**LAPORAN PROYEK**

***Indonesian Named-entity Recognition Using BiLSTM-CRF***



**Disusun oleh:**

1. 12S18014 – Giovanni Situmorang
2. 12S18015 – Della Cenovita Tarigan
3. 12S18022 – Alex Conro Manuel
4. 12S18042 – Indah Oktavia M. Sibarani
5. 12S18051 – Cindy Juniati Hutapea
6. 12S18062 – Ester Putri Dearest Sidabutar

**11S4037 – PEMROSESAN BAHASA ALAMI**

**FAKULTAS INFORMATIKA DAN TEKNIK ELEKTRO**

**INSTITUT TEKNOLOGI DEL**

**2021**

**DAFTAR ISI**

**BAB 1 PENDAHULUAN**

Pada bagian ini menyajikan latar belakang, tujuan, manfaat, dan ruang lingkup pengerjaan proyek

**1.1 Latar Belakang**

*Named Entity Recognition (NER)* adalah salah satu topik penelitian di bidang *Natural Language Processing (NLP)*. Penelitian ini bertujuan untuk mengidentifikasi entitas bernama (*named* *entity*) kata di dalam kalimat [1]. Istilah "*Named Entity*" telah digunakan secara luas di bidang *information extraction (IE), question answering (QA)*, dan berbagai bidang aplikasi *natural language processing (NLP)* lainnya. Istilah ini pertama kali digunakan di *Message Understanding Conference-6 (MUC-6)* pada tahun 1995 [3]. Terdapat 2 proses pengenalan entitas bernama dalam NER, yaitu identifikasi entitas kata yang tepat di dalam dokumen dan pengklasifikasian entitas-entitas tersebut ke dalam beberapa klasifikasi entitas umum dalam sistem NER seperti Organisasi, Orang, Lokasi, Tanggal, Waktu, Mata Uang, Persentase, Fasilitas dan *Geo Political Entities (GPE)* [2]. Tujuan dari NER adalah untuk mengidentifikasi atau mengklasifikasi sebuah entitas misalnya nama orang, organisasi, waktu, lokasi dan sesuatu entitas lain dalam sebuah teks yang sangat berguna dalam kasus ekstraksi informasi [7].

Penerapan NER dalam bahasa Inggris menunjukkan hasil yang baik dikarenakan banyaknya data yang tersedia dalam melakukan penelitian. Namun, penerapan NER dalam bahasa Indonesia tidak begitu baik, hal ini mungkin saja disebabkan oleh sifat bahasa Indonesia yang bervariasi dalam banyak hal karena sejarah morfologi [6]. Bahasa Indonesia memiliki karakteristik yang berbeda jika dibandingkan dengan bahasa Inggris. Bahasa Indonesia tidak memiliki indikasi gender seperti dalam bahasa Inggris. Bingung menentukan apakah kata ganti orang ketiga dalam bahasa Indonesia mengacu pada laki-laki atau perempuan. Kata ganti orang ketiga dalam bahasa Indonesia adalah-nya dan ia/dia. Ia/dia mewakili dia, dan-nya mewakili dia [5].

Pada penelitian yang dilakukan oleh Min Zhang, dkk, Penelitian ini, bertujuan untuk mengenali beberapa jenis entitas peninggalan budaya, termasuk nama peninggalan budaya, dinasti peninggalan budaya), lokasi yang digali, dan koleksi museum, yang merupakan konsep kritis dalam penemuan pengetahuan peninggalan budaya. Namun, kurangnya data berlabel di bidang peninggalan budaya menyulitkan model pembelajaran mendalam yang mengandalkan data berlabel untuk mencapai kinerja yang sangat baik. Untuk mengatasi masalah ini, maka diusulkan model *deep learning semi supervised* bernama 'SCRNER (*Semi-supervised model for Cultural Relics' Named Entity Recognition*) yang memanfaatkan memori *bidirectional long short term* (BiLSTM ) dan model *conditional random fields* (CRF) yang dilatih oleh data yang jarang diberi label dan banyak data tidak berlabel untuk mencapai kinerja yang efektif. *Bidirectional Long Short-Term Memory and Conditional Random Fields* (BiLSTM-CRF) pada penelitian ini dimulai dari data teks terlebih dahulu diubah ke dalam bentuk angka (*word embedding*). Word embedding dilakukan dengan menggunakan pendekatan ELMo. Hasil dari *word embedding* kemudian akan digunakan untuk melatih BiLSTM-CRF. Metode *Bidirectional Long Short-Term Memory* (BiLSTM) memanfaatkan informasi kontekstual masa lalu dan masa depan. Namun, dalam *sequence labeling* BiLSTM tidak menggeneralisasi korelasi antara label keluaran karena distribusi probabilitas BiLSTM tidak bergantung satu sama lain. *Sequence labeling* bertujuan untuk memberikan urutan dari label pada objek yang berurutan. Dalam *sequence labeling,* penting untuk mempertimbangkan korelasi antara label yang berdekatan. Oleh karena itu, pada penelitian ini model *sequence labeling* mengadopsi *Conditional Random Field* (CRF) di lapisan atas untuk memprediksi label dari semua kalimat secara bersamaan. *Conditional Random Field* (CRF) memperhitungkan letak label yang muncul sebelumnya kemudian setiap fitur (label) akan diberikan bobot dengan tujuan menghitung setiap probabilitas untuk label berikutnya[4].

