JURNAL TEKNOINFO
Volume 16, Nomor 2, Juli 2022, Page 313-322
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

MESIN PENTERJEMAH BAHASA INDONESIA-BAHASA SUNDA MENGGUNAKAN RECURRENT NEURAL NETWORKS

Yustiana Fauziyah[1)], Ridwan Ilyas[2)], Fatan Kasyidi[3)]

[1]Informatika/Fakultas Sains dan Informatika, Universitas Jenderal Achmad Yani
[2,3]Informatika/Fakultas Sains dan Informatika, Universitas Jenderal Achmad Yani
[1]Jalan Terusan Sudirman, Cimahi
[2,3]Jalan Terusan Sudirman, Cimahi
Email: [1]yustianafaa28@gmail.com, [2]rdwnilyas@gmail.com, [3]fatan.kasyidi@lecture.unjani.ac.id

Abstract

Translator is a process where one language is changed into another language. Translator in the last research was carried out using a Phrase-based Statistical Machine Translation (PSMT) approach. This research builds an Indonesian to Sundanese translator. The stages used start from pre-processing using text preprocessing and word embedding Word2Vec and the approach used is Neural Machine Translation (NMT) with Encoder-Decoder architecture in which there is a Recurrent Neural Network (RNN). Tests in the study resulted in the optimal value by the GRU of 99.17%. Models using Attention got 99.94%. The use of optimization model got optimal results by Adam 99.35% and BLEU Score results with optimal bleu 92.63% and brievity penalty 0.929. The results of the machine translator produce training predictions from Indonesian to Sundanese if the input sentences are in accordance with the corpus and the translation results are not suitable when the input sentences are different from the corpus.

Keyword: *machine translation, NMT, RNN, SeqToSeq, Attention..*

Abstrak

Penterjemah merupakan suatu proses dimana suatu bahasa diubah ke dalam bahasa lain. Penterjemah pada Penelitian lalu dilakukan dengan menggunakan pendekatan *Phrase-based Statistical Machine Translation* (PSMT). Penelitian ini membangun sebuah penerjemah Bahasa Indonesia ke Bahasa Sunda. Adapun tahapan yang digunakan dimulai dari pra proses menggunakan text preprocessing dan *word embedding Word2Vec* dan pendekatan yang digunakan yaitu *Neural Machine Translation* (NMT) dengan arsitektur *Encoder-Decoder* yang didalamnya terdapat sebuah *Recurrent Neural Network* (RNN). Pengujian pada penelitian menghasilkan nilai optimal oleh GRU sebesar 99,17%. Model dengan menggunakan *Attention* mendapat 99.94%. Penggunaan model optimasi mendapat hasil optimal oleh Adam 99.35% dan hasil *BLEU Score* dengan optimal bleu 92.63% dan *brievity penalty* 0.929. Hasil dari mesin penterjemah menghasilkan prediksi pelatihan dari Bahasa Indonesia ke Bahasa Sunda apabila input kalimat sesuai dengan korpus dan hasil terjemahan kurang sesuai ketika input kalimat berbeda dari korpus.

Kata Kunci: mesin penterjemah, NMT, RNN, SeqToSeq, Attention.

1. PENDAHULUAN

Negara Indonesia memiliki bahasa nasional yaitu Bahasa Indonesia yang digunakan dalam keseharian oleh mayoritas masyarakatnya, disamping bahasa nasional Indonesia juga memiliki bahasa daerah, salah satunya yaitu Bahasa Sunda yang ada di Jawa Barat. Menutur penelitian terdahulu, penutur Bahasa Sunda mengalami penurunan dari tahun ke tahun dan juga banyak karya ataupun sastra yang ditulis dalam Bahasa Sunda, namun tidak semua mengerti mengenai Bahasa Sunda, sehingga dibutuhkan suatu penterjemahan dalam lingkup Bahasa Sunda. Disisi lain masyarakatnya, beberapa belum terbiasa dengan penggunaan suatu bahasa daerah asing yang tidak digunakannya [1]. Maka dibutuhkan suatu penterjemahan dari bahasa nasional ke dalam bahasa daerah untuk dapat mengerti apa yang akan disampaikan. Penterjemahan tersebut dapat dibangun menggunakan suatu mesin penterjemah, yang di dalamnya memproses data kalimat korpus paralel atau data pasangan kalimat dari dua bahasa berbeda [2].

Penterjemahan termasuk pada kategori *Natural Language Processing* (NLP) atau pemrosesan bahasa alami [3], dalam bidangnya yang disebut dengan *Machine Translation* (MT) atau mesin penterjemah. MT melakukan sebuah konversi dari

JURNAL TEKNOINFO
Volume 16, Nomor 2, Juli 2022, Page 313-322
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

bahasa sumber atau bahasa yang digunakan ke bahasa target organisasi atau bahasa tujuan [4], banyak pendekatan yang dilakukan dalam MT yaitu *Statistical Machine Translation* (SMT), *Rule-based Machine Translation* (RMT), *Phrase-based Machine Translation* (PMT) dan *Neural Machine Translation* (NMT) [5]. Penelitian terdahulu mengenai penterjemahan Bahasa Sunda ke Bahasa Indonesia dilakukan dengan menggunakan pendekatan *Phrase-based Statistical Machine Translation* (PSMT). Sehingga menimbulkan celah untuk pengembangan dengan menggunakan pendekatan dan metode yang lain. Dari beberapa pendekatan, NMT merupakan suatu metodologi pengembangan terbaru pada MT dengan peningkatan yang lebih baik dari pendekatan sebelumnya [6], penelitian terdahulu menggunakan NMT menghasilkan model yang efektif untuk melakukan penterjemahan [7].

