Sistem Penerjemah Bahasa Indonesia ke Bahasa Kaili

Fendi Irfan Amorokhman[1], Ade Romadhony[2], Aditya Firman Ihsan[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[1]fendiirfan@student.telkomuniversity.ac.id, [2]aderomadhony@telkomuniversity.ac.id,
[3]adityaihsan@telkomuniversity.ac.id

Abstrak

Bahasa Kailinese, yang digunakan di Provinsi Sulawesi Tengah, Indonesia, menghadapi tantangan karena penggunaan sehari-hari yang terbatas. Menurut artikel yang diterbitkan di situs brin.go.id, beberapa dialek Kailinese hanya memiliki empat penutur keluarga. Untuk mengatasi masalah ini, penelitian ini memperkenalkan model terjemahan mesin dan kumpulan data khusus untuk Kailinese. Sistem terjemahan kami menggunakan model IndoBART-V2, yang memungkinkan terjemahan yang lancar antara bahasa Indonesia dan Kailinese. Kami melakukan dua skenario pengujian: satu dengan kumpulan data ulasan yang beragam dan topik acak, dan satu lagi fokus pada kumpulan data tipe ulasan. Dengan menggunakan parameter default dan teknik pra-pemrosesan menggunakan "Kamus Leksikon Indonesia Kolokial," model terjemahan kami mencapai skor SacreBLEU yang mengesankan dibandingkan dengan tidak menggunakan pra-pemrosesan atau mengubah parameter default. Untuk skenario 1 (terjemahan dari bahasa Indonesia ke Kailinese), model ini mencapai skor 19,8, sedangkan untuk skenario 2 (terjemahan dari Kailinese ke bahasa Indonesia), skornya adalah 23,0. Pada skenario 2, di mana data latihan dan pengujian terdiri dari kalimat tipe ulasan, model ini mencapai skor 18,4 dan 22,7 untuk terjemahan dari bahasa Indonesia ke Kailinese dan dari Kailinese ke bahasa Indonesia, secara berturut-turut. Hasil ini menunjukkan efektivitas dan akurasi model yang dikembangkan. Selain itu, analisis kami mengungkapkan bahwa komposisi kalimat secara signifikan mempengaruhi kinerja model, tanpa perbedaan yang mencolok antara skenario 1 dan skenario 2. Hal ini menekankan pentingnya mempertimbangkan jenis kalimat dalam model terjemahan.

Kata kunci : machine translation, Indonesian language, kailinese language, indobart-v2

1. Pendahuluan

Latar Belakang

Indonesia adalah negara yang kaya sumber daya manusia, sumber daya alam, budaya, dan juga bahasa daerah. Indonesia memiliki lebih dari 700 bahasa daerah, salah satunya adalah bahasa Kai [1]. Bahasa Kailinese merupakan bahasa daerah yang tumbuh dan berkembang di Provinsi Sulawesi Tengah [2] dan memiliki beberapa dialek seperti Kailinese Ledo, Kailinese Unde, Kailinese Rai, dan Kailinese Daa [3], antara lain. Secara khusus, penutur bahasa Kailinese Ledo sering ditemukan di tengah lembah kota Palu, sementara yang lain biasanya berada di luar kota Palu.

Menurut sensus tahun 2000, komposisi etnis kota Palu meliputi suku Kailinese 33,3%, suku Bugis 24,4%, suku Jawa 10,1%, suku Gorontalo 3,1%, suku Bali 1,2%, dan lain-lain 24,9% [4]. Keragaman suku yang tinggal di kota Palu membuat penggunaan bahasa Kailinese sulit dalam kehidupan sehari-hari di kota Palu. Berdasarkan artikel di kependudukan.brin.go.id, bahasa Kailinese merupakan bahasa daerah yang terancam punah dan bahkan pada tahun 1978, jumlah penutur dialek Kailinese Ende hanya empat keluarga [5].

Berdasarkan data penggunaan bahasa Kailinese, dapat disimpulkan bahwa diperlukan upaya pelestarian dari semua pihak dengan menghidupkan kembali penggunaan bahasa Kailinese. Salah satu cara untuk melestarikan bahasa adalah dengan mendokumentasikannya dalam bentuk sistem terjemahan bahasa. Namun, saat ini belum ada sistem terjemahan kalimat lengkap yang tersedia untuk bahasa Kailinese. Oleh karena itu, penelitian ini berfokus pada pengembangan sistem yang dapat menerjemahkan bahasa Indonesia ke Kailinese dan Kailinese ke bahasa Indonesia, termasuk dataset bahasa Kailinese yang diperlukan.

Berdasarkan penelitian NusaX, model IndoBART-V2 memiliki kinerja terbaik berdasarkan metrik SacreBLEU sebagai model terjemahan untuk bahasa daerah ke bahasa Indonesia [6]. Oleh karena itu, penelitian ini menggunakan model IndoBART-V2 dan berfokus pada pengembangan dataset bahasa Kailinese untuk penelitian terjemahan mesin.

Penelitian ini mengatasi beberapa rumusan masalah dalam ranah penelitian akademik. Hal ini mencakup pembuatan dataset bahasa Kailinese yang sesuai untuk terjemahan mesin, pengembangan model terjemahan mesin bahasa Indonesia-Kailinese, dan evaluasi kinerja model IndoBART-V2 serta efektivitas dataset dalam terjemahan bahasa Kailinese. Kontribusi utama dari penelitian ini terletak pada produksi dataset bahasa Kailinese yang dapat diterjemahkan oleh mesin, implementasi sistem terjemahan mesin bahasa Indonesia-Kailinese dengan menggunakan model IndoBART-V2, dan penilaian kinerja model IndoBART-V2 serta kesesuaian dataset untuk terjemahan bahasa Kailinese. Penting untuk dicatat bahwa penelitian ini menghadapi

beberapa keterbatasan. Dataset yang digunakan berasal dari proyek penelitian NusaX dan kemudian diterjemahkan ke dalam bahasa Kailinese. Selain itu, dialek Ledo dalam bahasa Kailinese menjadi fokus khusus dalam investigasi ini.

