Indonesian Journal of Multidiciplinary Expertise (IJME): Jurnal Multidisiplin Ilmu, 1(4),
Tahun 2023
e-ISSN: 3025-1583

Low-Resource Single-Domain Machine Translation untuk Bahasa Karo-Indonesia

Sengli Egani Sitepu[1]
Prodi. Bisnis Digital, Universitas Satya Terra Bhinneka, Medan, Indonesia
sengliegani@satyaterrabhinneka.ac.id

Abstract: *Indonesia has cultural and linguistic diversity that must be preserved. One way is to utilize currently developing technology, Neural Machine Translation, to build a translation machine. This research uses a Statistical Machine Translation (SMT) approach. This computer-based translation approach is an approach based on statistical methods introduced since the mid-20th century. The popularity of SMT arises due to a translation architecture based on strong mathematical theory, good quality translation in many test runs, and the existence of toolkits that can be used to build MT models in a short time. In this research, the author built a translation machine from Indonesian to Karo Batak language using the Karo regional language Bible domain (Pustaka Sibadia) as the Karo language corpus and the Contemporary Indonesian Bible (BIMK) as the Indonesian language corpus. The best blue score obtained in this study was 0.18. More in-depth research is needed to maximize the results*

Keywords: *Statistical machine translation, neural machine translation, Karo language*

Abstrak: Indonesia memiliki keanekaragaman budaya dan bahasa yang harus tetap dilestarikan. Salah satu caranya dengan memanfaatkan teknologi yang berkembang saat ini, Neural Machine Translation, untuk membangun mesin penerjemah. Penelitian ini menggunakan pendekatan Statistical Machine Translation (SMT). Pendekatan penerjemahan berbasis komputer ini adalah pendekatan yang didasarkan pada metode statistik yang diperkenalkan sejak pertengahan abad ke-20. Popularitas SMT muncul karena arsitektur terjemahan yang didasarkan pada teori matematika yang kuat, terjemahan berkualitas baik dalam banyak tes yang dijalankan, dan keberadaan toolkit yang dapat digunakan untuk membangun model MT dalam waktu singkat. Dalam penelitian ini, penulis membangun mesin penerjemah Bahasa Indonesia ke Bahasa batak Karo menggunakan domain Alkitab (Bible) berbahasa daerah Karo (Pustaka Sibadia) sebagai korpus berbahasa Karo dan Aklitab Bahasa Indonesia Masa Kini (BIMK) sebagai korpus berbahasa Indonesia. Hasil skor bleu terbaik yang berhasil diperoleh dalam penelitian ini yaitu 0.18. Diperlukan penelitian lebih dalam untuk memaksimalkan hasilnya

Kata Kunci: *Statistical machine translation, neural machine translation*, Bahasa Karo

Pendahuluan

Terjemahan mesin (bahasa Inggris: *machine translation*, MT) adalah suatu proses penerjemahan teks dari bahasa sumber ke bahasa lainnya (target) dengan menggunakan sebuah perangkat lunak. Saat ini penggunaan teknologi ini sudah semakin berkembang dan dapat digunakan oleh hampir semua orang dalam kesehariannya. Dalam perkembangannya saat ini, terdapat dua pendekatan yang paling canggih yaitu Terjemahan Mesin berbasis statistik (*Statistical Machine Translation*, SMT) dan Terjemahan Mesin berbasis neural (*Neural Machine Translation*, NMT). Terjemahan yang diproduksi oleh SMT berbasis pada analisis statistik dari korpora teks, sedangkan NMT menggunakan deep neural network untuk memodelkan dan menerjemahkan. Unit dasar pada penerjemahan berbasis statistik ialah kata sedangkan penerjemahan berbasis neural adalah sebuah vektor. NMT menggunakan word embedding untuk mengubah kata menjadi vektor sebelum diproses oleh NMT.

