nama Entitas	Pendekatan NER
[14]	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.	LOC (lokasi atau nama tempat), ORG (hotel atau nama akomodasi), FACILITY (fasilitas)	*spaCy*, BERT
[15]	Menggunakan data teks perjalanan yang diambil dari *Ctrip* dan *Mafengwo, Raiders*.	*Person, location, organization, time, thing, other*.	BBLC (BERT – BLSTM – CRF)
[16]	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.	*Nature, city, region, negative*.	*Naïve Bayes Classifier*
[17]	Korpus paralel yang terdiri dari bahasa Inggris dan bahasa India yang dikumpulkan dengan menggunakan *web crawler*, serta berasal dari domain *travel, tourism*, dan *culture*.	PE (nama orang), OE (nama organisasi), LE (nama lokasi)	Algoritma *phonetic matching, rule-based*
[18]	Korpus yang berisi 94.000 kata berbahasa Tamil yang dikumpulkan pada domain wisata.	Terdapat 106 label yang digunakan pada penelitian ini dan saling terkait satu sama lain secara hierarki.	CRF

Pendekatan yang dilakukan pada [17] tidak bergantung pada bahasa yang digunakan dan hanya membutuhkan seperangkat aturan untuk suatu bahasa. Peningkatan akurasi dapat dilakukan dengan menambahkan lebih banyak aturan umum. Namun, sistem yang dibangun masih harus diperbaiki agar dapat mengenali frasa nama entitas dan juga singkatan.

Pelabelan pada [18] menggunakan pelabelan bersarang (*nested*), dengan total terdapat 106 label di dalamnya. Data pelatihan dilatih menggunakan model CRF dan hasil F1-*Score* yang didapat secara keseluruhan sebesar 80,44%. 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 pelabelan bertingkat dapat dilihat pada Gambar 1.



Gambar 1. Contoh pelabelan bertingkat

Dari literatur yang dikaji pada penelitian ini, 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.

V. KESIMPULAN

Penelitian ini mengkaji 11 literatur mengenai NER dengan *deep learning* dalam dua tahun terakhir (2018-2020) dan 5 literatur mengenai NER pada domain wisata yang didapatkan melalui pencarian pada *Google Scholar*. Dari hasil kajian didapatkan informasi bahwa dalam dua tahun terakhir, penggunaan gabungan metode BLSTM dan CRF sangat populer, namun dalam pengembangan arsitektur *deep learning* terdapat beberapa perbedaan. *Deep learning* dengan *transfer learning* adalah yang paling banyak digunakan. Dari kajian literatur, ekstraksi fitur *Position-Dependent Entity Type* (PDET) dengan metode BLSTM+CRF terbukti memiliki performa yang cukup baik untuk melakukan NER. Penelitian NER pada domain wisata saat ini menggunakan *dataset* yang dibuat sendiri oleh para peneliti. Pelabelan yang digunakan terbagi menjadi dua jenis, yaitu *flat* NER dan *nested* NER.

REFERENSI

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

[2] 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.

[3] X. Dong *et al.*, "Deep learning for named entity recognition on Chinese electronic medical records: Combining deep transfer learning with multitask bi-directional LSTM RNN," *PLoS One*, vol. 14, no. 5, pp. 1–15, 2019, doi: 10.1371/journal.pone.0216046.

[4] Q. Wang, Y. Zhou, T. Ruan, D. Gao, Y. Xia, and P. He, "Incorporating dictionaries into deep neural networks for the Chinese clinical named entity recognition," *J. Biomed. Inform.*, vol. 92, no. February, p. 103133, 2019, doi: 10.1016/j.jbi.2019.103133.

[5] H. A. Nayel and H. L. Shashirekha, "Integrating Dictionary Feature into A Deep Learning Model for Disease Named Entity Recognition," *arXiv*, pp. 1–16, 2019.

[6] I. El Bazi and N. Laachfoubi, "Arabic named entity recognition using deep learning approach," *Int. J. Electr. Comput. Eng.*, vol. 9, no. 3, pp. 2025–2032, 2019, doi: 10.11591/ijece.v9i3.pp2025-2032.

[7] S. Zhao, Z. Cai, H. Chen, Y. Wang, F. Liu, and A. Liu, "Adversarial training based lattice LSTM for Chinese clinical named entity recognition," *J. Biomed. Inform.*, vol. 99, no. August, p. 103290, 2019, doi: 10.1016/j.jbi.2019.103290.

[8] X. Cai, S. Dong, and J. Hu, "A deep learning model incorporating part of speech and self-matching attention for named entity recognition of Chinese electronic medical records," *BMC Med. Inform. Decis. Mak.*, vol. 19, no. Suppl 2, 2019, doi: 10.1186/s12911-019-0762-7.

[9] D. S. Sachan, P. Xie, M. Sachan, and E. P. Xing, "Effective use of bidirectional language modeling for transfer learning in biomedical named entity recognition," *arXiv*, pp. 1–19, 2018.

[10] M. G. Sohrab and M. Miwa, "Deep exhaustive model for nested named entity recognition," *Proc. 2018 Conf. Empir. Methods Nat. Lang. Process. EMNLP 2018*, pp. 2843–2849, 2020.