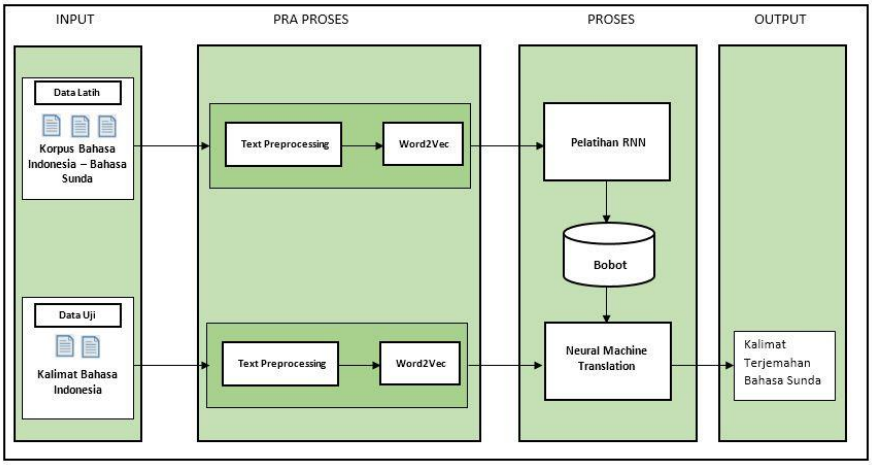
Pendekatan NMT tidak bisa begitu saja melakukan suatu penterjemahan bahasa, pada sebuah NMT berisi susunan model, lapisan proses yang terdiri dari metode yang digunakan. NMT menggunakan sebuah pemodelan yang disebut dengan *Sequence To Sequence* (SeqToSeq) [8], yang dapat menunjang terjadinya proses terjemahan bahasa. Pada model *SeqToSeq* terdapat dua buah tahapan yaitu *Encoder* dan *Decoder* [9], dimana *Encoder* merupakan sebuah lapisan proses yang akan menjadi masuknya bahasa sumber dan *Decoder* merupakan lapisan proses hasil terjemahan dari lapisan Encoder yang diubah menjadi sebuah bahasa target atau bahasa terjemah [10]. Lapisan *Encoder-Decoder* terdiri dari sebuah jaringan proses pembelajaran dengan metode *Recurrent Neural Network* (RNN) [11]. RNN merupakan sebuah metode yang sering digunakan dalam pengolahan data sekuensial seperti pengolahan teks dan lainnya [12]. RNN memiliki salah satu pengembangan variasi yaitu *Long Short-Term Memory* (LSTM) yang memiliki struktur relatif kompleks dengan penggunaan sebuah sel memori untuk menghubungkan setiap data masukan pertama dan seterusnya. Penelitian terdahulu menggunakan LSTM pada model *SeqToSeq* memberikan hasil cukup baik [13].

Namun model SeqToSeq biasa menghasilkan sebuah bottleneck dimana hasil dari encoder direpresentasikan hanya pada satu dimensi vektor sehingga model kurang mampu menghasilkan kata-kata yang jarang ditemui dalam sebuah korpus, untuk mengatasinya pada penelitian terdahulu model SeqToSeq ditambahkan sebuah mekanisme Attention [14]. Mekanisme Attention juga digunakan untuk meningkatkan NMT secara selektif dengan berfokus pada bagian-bagian dari bahasa sumber selama melakukan proses penterjemehan [15].

Penelitian ini mengusulkan pendekatan NMT pada mesin penterjemahkan kalimat Bahasa Indonesia ke Bahasa Sunda menggunakan arsitektur *Encoder-Decoder* dengan penggunaan metode RNN LSTM untuk memproses kata masukan pada mesin penterjemah dan mekanisme tambahan yaitu Attention Mechanism, selanjutnya dilakukan pengujian untuk pengukuran akurasi model dengan *Bilingual Evaluation Understudy* (BLEU).

2. METODE PENELITIAN

Penelitian ini mengusulkan pendekatan NMT pada mesin penterjemahkan kalimat Bahasa Indonesia ke Bahasa Sunda menggunakan arsitektur *Encoder-Decoder* dengan penggunaan metode RNN LSTM untuk memproses kata masukan pada mesin penterjemah dan mekanisme tambahan yaitu *Attention Mechanism*, selanjutnya dilakukan pengujian untuk pengukuran akurasi model dengan *Bilingual Evaluation Understudy* (BLEU).



