Pengenalan Entitas Bernama Otomatis untuk Bahasa Indonesia dengan Pendekatan Pembelajaran Mesin

Yudi Wibisono
Ilmu Komputer Universitas Pendidikan Indonesia
yudi@upi.edu

Masayu Leylia Khodra
Teknik Informatika Institut Teknologi Bandung
masayu@stei.itb.ac.id

ABSTRAK

*Pengenalan entitas bernama (*named-entity recognition *atau NER) adalah proses otomatis mengekstraksi entitas bernama yang dianggap penting di dalam sebuah teks dan menentukan kategorinya ke dalam kategori terdefinisi. Sebagai contoh, untuk teks berita, NER dapat mengekstraksi nama orang, nama organisasi, dan nama lokasi. NER bermanfaat dalam berbagai aplikasi analisis teks, misalnya pencarian, sistem tanya jawab, peringkasan teks dan mesin penerjemah. Tantangan utama NER adalah penanganan ambiguitas makna karena konteks kata pada kalimat, misalnya kata "Cendana" dapat merupakan nama lokasi (Jalan Cendana), atau nama organisasi (Keluarga Cendana), atau nama tanaman. Tantangan lainnya adalah penentuan batas entitas, misalnya "[Istora Senayan] [Jakarta]". Berbagai kakas NER telah dikembangkan untuk berbagai bahasa terutama Bahasa Inggris dengan kinerja yang baik, tetapi kakas NER bahasa Indonesia masih memiliki kinerja yang belum baik. Makalah ini membahas pendekatan berbasis pembelajaran mesin untuk menghasilkan model NER bahasa Indonesia. Pendekatan ini sangat bergantung pada korpus yang menjadi sumber belajar, dan teknik pembelajaran mesin yang digunakan. Teknik yang akan digunakan adalah LSTM - CRF* (Long Short Term Memory – Conditional Random Field). *Hasil terbaik (F-measure = 0.72) didapatkan dengan menggunakan word embedding GloVe Wikipedia Bahasa Indonesia.*

Kata Kunci: NER, entitas bernama, pembelajaran mesin, analisis teks

PENDAHULUAN

Pengenalan entitas bernama (*named-entity recognition* atau NER) memiliki peranan penting dalam task pencarian informasi (Khalid, 2008) yang digunakan mesin pencari seperti Google dan Bing. NER juga bermanfaat dalam sistem tanya-jawab otomatis atau *chatbot* (Adam, 2012) seperti yang digunakan Apple Siri, Amazon Alexa, Siri, Google Home. NER secara otomatis mengindentifikasi bagian teks yang dianggap sebagai entitas penting, seperti nama orang, nama organisasi dan nama lokasi. Sebagai contoh, pada kalimat berikut: "`Dirut Telkom Arwin Rasyid bertemu wartawan detikINET di Gedung Telkom Graha Cipta Caraka`", NER akan mengenali "`Telkom`" sebagai nama organisasi, "`Arwin Rasyid`" sebagai nama orang, "`detikINET`" sebagai nama organisasi, dan "`Gedung Telkom Graha Cipta Caraka`" sebagai nama lokasi. Dalam contoh ini, jika manusia melontarkan pertanyaan kepada sistem "`Di mana Arwin dan wartawan detikINET bertemu ?`", maka sistem dapat menjawab dengan akurat karena telah memiliki informasi lokasi.

Masalah utama pada NER adalah ambiguitas. Pada contoh ini, kemunculan kata "`Telkom`" pertama memiliki arti berbeda dengan kemunculan kata "`Telkom`" kedua. Kata pertama adalah nama sebuah perusahaan di Indonesia sedangkan kata kedua adalah nama sebuah gedung. Manusia memiliki kemampuan yang mampu membedakan arti kedua kata tersebut, tetapi tidak demikian dengan program komputer. Beberapa contoh ambiguitas lain dapat dilihat pada Tabel 1. Kesalahan penulisan, seperti tidak menggunakan tanda baca dan huruf kapital yang tepat, menambah tingkat kesulitan pengenalan entitas bernama.

Tabel 1 Contoh-contoh ambiguitas dalam NER

Contoh kalimat	Keterangan
Conte menginginkan Vidal di **Chelsea**	Chelsea bukan kota tetapi organisasi (klub sepakbola)
Akhir pekan lalu tersebar rumor di **Wall Street**	Wall Street bukan lokasi, tetapi pasar saham Amerika
Hubungan PDIP dengan **Istana** semakin dekat.	Istana bukan nama orang, bukan lokasi, tetapi presiden.
Muktamar itu digelar oleh kepengurusan hasil Mukmatar **Bandung**	Bandung bukan nama lokasi tetapi salah satu versi muktamar.
Badrodin mengatakan **operasi bersinar** ini merupakan perintah Presiden.	Walaupun penulis salah menggunakan huruf kecil, operasi bersinar adalah nama kegiatan.