Kedua model ini merupakan pendekatan yang sangat bergantung pada data, yaitu pendekatan yang belajar dari pararel korpus teks untuk membangun model terjemahan. Meskipun memiliki kesamaan namun kedua model ini memiliki kekuatan dan kelemahannya masing-masing. SMT memiliki

tingkat fleksibilitas yang tinggi dan menghasilkan tejemahan yang akurat dalam berbagai skenario, sementara NMT–yang saat ini telah menjadi pendekatan dominan dalam penelitian dan praktik MT – sensitif terhadap ukuran data pelatihan. Dalam banyak penelitian, NMT memang banyak menggungguli SMT, namun hal tersebut terjadi dikarenakan ketersediaan korpora paralel yang besar. Saat digunakan pada data yang terbatas, SMT masih bisa menggungguli NMT.

Saat ini, sudah ada beberapa penelitian yang dilakukan untuk menguji kinerja dari SMT dan NMT. Penelitian yang dilakukan Luong et al., 2015, membahas dua kelas mekanisme attention yang sederhana dan efektif: pendekatan global yang selalu memperhatikan semua kata sumber dan pendekatan lokal yang hanya melihat sebagian kata sumber pada satu waktu. Penelitian ini menunjukkan efektivitas kedua pendekatan pada tugas terjemahan WMT antara bahasa Inggris dan Jerman dan sebaliknya. Dengan *local attention*, penelitian ini mencapai perolehan signifikan sebesar 5,0 poin BLEU dibandingkan sistem *non-attention*, salah satunya *phrase-based MT*. Model yang diteliti menggunakan arsitektur perhatian yang berbeda menghasilkan hasil mutakhir baru dalam tugas terjemahan Bahasa Inggris ke Bahasa Jerman WMT'15 dengan 25.9 poin BLEU.

Pada penelitian selanjutnya oleh Koehn et al., 2017, bertujuan untuk memeriksa sejumlah tantangan untuk NMT dan memberikan hasil empiris tentang seberapa baik teknologi saat ini bertahan, dibandingkan dengan SMT. Penelitian ini menunjukkan bahwa, terlepas dari keberhasilan NMT baru-baru ini, terjemahan mesin berbasis neural masih harus mengatasi berbagai tantangan, terutama kinerja di luar domain dan di bawah kondisi sumber daya yang rendah. Kesamaan dari banyak masalah adalah bahwa model terjemahan neural tidak menunjukkan perilaku yang kuat ketika dihadapkan dengan kondisi yang berbeda secara signifikan dari kondisi pelatihan.

Penelitian selanjutnya yang dilakukan oleh Ahmadnia et al., 2020, yang melakukan percobaan untuk menemukan model terbaik untuk terjemahan mesin berbasis multi-domain Bahasa Spanyol-Farsi dengan sumber daya yang minim. Penelitian ini menunjukkan portabilitas sistem MT ke beberapa domain di bawah kondisi sumber daya rendah untuk bahasa target, di mana data domain bahasa target tidak tersedia. Peneliti menyebut kasus ini sebagai pasangan bahasa dengan sumber daya rendah dwibahasa. Peneliti membandingkan kinerja sistem Neural MT (NMT) berbasis perhatian dengan sistem MT (SMT) Statistik berbasis frase dalam kondisi ini, melatih masing-masing pada korpora paralel yang terdiri dari domain yang berbeda. Hasil eksperimen pada Spanyol-Farsi sebagai pasangan bahasa sumber daya rendah bilingual menunjukkan bahwa paradigma SMT masih mengungguli NMT.

Dari ketiga penelitian tersebut, maka dapat disimpulkan bahwa suatu pendekatan belum tentu memberikan hasil yang lebih baik dibanding dengan pendekatan lainnya untuk semua kasus yang berbeda. Dalam penelitian ini, kasus digunakan berbeda dengan kasus dari ketiga penelitian sebelumnya, yaitu sumber data yang terbatas dan domain tunggal (spesifik). Oleh karena itu perlu

dilakukan pengujian pada kedua pendekayan baik terjemahan mesin berbasis statistik maupun neural untuk memperoleh hasil yang terbaik

Penelitian ini membandingkan performa dari NMT berbasis perhatian (*attention-based NT)* dan SMT berbasis frasa (*phrase-based NT*) pada domain yang spesifik yaitu realigi (Bible). Umumnya, *attention-based NT* mengacu pada fokus selektif pada sub-bagian kalimat selama terjemahan, sedangkan *phrase-based NT* terjemahan yang mana frase berfungsi sebagai unit terkecil