Gambar 1. Rancangan Sistem Mesin Penterjemah Bahasa Indonesia-Bahasa Sunda

JURNAL TEKNOINFO
Volume 16, Nomor 2, Juli 2022, Page 313-322
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

Mesin penterjemah memiliki empat proses yaitu *input*, praproses, proses dan *output*. Proses terbagi ke dalam data latih dan data uji, *input* data latih merupakan korpus Bahasa Indonesia dan Bahasa Sunda sedangkan data uji merupakan kalimat Bahasa Indonesia, keduanya akan melewati tahap kedua yaitu pra proses data dan setelah itu masuk pada pemblejaran RNN dengan model SeqToSeq untuk menghasilkan *output* yaitu terjemahan dalam Bahasa Sunda.

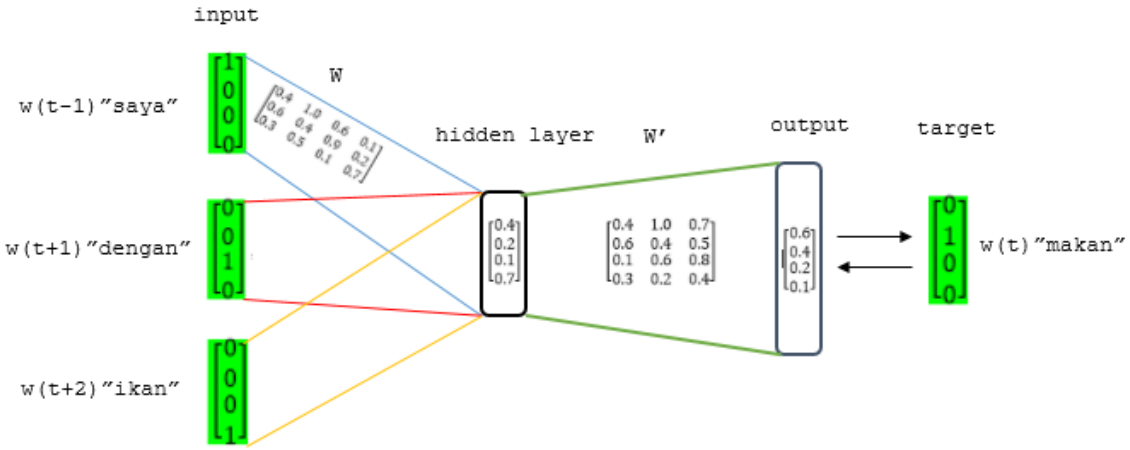
2.1 *Dataset*

Data yang digunakan pada penelitian ini berasal dari penelitian terdahulu yang merupakan korpus *low-resource* Bahasa Sunda loma. Adapun beberapa modifikasi pada data yaitu mengubah bentuk data dan melakukan penambahan data korpus. Pengubahan data dari penelitian terdahulu menghasilkan 1187 korpus dan pengembangan data manual menghasilkan 2309 korpus, sehingga jumlah data yang digunakan yaitu 3496 korpus data Bahasa Indonesia dan Bahasa Sunda dengan format .csv.

2.2 Praproses Data

Praproses data dilakukan untuk mengolah data agar siap digunakan di tahap proses pembelajaran pada model NMT mesin penterjemah. Praproses data dilakukan dengan *text preprocessing* yang digunakan untuk seleksi data agar lebih terstruktur seperti normalisasi, *text cleaning, tokenizing, case folding* dan tahapan *text preprocessing* lainnya. Selain *text preprocessing* praproses lainnya adalah pengubahan bentuk data menjadi sebuah vektor kata. Praproses ini terdapat pada suatu *embeding layer* yang berisikan nilai vektor untuk dilakukan proses pembelajaran.

Penelitian ini menggunakan Wor2Vec sebagai mekanisme pengubahan data korpus menjadi sebuah vektor. Setiap kata akan direpresentasikan pada suatu vektor *one-hot-encoding* yang berisi nilai biner 0 dan 1 dengan nilai 1 merepresentasikan letak kata dalam vektor. Arsitektur Word2Vec yang digunakan adalah *Continous Bag-Of-Words* (CBOW) yang merupakan sebuah model untuk memprediksi kata target dari beberapa konteks kata yang diberikan. Arsitektur CBOW ditunjukan oleh Gambar 2.