Masalah lain dalam NER adalah penentuan batas frasa. Misalnya pada kalimat berikut yang diambil dari berita online: "`CEO PT Cyrus Nusantara Hasan Nasbi menyerahkan uang Rp 1,4 miliar kepada KPK`", NER harus dapat mengenali "`PT Cyrus Nusantara`" dan "`Hasan Nasbi`". Tidak adanya tanda baca membuat batasan antara nama organisasi dengan nama CEO menjadi tidak jelas.

Pendekatan umum NER adalah pendekatan berbasis aturan yang memerlukan pendefinisian banyak aturan secara manual. Misalnya dapat dibuat aturan bahwa kata atau frasa setelah "di" kemungkinan besar adalah lokasi, kecuali untuk kasus-kasus tertentu. Tetapi cara ini memerlukan waktu banyak dan sulit untuk mencakup semua kasus.

Sejalan dengan data yang semakin mudah diperoleh membuat teknik pembelajaran mesin untuk NER berkembang. Pembelajaran mesin membuat model probabilitas berdasarkan data yang diketahui labelnya untuk memprediksi. Teknik pembelajaran mesin yang digunakan adalah Hidden Markov Model, Decision Tree, Maximum Entropy, Support Vector Machine, Conditional Random Field (Nadeau, 2007). Kelemahan dari pendekatan ini adalah tetap diperlukan pemilihan fitur secara manual dan pengetahuan tentang domain. Untuk mengatasi ini, berkembang teknik *deep learning* yang berbasis jaringan syaraf tiruan. Deep learning menggunakan banyak lapisan dan beberapa teknik yang membuat fitur dapat diplajari secara otomatis. Teknik bidirectional LSTM - Conditional Random Fields (CRF) memperoleh hasil tertinggi untuk NER Bahasa Inggris (Lample dkk., 2016). Makalah ini mengeksplorasi pendekatan ini untuk Bahasa Indonesia.

Penelitian NER untuk Bahasa Indonesia masih terbatas. Wibawa (2016) meneliti NER untuk 15 kelas entitas bernama pada 457 berita dengan teknik ensembled dan mendapatkan F-Measure tertinggi 0.50 Budi (2015) menggunakan pendekatan aturan yang dibuat manual untuk tiga kelas (nama orang, lokasi, organisasi) dengan F-Measure tertinggi 0.67

TEORI & METODOLOGI

Di dalam pembelajaran mesin, NER dapat dianggap sebagai masalah *sequence labeling*. *Sequence labeling* adalah memberikan urutan label pada objek yang berurutan. Selain untuk bidang bahasa *sequence labeling* juga digunakan dalam bidang biologi komputasional, misalnya untuk mendeteksi urutan proses mitosis (Liu dkk, 2010).

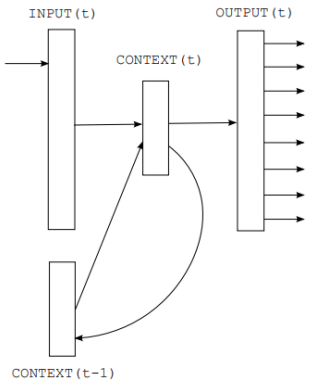
Sequence labeling memerlukan pelabelan data (encoding). NER menggunakan format standard BIO (Begin, Inside, Other). Jadi untuk nama orang, akan dilabeli dengan B-PER dan I-PER, nama organisasi dengan B-ORG dan I-ORG dan entitas lain seperti judul akan dilabeli dengan B-MISC I-MISC. Tabel 2 memperlihatkan contoh pelabelan kalimat " Eks bek Barcelona, Eric Abidal mengkritik bek Paris Saint-Germain " dengan format BIO.

Tabel 2 Contoh pelabelan dengan format BIO

Token	Eks	bek	Barcelona	,	Eric	Abidal	mengkritik	bek	Paris	Saint-Germain
Label	O	O	B-ORG	O	B-PER	I-PER	O	O	B-ORG	I-ORG

LSTM (Long Short Term Memory) network adalah pengembangan dari RNN (Recurrent Neural Network). Pada jaringan syaraf tiruan klasik, jaringan tidak dapat menyimpan informasi dari proses sebelumnya. Jika diterapkan pada pemrosesan bahasa, saat jaringan memproses suatu kata pada kalimat, maka kata-kata yang muncul sebelumnya akan "dilupakan", tentu ini akan mengurangi kinerja sistem. RNN mengatasi hal ini dengan mengunakan loop. Jaringan untuk kata yang muncul sebelumnya akan diberikan pada jaringan kata berikutnya. Gambar 1 memperlihatkan arstektur RNN.