Metode

A. Statistical Machine Translation

Statistical Machine Translation (SMT) adalah salah satu dari pendekatan penerjemahan berbasis komputer yang didasarkan pada metode statistik yang diperkenalkan sejak pertengahan abad ke-20. Popularitas SMT muncul karena arsitektur terjemahan yang didasarkan pada teori matematika yang kuat, terjemahan berkualitas baik dalam banyak tes yang dijalankan, dan keberadaan toolkit yang dapat digunakan untuk membangun model MT dalam waktu singkat.

SMT memperoleh sebuah kalimat (berbahasa sumber), $S = s_1, s_2, s_3, \dots s_n$, dan menghasilkan sebuah kalimat target $T = t_1, t_2, t_3, \dots t_n$, sesuai dengan Bahasa target. Dalam model probabilitas, kalimat target terbaik, T* adalah ninai probabilitas tertinggi yang diperoleh dari P(T|S). Rumusnya akan diturunkan dengan menggunakan teorema Bayes, yaitu :

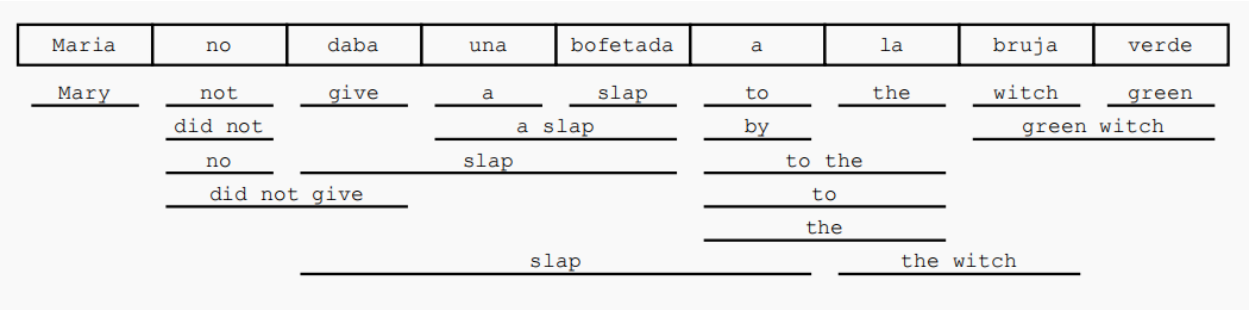
$$T^* = argmax\ (P(T|S)) \qquad (1)$$

$$T^* = argmax\ \left(\frac{P(S|T)\times P(T)}{P(S)}\right) \qquad (2)$$

$$T^* = argmax(P(S|T) \times P(T) \qquad (3)$$

P(T) adalah probabilitas dari kalimat target dan dievaluasi menggunakan sebuah model Bahasa. Model bahasa untuk suatu bahasa dibangun menggunakan korpus teks. Model bahasa menyimpan statistik untuk urutan kata. Biasanya, probabilitas untuk urutan 3 kata, yang disebut 3-*gram*, digunakan. Oleh karena itu, model bahasa berfungsi sebagai tata bahasa dalam bentuk statistik untuk SMT.

Di sisi lain, P(S|T) adalah probabilitas kalimat bahasa sumber terhadap kalimat bahasa target. Model terjemahan terdiri dari tabel terjemahan frase dan tabel penataan ulang. Tabel terjemahan frasa berisi frasa dan terjemahannya. Setiap terjemahan diberi probabilitas. Sedangkan tabel penataan ukang (*reordering*) menyimpan informasi, mengenai penataan ulang frase target. Sebuah model terjemahan harus dibangun menggunakan corpus paralel.