[11] Y. Shen, H. Yun, Z. C. Lipton, Y. Kronrod, and A. Anandkumar, "Deep active learning for named entity recognition," *arXiv*, pp. 1–15, 2018, doi: 10.18653/v1/W17-2630.

[12] P. Cao, Y. Chen, K. Liu, J. Zhao, and S. Liu, "Adversarial transfer learning for Chinese named entity recognition with self-attention mechanism," *Proc. 2018 Conf. Empir. Methods Nat. Lang. Process. EMNLP 2018*, pp. 182–192, 2020, doi: 10.18653/v1/d18-1017.

[13] J. A. Kosasih and M. L. Khodra, "Transfer Learning for Indonesian Named Entity Recognition," in *Proceeding - 2018 International Symposium on Advanced Intelligent Informatics: Revolutionize Intelligent Informatics Spectrum for Humanity, SAIN 2018*, Mar. 2018, pp. 173–178, doi: 10.1109/SAIN.2018.8673345.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] Y. Wibisono and M. L. Khodra, "Pengenalan Entitas Bernama Otomatis untuk Bahasa Indonesia dengan Pendekatan Pembelajaran Mesin," 2018, doi: 10.31227/osf.io/vud2p.

Kajian Literatur Named Entity Recognition pada Domain Artikel Wisata

ORIGINALITY REPORT

11%	10%	9%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Honglei Liu, Yan Xu, Zhiqiang Zhang, Ni Wang, Yanqun Huang, Yanjun Hu, Zhenghan Yang, Rui Jiang, Hui Chen. "A Natural Language Processing Pipeline of Chinese Free-text Radiology Reports for Liver Cancer Diagnosis", IEEE Access, 2020 Publication	1%
2	repub.eur.nl Internet Source	1%
3	ijcsi.org Internet Source	1%
4	gssrr.org Internet Source	1%
5	www.ijrte.org Internet Source	1%
6	Ziqi Lin, Haidong Zhang, Wancheng Ni, Yiping Yang. "Progressive Joint Framework for Chinese Question Entity Discovery and Linking	1%

With Question Representations", IEEE Access, 2019
Publication

7	Sufen Wang, Minmin Pang, Changqing Pan, Junyi Yuan, Bo Xu, Ming Du, Hong Zhang. "Information Extraction for Intestinal Cancer Electronic Medical Records", IEEE Access, 2020 Publication	1%
8	e-journal.stmiklombok.ac.id Internet Source	1%
9	tel.archives-ouvertes.fr Internet Source	<1%
10	stay.eu.com Internet Source	<1%
11	www.lib.unair.ac.id Internet Source	<1%
12	www.scilit.net Internet Source	<1%
13	Donghyeon Kim, Jinhyuk Lee, Chan Ho So, Hwisang Jeon, Minbyul Jeong, Yonghwa Choi, Wonjin Yoon, Mujeen Sung, Jaewoo Kang. "A Neural Named Entity Recognition and Multi-Type Normalization Tool for Biomedical Text Mining", IEEE Access, 2019 Publication	<1%

14	www.groundai.com Internet Source	<1%
15	www.hindawi.com Internet Source	<1%
16	Xiao Li, Wenteng Liang, Yifeng Li, Yuxuan Zhao, Zhizheng Zhang, Hengguang Yang. "RoBERTa Word Embedding Based Power Grid Dispatching Entity Recognition", 2020 12th IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC), 2020 Publication	<1%
17	export.arxiv.org Internet Source	<1%
18	www.coursehero.com Internet Source	<1%
19	Xishuang Dong, Uboho Victor, Lijun Qian. "Two-Path Deep Semisupervised Learning for Timely Fake News Detection", IEEE Transactions on Computational Social Systems, 2020 Publication	<1%
20	Chantana Chantrapornchai, Aphisit Tunsakul. "Information Extraction based on Named Entity for Tourism Corpus", 2019 16th International Joint Conference on Computer Science and Software Engineering (JCSSE), 2019 Publication	<1%

21	Abhishek Pradhan, Ketan Kumar Todi, Anbarasan Selvarasu, Atish Sanyal. "Knowledge Graph Generation with Deep Active Learning", 2020 International Joint Conference on Neural Networks (IJCNN), 2020 Publication	<1%
22	moam.info Internet Source	<1%
23	amenroom.com Internet Source	<1%
24	pembelajaran-mas-dewantara.blogspot.com Internet Source	<1%
25	id.123dok.com Internet Source	<1%
26	thesis.umy.ac.id Internet Source	<1%
27	Ismail El Bazi, Nabil Laachfoubi. "Arabic named entity recognition using deep learning approach", International Journal of Electrical and Computer Engineering (IJECE), 2019 Publication	<1%
28	Mohammad Al-Smadi, Saad Al-Zboon, Yaser Jararweh, Patrick Juola. "Transfer Learning for Arabic Named Entity Recognition With Deep Neural Networks", IEEE Access, 2020 Publication	<1%

29	dblp.uni-trier.de Internet Source	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		