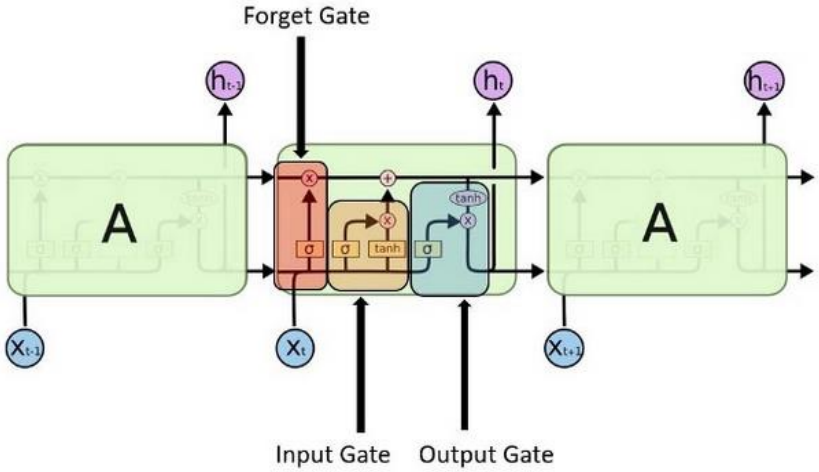
Gambar 2. Arsitektur CBOW

Dimana *w* merupakan suatu konteks kata yang berisi nilai dari *one-hot-encoding* dan *t* merepresentasikan konteks sebelum dan sesudah dari kata target *w(t)* yang dicari. Setiap nilai konteks akan dikalikan dengan matriks bobot pada bagian *input layer* lalu menghasilkan nilai bobot pada *hidden layer* atau *projection,* selanjutnya mengalikan kembali bobot *hidden layer* dengan bobot *output* sehingga menghasilkan vektor akhir dan menjadikan bobot tersebut nilai vektor pada *w(t).*

2.3 *Recurrent Neural Networks*

RNN merupakan merupakan suatu metode dalam deep learning yang digunakan untuk memproses data sekuensial dengan pemanggilan berulang. Pada arsitektur RNN memiliki beberapa jaringan neuron yang berbalik dari dirinya sendiri atau ke *neuron* di *layer input* sehingga jaringan dapat menyimpan nilai yang memberi dampak pada cara *input* untuk menghasilkan nilai sebelumnya ke dalam jaringan. Lapisan pertama memiliki bobot dari lapisan *input*, lalu lapisan selanjutnya akan menerima bobot dari lapisan sebelumnya. Penelitian ini menggunakan RNN dengan variasi LSTM yang memiliki *memory cells* untuk dapat menyimpan informasi dengan jangka waktu yang panjang. LSTM memiliki beberapa *gate* diantaranya : *cell state, forget gate, input gate* dan *output gate*.

JURNAL TEKNOINFO
Volume 16, Nomor 2, Juli 2022, Page 313-322
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index



Gambar 3. Arsitektur LSTM

Pada Gambar 3 memperlihatkan proses pada setiap gerbang. Kata *input* direpresentasikan sebagai *x* untuk setiap *time step* ke-*t*. Kata akan melalui gerbang pertama yaitu *forget gate* dengan Persamaan (1).

$$f_t = \sigma\,(U_f h_{t-1} + W_f x_t + b_f) \quad (1)$$

Dimana σ merupakan fungsi aktivasi sigmoid, $U_f h_{t-1}$ merupakan bobot pada *hidden layer* dan $W_f x_t$ merupakan bobot matriks kata dan b_f adalah bias. Selanjutnya masuk pada proses *input layer* dengan Persamaan (2) dan Persamaan (3).

$$i_t = \sigma\,(U_i h_{t-1} + W_i x_t + b_i) \quad (2)$$

$$\hat{h}_t = tanh(U h_{t-1} + W x_t + b) \quad (3)$$

Sama seperti notasi sebelumnya untuk fungsi pada *input gate* terdapat matriks *U* dan *W*. Terdapat dua fungsi aktivasi yang ada pada *input gate*. Pada Persamaan (3), $\hat{h}_t$ merupakan sebuah kandidat *hidden layer*. Setelah proses *input gate* selesai terdapat pembaharuan *cell state* untuk melakukan *update* terhadap informasi yang dimiliki pada *cell* dengan mengalikan *cell state* sebelumnya sengan nilai vektor *forget gate* lalu mengambil nilai dari *input gate* dan kandidat *hidden state*. Proses dilakukan dengan Persamaan (4).

$$\tilde{c}_t = f_t \odot c_{t-1} + i_t \odot \hat{h}_t) \quad (4)$$

Dimana $\tilde{c}_t$ merupakan *cell state* dan c_{t-1} merupakan *cell state* sebelumnya. Proses terakhir yaitu *output gate* dilakukan dengan menggunakan Persamaan (5) dan Persamaan (6).

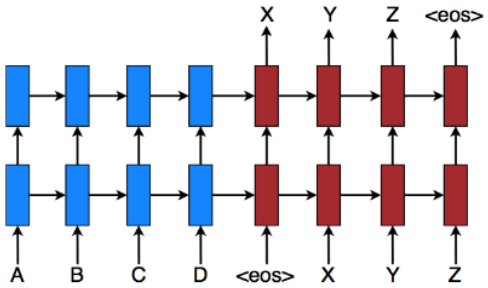
$$o_t = \sigma\,(U_o h_{t-1} + W_o x_t + b_o \quad (5)$$

$$h_t = o_t \odot \tanh(c_t) \quad (6)$$

2.4 *Sequence to Sequence*

Dalam mesin penterjemah terdapat suatu pemetaan dari urutan kata *input* ke urutan kata *output*. Namun dalam bentuk nyata penterjemahan urutan kata *input* terkadang berbeda dengan kata ruang *output*, selain itu panjang ukuran *input* terkadang tidak sama dengan panjang ukuran *output* dan letak urutan yang dihasilkan. Untuk memetakan urutan *input* ke ruang *output* munculah sebuah model yang dinamakan dengan *SeqToSeq*. Terdapat arsitektur pada *SeqToSeq* untuk menangani proses penterjemahan yaitu *Encoder* dan *Decoder*.