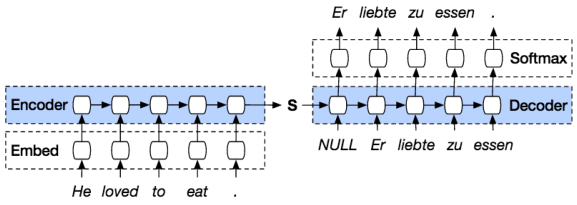
Gambar 1: Beberapa pilihan terjemahan untuk kalimat berbahasa Spanyol *Maria no daba una bofetada a ala bruja verd* ke kalimat berbahasa Inggris menggunakan SMT

P(T) adalah probabilitas kalimat bahasa target, yang dimodelkan oleh model bahasa. Model bahasa untuk bahasa sasaran dapat dibangun dengan korpus teks bahasa sasaran. Di sisi lain, P(S|T) adalah probabilitas kalimat sumber yang diberikan kalimat target, yang dimodelkan oleh model terjemahan. Model dibangun menggunakan korpus paralel. Dengan demikian, kualitas sistem terjemahan mesin sangat bergantung pada ketersediaan sejumlah besar sumber daya untuk membangun model bahasa dan model terjemahan yang kuat. Untuk bahasa dengan sumber daya rendah, jumlah sumber daya yang terbatas ini akan membuktikan sulitnya membangun sistem SMT yang baik dan masuk akal.

B. Neural Machine Translation

Sequence to sequence di kenalkan pada paper (Cho et al., 2014) sehingga mulai saat itu RNN dianggap sebagai model machine learning translation yang memiliki encoder dan decoder.

Encoder menerima satu elemen dari urutan input pada setiap langkah waktu, memprosesnya, mengumpulkan informasi untuk elemen tersebut dan menyebarkannya ke depan. Vektor perantara (*S*), Hasil dari encoder, berisi informasi tentang seluruh urutan input untuk membantu decoder membuat prediksi yang akurat. Decoder memprediksi output pada setiap langkah waktu.



Gambar 2. Contoh Arsitektur Sequence to sequence

Salah satu batasan dari pendekatan sequence to sequence (seq2seq) ini adalah bahwa RNN mencoba mengingat seluruh urutan input melalui satu unit tersembunyi (*S*) sebelum menerjemahkannya. Hal ini dapat menyebabkan hilangnya informasi, terutama untuk urutan yang

panjang. Untuk mengatasi hal ini, penambahan mekanisme *attention* (atensi) memungkinkan RNN mengakses semua elemen input pada setiap langkah waktu tertentu. Namun, memiliki akses ke semua elemen urutan input pada setiap langkah waktu (time-step) bisa sangat melelahkan. Jadi, untuk membantu RNN fokus pada elemen paling relevan dari urutan input, mekanisme perhatian memberikan bobot perhatian yang berbeda untuk setiap elemen input. Bobot perhatian ini menunjukkan seberapa penting atau relevan elemen urutan input yang diberikan pada langkah waktu tertentu.

C. Dataset

Indonesia merupakan negara yang memiliki beragam adat dan budaya dan setiap budaya memiliki keunikannya masing-masing. Salah satu contoh keunikan yang diwariskan dari suatu budaya adalah Bahasa daerah. Suku Karo atau yang lebih dikenal dengan sebutan Batak Karo adalah suku bangsa atau kelompok etnik yang mendiami Dataran Tinggi Karo, Kabupaten Deli Serdang, Kota Binjai, Kabupaten Langkat, Kabupaten Dairi, Kota Medan, dan Kabupaten Aceh Tenggara. Suku ini memiliki bahasa yang disebut Bahasa Karo atau Cakap Karo.

Dalam penelitian ini, penulis akan melakukan pengujian terhadap kinerja kedua model MT - SMT dan NMT- untuk melakuka terjemahan kalimat berbahasa Karo ke kalimat berbahasa Indonesia dengan menggunakan domain Alkitab (*Bible*) berbahasa daerah Karo (*Pustaka Sibadia*) sebagai korpus berbahasa Karo dan Aklitab Bahasa Indonesia Masa Kini (BIMK) sebagai korpus berbahasa Indonesia. Kedua korpus ini akan dibentuk untuk menjadi korpus paralel untuk dapat melatih dan menguji kinerja kedua metode MT.