Topik dan Batasannya

Set Data yang digunakan berasal dari penelitian NusaX yang akan translate kedalam bahasa Kaili. Bahasa Kaili yang digunakan merupakan bahasa Kaili dialek Ledo.

Tujuan

Menghasilkan set data Bahasa Kaili yang dapat yang dapat digunakan untuk mesin penerjemah. Mengimplementasikan mesin penerjemah Bahasa Kaili dengan menggunakan model IndoBARTv2. Mengetahui performa dari model IndoBARTv2 dan set data yang digunakan dalam penerjemahan Bahasa Kaili.

Organisasi Tulisan

Pada jurnal ini berisi bagian abstrak, pendahuluan, studi terkait, sistem yang dibangun, evaluasi, kesimpulan.

2. Studi Terkait

Beberapa penelitian telah fokus pada terjemahan bahasa, salah satunya dilakukan pada tahun 2018, dengan tujuan menerjemahkan kalimat-kalimat dari bahasa Lampung ke bahasa Indonesia [7]. Penelitian ini menggunakan metode terjemahan mesin neural berbasis perhatian (NMT). NMT adalah pendekatan terjemahan mesin yang menggunakan komponen encoder dan decoder untuk memahami seluruh kalimat dengan efektif dan menangkap informasi kontekstual, sehingga menghasilkan keluaran terjemahan. Dalam penelitian ini, metode NMT dievaluasi menggunakan BLEU sebagai metrik evaluasi.

Salah satu model yang umum digunakan dalam bidang terjemahan bahasa adalah model penggantian kata langsung, di mana setiap kata diterjemahkan berdasarkan kamus. Studi serupa yang menggunakan model penggantian kata untuk terjemahan telah dilakukan [6]. Namun, metode ini memiliki beberapa kelemahan, karena model tidak dapat memperoleh konteks suatu kalimat atau paragraf akibat berbagai makna yang dapat dimiliki oleh suatu kata dalam konteks yang berbeda. Perkembangan terjemahan mesin telah menghasilkan model-model yang lebih baru dan umum digunakan, seperti model IndoBART-V2. Model IndoBART-V2 adalah model yang telah dipretraining dari korpus bahasa Indonesia, Jawa, dan Sunda [8]. Model ini didasarkan pada arsitektur transformer standar mBART [9]. Model-model dengan dasar berbasis transformer dapat mengatasi kelemahan model-model sebelumnya, seperti ketergantungan sekuensial jangka panjang dan waktu pelatihan yang lama, sehingga membuat IndoBART-V2 sangat efisien dalam tugas terjemahan mesin.

2.1 Mesin Penerjemah

Terjemahan mesin adalah salah satu tantangan dalam bidang pemrosesan bahasa alami, yang bertujuan untuk memfasilitasi terjemahan otomatis dari satu bahasa ke bahasa lain [10]. Sistem terjemahan mesin menerima input berupa kalimat dalam bahasa tertentu dan kemudian menjalani proses terjemahan ke bahasa lain, yang akhirnya menghasilkan keluaran terjemahan yang diinginkan.

Terjemahan mesin memiliki beberapa pendekatan, termasuk Terjemahan Mesin Langsung (DMT), Terjemahan Mesin Berbasis Aturan (RBMT), dan pendekatan berbasis data [11]. Pendekatan DMT membutuhkan kamus dwibahasa, sedangkan RBMT erat kaitannya dengan aturan untuk menganalisis dan mengubah representasi bahasa sumber serta menghasilkan kalimat-kalimat dalam bahasa target.

2.2 Sistem Mesin Penerjemah Menggunakan Model IndoBART-V2

Pengembangan model terjemahan mesin telah berkembang pesat selama bertahun-tahun. Jaringan saraf, seperti RNN, LSTM, dan lainnya, adalah beberapa konsep awal yang digunakan dalam penelitian terjemahan mesin [12]. Jaringan saraf adalah metode yang terinspirasi dari cara kerja otak manusia, yang terdiri dari banyak neuron atau lapisan tersembunyi. Arsitektur jaringan saraf sangat bergantung pada bobot, input, dan output yang didefinisikan pada unit neuron [13]. Konsep lapisan tersembunyi, bobot, dan input/output digunakan sebagai dasar pembelajaran dalam jaringan saraf.

Metode pembaruan berikutnya adalah model perhatian. Model perhatian, juga dikenal sebagai mekanisme perhatian, adalah teknik pembelajaran mendalam yang fokus pada komponen-komponen khusus seperti perhatian diri antara input dan perhatian umum antara input dan output. Dalam terjemahan mesin, model perhatian mengevaluasi pentingnya kata-kata dalam sebuah kalimat dengan memberikan bobot.

Dengan diperkenalkannya model perhatian, penelitian lebih lanjut mengembangkan konsep tersebut menjadi arsitektur terkini di bidang Pemrosesan Bahasa Alami (NLP), yang dikenal sebagai arsitektur transformer, yang memiliki performa lebih baik dibandingkan dengan model-model sebelumnya seperti RNN.