JURNAL TEKNOINFO
Volume 16, Nomor 2, Juli 2022, Page 313-322
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index



Gambar 4. Model *SeqToSeq*

Pada Gambar 4 menunjukan sebuah proses *Encoder* dan *Decoder*, keduanya mewakili pemprosesan sebuah bahasa sumber yang menjadi *input* untuk menghasilkan bahasa target yang menjadi *output*.

2.4.1 Encoder

Encoder membaca kalimat *input* dari bahasa sumber dan merangkum informasi tersebut dengan mendekodekan ke dalam suatu vektor di *hidden state* LSTM hasilnya yaitu sebuah *context vector*. Notasi *Encoder* pada *SeqToSeq* digambarkan menjadi (x_i , ... ,$x_{T)}$.

2.4.2 Decoder

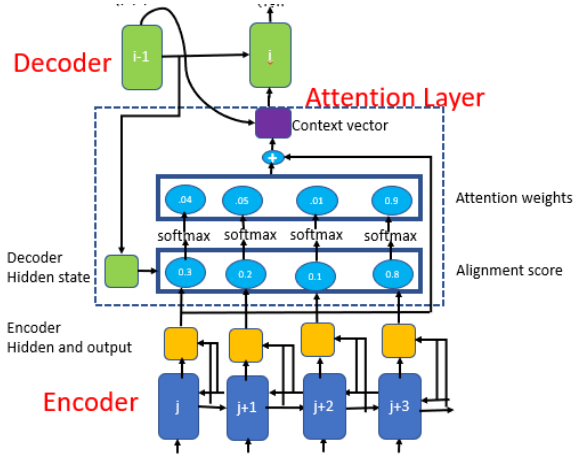
Decoder berperan untuk mendekode informasi yang diproses dari *Encoder*. *Context vector* yang diperoleh pada proses sebelumnya digabungkan dengan keluaran Decoder sebelumnya dan masuk pada *cells* RNN *Decoder* untuk menghasilkan *hidden state* baru. Proses berulang sampai dengan Decoder menemukan token "<EOS>". Notasi Decoder pada SeqToSeq digambrakan menjadi (y_i , ... ,$y_{T'}$). Maka mekanisme *SeqToSeq* dimodelkan seperti pada Persamaan (7).

$$p(y_1,\ldots,y_{T'}|x_1,\ldots,x_T) = \prod_{t=1}^{T'} p(y_1|v,y_1,\ldots,y_{t-1}) \quad (7)$$

Dimana *v* merupakan *conditional probability* yang memperoleh dimensi *fix-length* vektor dari *hidden state* terakhir yang diberikan oleh LSTM.

2.5 Attention Mechanism

Mekanisme Attention merupakan upaya dalam mempelajari representasi vektor tunggal pada kalimat atau mengarahkan fokus pada faktor penting sambil memprediksi *output* dalam model SeqToSeq. Terdapat tiga *layer* utama pada Attention yaitu *alignment layer*, *attention weights* dan *context vector*. Dimana *hidden state* Encoder bersama dengan *hidden state* Decoder digunakan untuk menghasilkan *context vector*. mekanisme ini memperhatikan spesifik vektor *input* dari urutan masukan berdasarkan *attention weights*.



Gambar 5. Mekasnisme *Attention*

JURNAL TEKNOINFO
Volume 16, Nomor 2, Juli 2022, Page 313-322
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

Penelitian ini menggunakan mekanisme Attention yang dikemukakan oleh Bahdanau dengan arsitektur yang ditunjukan oleh Gambar 5. *Alignment score* memetakan seberapa baik *input* pada posisi *j* dan *output* pada posisi *i*. Skor berdasarkan dari *hidden state* Decoder sebelumnya, dimana $s_{(i-1)}$ tepat sebelum memprediksi kata target dan *hidden state* h_j dari kalimat *input. Alignment score* dihitung menggunakan Persamaan (8).

$$e_{ij} = a(s_{i-1}, h_j) \tag{8}$$

Selanjutnya hasil *alignment score* masuk pada fungsi *softmax* untuk menghasilkan *attention weights*, fungsi aktivasi *softmax* akan mendapatkan probabilitas yang membantu dalam mewakili bobot pengaruh pada setiap urutan input. Semakin tinggi nilai *attention weights* maka semakin tinggi pengaruh yang diberikan untuk memprediksi. Perhitungan dilakukan menggunakan Persamaan (9)

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})} \tag{9}$$

Terakhir yaitu *context vector* yang digunakan untuk menghitung keluaran akhir dari Decoder, *context vector* merupakan jumlah *attention weights* dan *hidden state* Decoder yang memetakan ke kalimat *output* dihitung dengan menggunakan Persamaan (10).

$$c_i = \sum_{j=1}^{T_x} a_{ij} h_j \tag{10}$$

Terlihat bahwa a_{ij} mengartikan seberapa banyak h_j berpengaruh pada *context*. Untuk melakukan suatu prediksi pada kata target, Decoder menggunakan Persamaan (11).