Kelemahan dari RNN adalah tidak dapat mempelajari informasi yang terlalu jauh atau *long-term dependencies*. LSTM mengatasi ini dengan menambahkan pemrosesan pada repeating module. Jika pada RNN repeating module hanya terdiri dari lapisan aktiviasi sederhana, LSTM menggunakan beberapa lapisan dan gerbang yang memungkinkan proses yang lebih kompleks. Ini memungkinan LSTM menambah informasi untuk "memperkuat" atau menghapus informasi untuk "melupakan". *Bidirectional* LSTM mengembangkan lebih lanjut dengan menggunakan tambahan sel untuk arah sebaliknya. Jadi tidak hanya kata sebelum yang diperhitungkan, tapi kata sesudahnya (*future context*).



Gambar 1 Arsitektur RNN (Mikolov dkk, 2010)

Teknik pembelajaran mesin terbaik yang digunakan untuk *sequence labeling* sebelum teknik *deep learning* adalah CRF (Conditional Random Field). CRF memperhitungkan label yang muncul sebelummya dan setiap fitur diberi bobot untuk menghitung probabilitas label berikutnya.

Makalah ini menggunakan pendekatan yang digunakan Lample (2016) yang menggunakan *bidirectional* LSTM yang dikombinasikan dengan CRF. *Library* yang digunakan adalah anaGO[1]. Input dari sistem adalah *word embedding. Word embedding* adalah pemetaan dari kata ke dalam vector berdasarkan distribusinya.

Ada dua tahap eksperimen, pertama adalah mengumpulkan dan memberikan label data latih. Kedua adalah membuat model klasifikasi berdasarkan data latih. Pada tahap pertama, data NER yang diperoleh[2] diperbaiki dan dikonversi ke format BIO dan dipisahkan menjadi data latih, validasi dan tes (Tabel 3). Tabel 4 memperlihatkan rincian jumlah untuk setiap label. Dapat dilihat jumlah label O (Other) jauh lebih besar dibandingkan label lain. Hal ini akan mempersulit pembuatan model.

Tabel 3 Jumlah kalimat di dalam korpus

	Training	Validasi	Tes	Total
Kalimat	1253	420	419	2092

Tabel 4 Label BIO corpus

BIO	Training		Validasi		Tes		Total	
	Frek	%	Frek	%	Frek	%	Frek	%
B-PER	1080	4,03%	343	3,75%	438	4,83%	1861	4,14%
I-PER	540	2,02%	164	1,79%	214	2,36%	918	2,04%
B-ORG	1245	4,65%	430	4,70%	296	3,26%	1971	4,38%
I-ORG	732	2,73%	226	2,47%	151	1,66%	1109	2,47%
B-LOC	537	2,01%	162	1,77%	218	2,40%	917	2,04%
I-LOC	320	1,20%	99	1,08%	141	1,55%	560	1,24%
B-MISC	420	1,57%	173	1,89%	141	1,55%	734	1,63%
I-MISC	415	1,55%	178	1,95%	154	1,70%	747	1,66%
O	21485	80,25%	7369	80,59%	7318	80,67%	36172	80,40%
Total	26774	100,00%	9144	100,00%	9071	100,00%	44989	100,00%

Setelah data disiapkan, maka akan dilanjutkan dengan pembuatan model dapat dimulai untuk mencari hyperparameter yang terbaik. Untuk *word embedding*, akan dicoba menggunakan data corpus, data wikipedia Bahasa Indonesia dengan Word2Vec (Mikolov 2013) dan GloVe (Pennington 2014). Pengukuran kinerja sistem dilakukan dengan F-measure.

[1] https://github.com/Hironsan/anago
[2] https://github.com/yohanesgultom/

TEMUAN & PEMBAHASAN

Tabel 5 memperlihatkan konfigurasi hyperparameter dengan hasil terbaik dengan nilai F-measure terbaik 0.73. Kinerja turun jika CRF tidak digunakan. Pengaruh paling besar didapat dengan mengatur parameter batch_size dan learning_rate.

Tabel 5 Konfigurasi hyperparameter terbaik

Hyperparameter
char_emb_size=100, word_emb_size=50, char_lstm_units=25, word_lstm_units=100, dropout=0.5, char_feature=True, crf=True, batch_size=3, optimizer='adam', learning_rate=0.005, lr_decay=0.7, clip_gradients=5.0.

Efek penggunaan berbagai jenis *word embedding* dapat dilihat pada Tabel 6. Penggunaan pre-trained *word embedding* akan meningkatkan kinerja karena struktur yang terdapat di Wikipedia Bahasa Indonesia ikut dimanfaatkan. GloVe memberikan hasil tertinggi dengan ukuran dimensi 50. Saat ukuran dimensi diperbesar, nilai F-Measure justru turun.