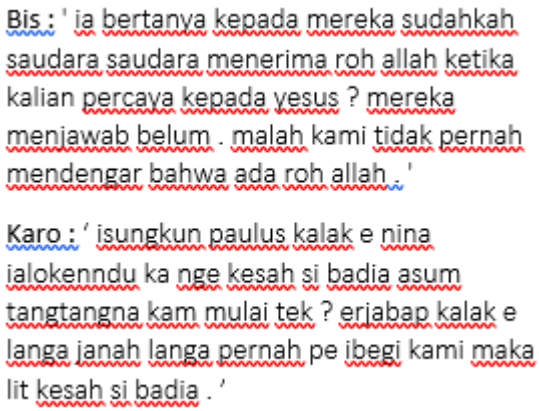
Alkitab sendiri terdiri dari 39 kitab Perjanjian Lama (PL) dan 27 kitab Perjanjian Baru (PB). Sebagai data uji, akan digunakan 38 kitab (PL) dan 26 kitab (PB). Untuk teks Alkitab berbahasa dan teks Alkitab berbahasa Indonesia Masa Kini diperoleh dari https://github.com/erwindosianipar/api-alkitab.

Untuk mengevaluasi kinerja dari kedua metode tersebut, penulis menggunakan *BLUE score*. Skor BLEU adalah metrik yang biasa digunakan untuk mengevaluasi kualitas terjemahan dengan membandingkan hipotesis/output terjemahan dengan terjemahan referensi.

Hasil dan Pembahasan

A. Pembuatan Korpus Paralel

Teks Alkitab yang akan diolah menjadi korpus diperoleh melalui API yang dapat diakses secara umum. Teks disimpan dalam format .xls untuk membatu pengolahan data. File .xls masing-masing bahasa kemudian dipecah perayat dan dipasangkan sesuai dengan susunan ayatnya, sehingga pada satu bagian ayat (baris) terdapat dua bagian yaitu ayat berbahasa Indonesia dan berbahasa Batak Karo. Dari proses ini diperoleh 650.938 pasang kalimat.



Gambar 3 : Contoh pasangan ayat pada korpus paralel Bahasa Indonesia-Karo

Text ayat yang telah berpasangan tersebut kemudian dilakukan normalisasi dengan menghapus seluruh karakter yang tidak terpakai beserta angka-angka. Kemudian dilakukan penyaringan dimana ayat yang dipakai ialah ayat yang memiliki panjang tidak lebih dari 50 kata dan tidak kurang dari 2 kata. Dari proses ini diperoleh 131.792 pasang kalimat, yang kemudian data akan dibagi 2 untuk data pelatihan (90%) dan data pengujian (10%).

B. Pelatihan dan pengujian Model

Proses pelatihan dan pengujian model dilakukan di Google Colab dengan menggunakan *framework* PyTorch. Model yang menggunakan encoder RNN regular dan penambahan mekanisme *attention* pada bagian decodernya. Algoritma Beam Search digunakan sebagai basis decoder untuk menentukan urutan kalimat. Dari beberapa percobaan yang dilakukan, peneliti peneliti melakukan percobaan untuk menentukan parameter terbaik, dan di akhir percobaan, Model dilatih dengan rincian sebagai berikut:

- Learning rate: 0.01
- Hidden_size nodes : 128
- Optimizer : Stochastic gradient descent
- Loss function : Negative log likelihood loss
- Training step : 60.000

Dari hasil pengujian, diperoleh nilai loss 3,534, dimana hasil nilai loss ini belum bisa dikatakan cukup baik. Setelah diuji dengan matrix bleu, model hanya memperoleh nilai 0.188. Untuk meningkatkan kemampuan model mesin translasi, peneliti mencoba untuk mencari model berbahasa Batak lainnya, yang telah dilatih sebelumnya, dengan tujuan melakukan *transfer learning* ke model yang digunakan. Sayangnya model tersebut belum berhasil ditemukan.