$$s_i = f(s_{i-1}, c_i, y_{i-1}) \tag{11}$$

Dimana s_{i-1} merupakan Decoder *hidden state* sebelumnya, y_{i-1} merupakan *output* Decoder pada *time step* sebelumnya dan c_i merupakan *context vector.*

2.5 BLEU Score

Bilingual Evaluation Understudy (BLEU) merupakan sebuah pengujian yang banyak digunakan pada sebuah mesin penterjemah. Matriks BLEU dirancang untuk mengukur seberapa dekat keluaran yang dihasilkan dengan melakukan pencocokan panjang frasa variabel keluaran dari mesin penterjemah dengan referensi terjemahan. Matriks dasar memerlukan sebuah kalkulasi *brevity penalty* dengan perhitungan pada Persamaan (12) dan Persamaan (13).

$$P_B = \begin{cases} 1, c > r \\ e^{\left(1-\frac{r}{c}\right)}, c \leq r \end{cases} \tag{12}$$

$$BLEU = P_B \exp\left(\sum_{n=0}^{N} w_n \log P_n\right) \tag{13}$$

Dimana *r* merupakan panjang korpus referensi, *c* merupakan kandidat panjang terjemahan, w_n adalah bobot positif yang dijumlahkan menjadi satu dan P_n yaitu kalkulasi n-gram dengan maksimum N.

3. HASIL DAN PEMBAHASAN

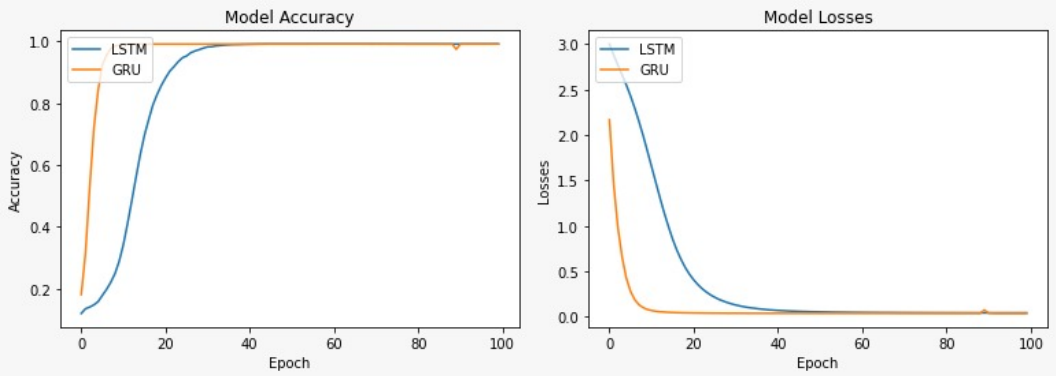
Dalam penelitian ini terdapat beberapa pengujian yang dilakukan yaitu pengujian arsitektur RNN, pengujian model NMT dengan Attention dan tanpa Attention, pengujian model optimasi dan pengujian BLEU *score*. Dataset yang digunakan berasal dari pengembangan data dengan penggunaan Bahasa Sunda jenis loma dengan jumlah 3496 data. Untuk melihat optimal model pembelajaran yang digunakan, pengujian dilakukan dengan menggunakan dua arsitektur model variasi dari Recurrent Neural Network untuk dilakukan perbandingan yaitu LSTM dan GRU dengan nilai pembelajaran 0.001 dengan 100 *epoch* pada masing-masing arsitektur. Hasil perbandingan ditunjukan oleh Tabel 1.

JURNAL TEKNOINFO
Volume 16, Nomor 2, Juli 2022, Page 313-322
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

Tabel 1. Perbandingan Learning LSTM dan GRU

Model	Accuracy (%)	Losses
LSTM	99.18	0.0426
GRU	99.17	0.0361

Setelah mencoba melakukan pengujian dengan penggunaan varian lain dari RNN, didapat hasil pengujian seperti pada Tabel 1. LSTM dan GRU keduanya memiliki pembelajaran yang baik meski GRU sedikit lebih optimal dengan nilai *accuracy* 99.17% dengan nilai *losses* yang lebih kecil dari LSTM. Namun hasil yang didapat LSTM pun menunjukan akurasi yang baik saat pembelajaran yaitu 99.18%.