Secara umum, terdapat empat tantangan atau permasalahan NER dalam proses pengenalan entitas bernama [1], yaitu:

1. Deteksi batasan dari entitas bernama (NE *boundary detection*), jika entitas bernama berupa gabungan dua atau lebih kata.
2. Homonim, kata yang memiliki lafal dan ejaan yang sama dengan kata lain tetapi berbeda maknanya. Kedua kata ini memungkinkan untuk memiliki entitas bernama yang berbeda.
3. *Spelling mistakes*, biasanya permasalahan ini ada jika teks merupakan hasil dari forum diskusi online, bahasa yang tidak formal dan kurang terstruktur.
4. Struktur kalimat yang tidak baku: Struktur kalimat adalah penting untuk penandaan *part-of-speech*, yang digunakan untuk mendeteksi *frase* kata benda dari teks, untuk membantu dalam deteksi batas

Selain permasalah diatas, masalah yang sering muncul pada *named-entity recognition* adalah *Ambiguity*. Pertama, kata yang sama dapat berarti dua entitas yang berbeda. Misalnya kata Soekarno dapat berarti presiden pertama Indonesia, atau nama belakang seorang seniman (Enrico Soekarno), keduanya entitas berbeda walaupun tipenya sama (orang/*person*). Jenis ambiguitas kedua adalah nama yang sama tapi tipe berbeda. Contohnya adalah Bung Karno sebagai stadion dengan Bung Karno sebagai orang. Berdasarkan permasalahan tersebut maka dibutuhkan sebuah metode yang dapat mengatasi permasalahan diatas.

Pada penelitian ini, metode yang digunakan adalah metode Bidirectional LSTM-CRF. Bidirectional LSTM menggabungkan konteks sebelumnya dan konteks setelahnya dengan memproses data dari dua arah yang selanjutnya diklasifikasi menggunakan CRF [7]. *Bidirectional long short term* (BiLSTM) mempelajari bobot keluaran dari konteks sebelumnya dan masukan dari setiap urutan pada saat berlangsung[4]. Selain itu, informasi dari kata sebelumnya dan kata selanjutnya dari urutan kalimat dapat ditangkap secara bersamaan oleh *forward network* yang merepresentasikan konteks sebelumnyadan *backward network* yang merepresentasikan konteks setelahnyayang diikuti dengan memproses data dari dua arah dengan hidden layer, sehingga memperoleh informasi konteks dalam proses pemodelan urutan kalimat. *Conditional Random Field* merupakan sebuah model probabilistik yang digunakan untuk memprediksi data terstruktur yang telah digunakan dalam berbagai tugas, seperti *natural language processing* [7]. Model CRF melakukan training untuk memprediksi sebuah vektor dari sebuah kalimat. *Conditional random fields* (CRFs) telah terbukti efektif di banyak area pemrosesan bahasa alami (NLP), termasuk tugas tag urutan dan pengenalan entitas bernama (NER) [4]. Oleh karena itu, dengan mengkombinasikan pendekatan *bidirectional* LSTM dan model CRF dapat digunakan untuk menangkap semua informasi selama pemodelan urutan kalimat panjang. Berdasarkan uraian tersebut maka pada penelitian ini akan dilakukan Implementasi pendekatan *Bidirectional* LSTM-CRF pada kasus *Named Entity Recognition* dalam teks bahasa Indonesia dengan batasan data yang akan digunakan pada penelitian ini adalah dataset singgalang.