Peneliti juga telah mencoba untuk mentraining model dengan dataset yang sudah lebih umum digunakan dengan bermaksud melakukan transfer learning jika memperoleh hasil yang baik. Dataset yang digunakan yaitu korpus parallel Prancis-Inggris. (https://www.manythings.org/anki/fra-eng.zip)

dengan parameter yang sama dengan model diatas. Namun hasil skor bleu yang diperoleh juga kurang baik, yaitu 0.02. Peneliti juga telah mencoba untuk mentraining model dengan dataset yang sudah lebih umum lainya yaitu korpus paralel Indonesia-Inggris (https://www.manythings.org/anki/ind-eng.zip) dengan parameter yang sama dengan model diatas. Hasil skor bleu yang diperoleh lebih baik dari Prancis-Inggris, yaitu 0.19.

Dari skor bleu yang diperoleh, hasilnya tidak ada yang melebihi berhasil mengungguli skor bleu Indonesia-Karo, sehingga peneliti tidak melakukan transfer learning dari kedua percobaan tersebut.

Untuk pelatihan dan pengujian model berbasis statistical machine translation akan dilakukan di penelitian selanjutnya. Pengujian ini penting dilakukan karena model yang berbasis statistik memberikan hasil yang lebih baik pada kondisi data pelatihan yang sedikit.

Kesimpulan

Berdasarkan percobaan yang telah dilakukan, korpus paralel Indonesia-karo masih belum benar-benar bersih, terbukti dari 650.938 pasangan kalimat yang dapat digunakan untuk proses training hanya 131.792 pasang kalimat (20,24 %). Dari metode yang digunakan masih belum bisa memberikan hasil yang baik. Skor *bleu* yang terbaik yang berhasil diperoleh ialah 0.188. Dari hasil pelatihan model dengan korpus bahasa asing (Prancis-Inggris dan Indonesia-Inggris), dapat kita lihat bahwa pemahaman akan karakter tulisan dari setiap negara cukup berpengaruh pada kualitas korpus dan tentunya akan mempengaruhi skor *bleu*, sehingga butuh pendekatan khusus pada proses normalisasi dan filtrasi data.

Referensi

Abidin, Z. (2017). Penerapan Neural Machine Translation untuk Eksperimen Penerjemahan secara Otomatis pada Bahasa Lampung – Indonesia. *Prosiding Seminar Nasional Metode Kuantitatif 2017, 53-68*. ISBN No. 978-602-98559-3-7

Benyamin Ahmadnia, Bonnie J. Dorr, Low-Resource Multi-Domain Machine Translation for Spanish-Farsi: Neural or Statistical?, Procedia Computer Science, Volume 177, 2020, Pages 575-580, ISSN 1877-0509

Cho, K. et al. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation

Hermanto, A., Adji, T. B., & Setiawan, N. A. (2015). Recurrent Neural Network Language Model for English-Indonesian Machine Translation: Experimental Study. *Proceeding of International Conference on Science in Information Technology*. Yogyakarta, Indonesia. 27 – 28 October 2015

Junczys-Dowmunt, M., Dwojak, T., & Hoang, H. (2016). *Is Neural Machine Translation Ready for Deployment ? A Case Study on 30 Translation Directions* [diakses 1Agutus 2017]. Retrieved from: https://workshop2016.iwslt.org/downloads/WSLT_2016_paper_4.pdf

Koehn, P. (2004). Pharaoh: A Beam Search Decoder for Phrase-Based Statistical Machine Translation Models. In: Frederking, R.E., Taylor, K.B. (eds) Machine Translation: From Real Users to Research. AMTA 2004. Lecture Notes in Computer Science, vol 3265. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-30194-3_13

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. *Proc. Workshop on Neural Machine Translation, pp. 28–39, 2017*.

Luong, T., Pham, H. and Manning, D. C. (2015). Effective approaches to attention-based neural machine translation. *Proc. Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal, pp. 1412–1421, Sept. 2015*.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.J. (2002). BLEU: a method for automatic evaluation of machine translation. *Proceeding of the 40th annual meeting on association or computational linguistics*. Philadelphia, Pennsylvania. 7 - 12July 2002

Vaswani, A. et al. (2017). Attention Is All You Need. [online] arXiv.org. Available at: https://arxiv.org/abs/1706.03762. [Accessed 18 Aug. 2021]

Zhang, J., & Zong, C. (2015). Deep Neural Networks in Machine Translation: An Overview. *IEEE Intelligent Systems.Vol.30, Issue: 5, pp. 16 – 25*.