Gambar 6. Perbandingan LSTM dan GRU

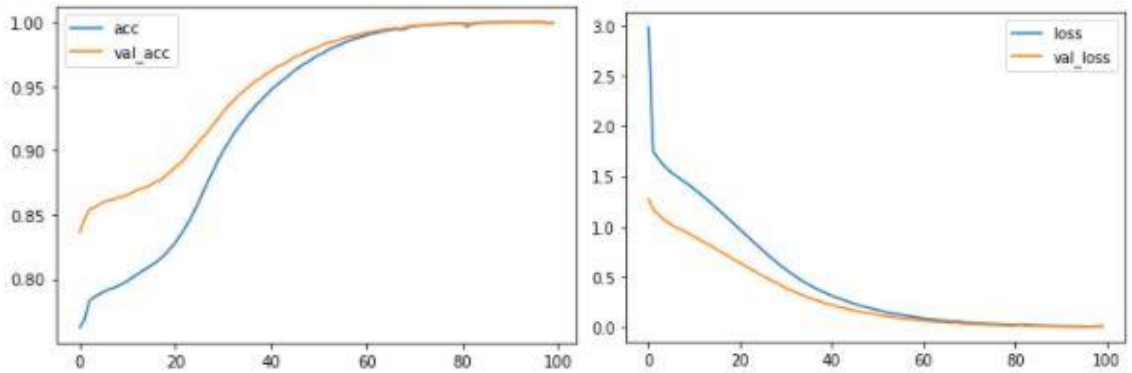
Hasil perbandingan digambarkan pada grafik yang ditunjukan oleh Gambar 6 mendapati garis model GRU sedikit lebih dulu menaik dalam proses pembelajaran dan model LSTM menaik setelahnya, meski begitu keduanya mencapai titik optimal untuk mengenali pola pada data yang diberikan. Begitupun pada nilai *losses,* GRU mengoptimalkan nilai *losses* lebih dulu dan LSTM setelahnya.

Pengujian selanjutnya menggunakan mekanisme Attention dan tanpa mekanisme Attention dengan hasil yang didapat ditunjukan oleh Tabel 2.

Tabel 2. Perbandingan Penggunaan Mekanisme Attention

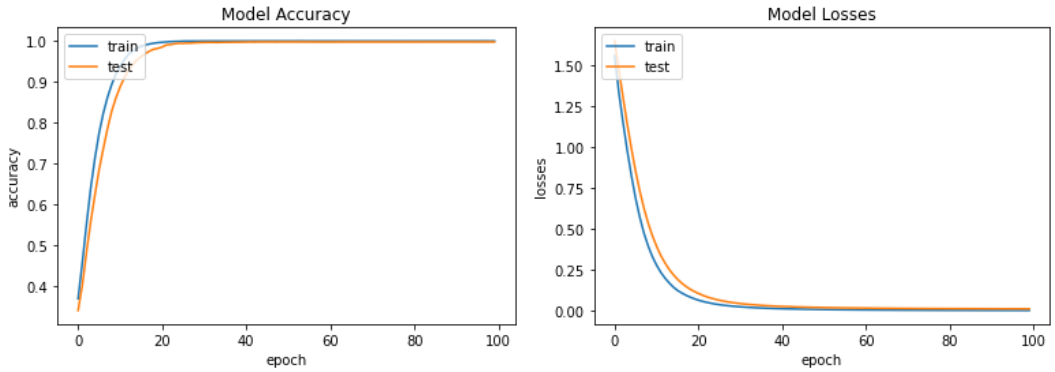
Model	Accuracy (%)	Losses
Attention	99.94	0.0075
Non-Attention	99.73	0.0116

Hasil pengujian dengan melakukan perbandingan dengan non-Attention didapatkan nilai optimal didapatkan oleh model dengan menggunakan mekanisme Attention yaitu akurasi sebesar 99.94% sedikit optimal dibandingkan dengan menggunakan model SeqToSeq biasa.



Gambar 7. Model dengan Attention

JURNAL TEKNOINFO
Volume 16, Nomor 2, Juli 2022, Page 313-322
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index



Gambar 8. Model Tanpa Attention

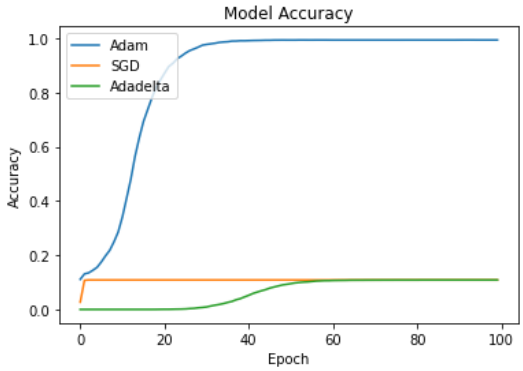
Pada Gambar 7 model dengan mekanisme Attention mengalami kerenggangan saat memulai pembelajaran karena adanya penggunaan tambahan mekanisme Attention pada sistem, sehingga terjadi proses tambahan sebelum akhirnya garis grafik selaras saat pembelajaran. Pada Gambar 8 model dengan Non-Attention garis grafik cenderung cepat menemukan titik optimal bersamaan karena sistem melakukan pembelajaran dengan SeqToSeq biasa tanpa penambahan mekanisme.

Pengujian selanjutnya yaitu dengan membandingkan penggunaan model optimasi pada saat pembelajaran, model optimasi yang digunakan yaitu Adam, SGD dan Adadelta.

Tabel 3. Hasil Penggunaan Model Optimasi

Model	Accuracy (%)	Losses
Adam	99.35	0.0396
SGD	10.87	3.3593
Adadelta	10.57	4.0068

Hasil perbandingan pada Tabel 3 mendapatkan nilai optimal oleh model optimasi Adam dengan akurasi 99.35%, sedangkan SGD dan Adadelta mendapati hasil yang kurang baik ketika dilakukan pengujian pada sistem.