Tabel 6 Pengaruh word embedding

Jenis Word Embedding	F-Measure	Ukuran Vector
training corpus data training	0.69	50
Word2Vec Wikipedia Bahasa Indonesia	0.69	400
Word2Vec Wikipedia Bahasa Indonesia	0.70	50
GloVe Wikipedia Bahasa Indonesia	0.73	50

Tabel 7 memperlihatkan ouput sistem. Dengan F-measure = 0.73, masih terdapat beberapa kesalahan tagging. Pada kalimat pertama, frase "Budi Martami" hanya dikenali "Martami" saja untuk label PER. Pada kalimat ke-4 dan ke-5, frase "Chelsea" dan "Wall Street" masih dikenali sebagai LOC, bukan ORG. Pada kalimat terakhir frase "operasi bersinar" tidak berhasil dikenali sebagai MISC.

Tabel 7 Contoh output sistem

Contoh kalimat	Output sistem
Budi Martami kuliah di UPI yang berlokasi di Bandung	Budi [Martami] kuliah di [UPI] yang berlokasi di [Bandung] PER: Martami LOC: UPI, Bandung
PDIP yang dikawal Megawati menang dalam Pilkada DKI Jakarta	[PDIP] yang dikawal [Megawati] menang dalam [Pilkada DKI Jakarta] ORG: PDIP PER: Megawati MISC: Pilkada DKI Jakarta
Badrodin mengatakan operasi bersinar ini merupakan perintah Presiden.	[Badrodin] mengatakan operasi bersinar ini merupakan perintah [Presiden]. PER: Badrodin MISC: Presiden
Conte menginginkan Vidal di Chelsea	[Conte] menginginkan [Vidal] di [Chelsea] PER: Conte, Vidal LOC: Chelsea
Akhir pekan lalu tersebar rumor di Wall Street	Akhir pekan lalu tersebar rumor di [Wall Street] LOC: Wall Street
Hubungan PDIP dengan Istana semakin dekat.	Hubungan [PDIP] dengan [Istana] semakin dekat. ORG: PDIP, Istana
Muktamar itu digelar oleh kepengurusan hasil Mukmatar Bandung	Muktamar itu digelar oleh kepengurusan hasil [Mukmatar Bandung] MISC: Mukmatar Bandung
Badrodin mengatakan operasi bersinar ini merupakan perintah Presiden.	[Badrodin] mengatakan operasi bersinar ini merupakan perintah [Presiden]. PER: Badrodin MISC: Presiden

KESIMPULAN & SARAN

Makalah ini mengembangkan model NER dengan arsitektur BiLSTMs-CRF dengan melihat pengaruh tiga jenis *word embedding*. Nilai F-measure terbaik 0.73. Untuk meningkatkan kinerja model NER, masih perlu dilakukan peningkatan ukuran korpus berlabel, dan mengevaluasi kinerja dibandingkan dengan teknik lain.

DAFTAR PUSTAKA

Khalid, M.A., Jijkoun, V. and De Rijke, M., 2008, March. The impact of named entity normalization on information retrieval for question answering. In *European Conference on Information Retrieval* (pp. 705-710). Springer, Berlin, Heidelberg.

Adam, Mitchell, M., & Sproat, R. (2012, June). Discourse-based modeling for aac. In *Proceedings of the Third Workshop on Speech and Language Processing for Assistive Technologies* (pp. 9-18). Association for Computational Linguistics.

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1), 3-26.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. 2016. *Neural architectures for named entity recognition*. arXiv preprint arXiv:1603.01360.

Wibawa, A. S., & Purwarianti, A. (2016). Indonesian named-entity recognition for 15 classes using ensemble supervised learning. *Procedia Computer Science*, *81*, 221-228.

Budi, I., Bressan, S., Wahyudi, G., Hasibuan, Z. A., & Nazief, B. A. (2005, October). Named entity recognition for the Indonesian language: combining contextual, morphological and part-of-speech features into a knowledge engineering approach. In *International Conference on Discovery Science* (pp. 57-69). Springer, Berlin, Heidelberg.

Liu, A. A., Li, K., & Kanade, T. (2010, April). Mitosis sequence detection using hidden conditional random fields. In *Biomedical Imaging: From Nano to Macro, 2010 IEEE International Symposium on*(pp. 580-583). IEEE.

Mikolov, T., Karafiát, M., Burget, L., Černocký, J., & Khudanpur, S. (2010). Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*.

Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." *Advances in neural information processing systems*. 2013.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

RIWAYAT HIDUP

Nama Lengkap	Institusi	Pendidikan	Minat Penelitian
Yudi Wibisono	UPI	S3	Linguistik Komputasional
Masayu Leylia Khodra	ITB	S3	Linguistik Komputasional, Pembelajaran Mesin, Inteligensi Buatan