**1.2 Tujuan**

1. Membuat model *Indonesian Named Entity Recognition* yang dapat melakukan pengenalan pada teks bahasa indonesia serta mendeteksi dan mengklasifikasikan kata kedalam entitas yang seharusnya.
2. Untuk mengetahui tingkat akurasi penggunaan metode BiLSTM-CRF dalam membangun model NER untuk bahasa Indonesia.

**1.3 Manfaat**

Manfaat dari pengerjaan proyek ini yaitu:

1. Proyek ini diharapkan nantinya dapat memberi wawasan serta pengetahuan bagi tim proyek dalam menerapkan aplikasi pemrosesan bahasa alami khususnya *named entity recognition* pada bahasa Indonesia
2. Proyek ini diharapkan dapat digunakan dalam membantu mendeteksi informasi penting dan mengidentifikasi elemen kunci dalam teks dengan mudah dan tepat sesuai entitasnya, seperti nama orang, tempat, dan lainnya.

**1.4 Ruang Lingkup**

Ruang lingkup yang akan dibahas dalam proyek ini adalah model yang dibagun hanya bisa menerima *input* berupa kalimat Bahasa Indonesia, dimana dataset yang digunakan dalam proyek ini adalah dataset singgalang , dengan link berikut :

<https://github.com/ialfina/ner-dataset-modified-dee/tree/master/singgalang>.

**BAB 2**

**ISI**

Bab ini menjelaskan mengenai analisis yang akan dilakukan terhadap data dan metode, desain pemrosesan bahasa alami dan implementasi berupa kode program dan cuplikan hasil, dan hasil evaluasi kuantitatif terhadap implementasi pemrosesan bahasa alami yang dilakukan.

**2.1 Analisis**

Pada sub bab ini dijelaskan analisis yang dilakukan terhadap data dan metode yang digunakan dalam pengimplementasian pemrosesan bahasa alami.

**2.1.1 Analisis Data**

*Dataset* yang digunakan dalam proyek ini dapat dilihat pada link berikut: <https://github.com/ialfina/ner-dataset-modified-dee/tree/master/singgalang>.

**2.1.2 Analisis Metode**

**2.2 Implementasi**

**2.2.1 Data Preprocessing**

**2.2.2 Feature Extraction**

**2.2.3 Feature Selection CRF**

**2.2.4 Modeling with BiLSTM**

**2.3 Hasil**

**BAB 3**

**PENUTUP**

3.1 Pembagian Tugas dan Tanggung Jawab

3.2 Kesimpulan

3.3 Saran

# **DAFTAR PUSTAKA**

1. SUFA, M., Yusliani, N., & Buchari, M. A. (2020). NAMED-ENTITY RECOGNITION PADA TEKS BERBAHASA INDONESIA MENGGUNAKAN METODE HIDDEN MARKOV MODEL DAN PART-OF-SPEECH TAGGING (Doctoral dissertation, Sriwijaya University).
2. Aryoyudanta, B., Adji, T. B., & Hidayah, I. (2016, July). Semi-supervised learning approach for Indonesian Named Entity Recognition (NER) using co-training algorithm. In 2016 International Seminar on Intelligent Technology and Its Applications (ISITIA) (pp. 7-12). IEEE.
3. Aditya Satrya Wibawa, Ayu Purwarianti. (2016). Indonesian Named-entity Recognition for 15 Classes Using Ensemble Supervised Learning.
4. Zhang, M., Geng, G., & Chen, J. (2020). Semi-supervised bidirectional long short-term memory and conditional random fields model for named-entity recognition using embeddings from language models representations. Entropy, 22(2), 252.
5. I. Budi and S. Bressan, "Application of association rules mining to Named Entity Recognition and co-reference resolution for the Indonesian language," International Journal of Business Intelligence and Data Mining, vol. 2, no. 4, p. 426–446, 2007.
6. Gunawan, W., Suhartono, D., Purnomo, F., & Ongko, A. (2018). Named-entity recognition for indonesian language using bidirectional lstm-cnns. *Procedia Computer Science*, *135*, 425-432.
7. Permana, H. (2019). *Named Entity Recognition Menggunakan Metode Bidirectional Lstm-Crf Pada Teks Bahasa Indonesia* (Doctoral dissertation, Universitas Komputer Indonesia).