Gambar 9. Hasil Model Optimasi

Hasil pengujian yang disajikan oleh Gambar 9 mendapati garis grafik optimizer Adam lebih dulu menghasilkan nilai akurasi optimal dibandingkan dengan dua model optimasi lainnya yang digunakan. SGD menghasilkan nilai akurasi yang kurang optimal dimana SGD mengambil sampel data secara acak dan tidak melakukan perubahan pada nilai *learning rate* dan *update* bobot. Begitupun Adadelta yang menghasilkan akurasi kurang baik dalam pengujian meski Adadelta merupakan suatu teknik adaptif untuk memperbaiki bobot, Adadelta berfokus pada penurunan tingkat pembelajaran yang terus menerus selama pelatihan dan kebutuhan akan tingkat pembelajaran global yang dipilih secara manual.

Pengujian selanjutnya digunakan BLEU Score untuk melihat seberapa sesuai model meghasilkan terjemahan. Dalam evaluasi BLEU Score mesin melakukan evaluasi secara otomatis pada model dengan hasil nilai BLEU optimal 92.63 dengan *brevity penalty* 0.929 yang menunjukan bahwa korpus referensi sesuai dengan kandidat terjemahan.

Dari beberapa pengujian menghasilkan hasil optimal pada masing-masing model yang dibandingkan untuk mendapatkan hasil penterjemahan yang lebih optimal. Namun hasil terjemah dari mesin penterjemah menghasilkan

JURNAL TEKNOINFO
Volume 16, Nomor 2, Juli 2022, Page 313-322
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

terjemahan yang sesuai ketika kalimat yang dimasukan merupakan kalimat sama yang ada pada korpus. Ketika mencoba untuk memasukan urutan kalimat berbeda dengan korpus maka hasil ditemukan kurang sesuai atau tidak sesuai. Ini dikarenakan penggunaan korpus yang masih termasuk pada korpus *low-resource* sementara untuk menghasilkan pembelajaran yang optimal dibutuhkan korpus yang dengan jumlah yang lebih besar agar model dapat mengenali bentuk-bentuk dari kalimat masukan, sehingga mesin masih lebih banyak menghasilkan kata ataupun kalimat yang tidak terdapat di dalam korpus. Sebagai contoh terjemahan sesuai dengan menggunakan kalimat yang sama dengan korpus ditunjukan oleh Tabel 4, sementara hasil dari kalimat kurang sesuai atau tidak sesuai ditunjukan oleh Tabel 5.

Tabel 4. Hasil Penterjemah yang Sesuai

Input	Hasil Prediksi
Input kalimat Bahasa Indonesia	Datuk Makotta dan istrinya Tuan Sitti sebagai cikal bakal keluarga Minangkabau di Sulawesi
Actual Sunda Translation	datuk makotta jeung istrina tuan sitti salaku cikal bakal kulawarga minangkabau di sulawesi
Predicted Sunda Translation	datuk makotta jeung istrina tuan sitti salaku cikal bakal kulawarga minangkabau di sulawesi
Input kalimat Bahasa Indonesia	Diawali dengan Shalawat Nabi sejarah riwayat Nabi sampai ke ceramah sejumlah ulama serta habib waktu yang tunggu yaitu waktu warga dipersilahkan merasakan hidangan yang sedia sepuasnya
Actual Sunda Translation	dimimitian kalayan shalawat nabi sajarah riwayat nabi nepi ka ceramah sajumlah ajengan sarta habib nepi ka waktu anu ditunggu nyaeta waktu warga dihaturan ngarasakeun hidangan anu sadia sawaregna
Predicted Sunda Translation	dimimitian kalayan shalawat nabi sajarah riwayat nabi nepi ka ceramah sajumlah ajengan sarta habib nepi ka waktu anu ditunggu nyaeta waktu warga dihaturan ngarasakeun hidangan anu sadia sawaregna

Pada Tabel 4 terlihat bahwa actual sunda translation merupakan kalimat referensi dari korpus menghasilkan prediksi kalimat yang sesuai dengan kalimat referensi Bahasa Sunda yang ada pada korpus, begitupun contoh lainnya ketika memasukan kalimat yang sama dari korpus akan menghasilkan prediksi yang sesuai.

Tabel 5. Hasil Penterjemah yang Tidak Sesuai

Input	Hasil Prediksi
Input kalimat Bahasa Indonesia	Saya ingin makan dan minum bersama teman-teman
Actual Sunda Translation	dina urang anu aya anu aya anu aya anu aya anu
Predicted Sunda Translation	abdi hoyong miceun jeung

Pada Tabel 5 diberikan sebuah kalimat yang berbeda dengan urutan kalimat pada korpus menghasilkan ketidaksesuaian dalam prediksi yang dihasilkan.

4. KESIMPULAN

Penelitian ini mengusulkan sistem mesin penterjemah Bahasa Indonesia ke Bahasa Sunda dengan menggunakan pendekatan Neural Machine Translation dengan aristektur Encoder-Decoder dimana terdapat RNN di dalamnya. Data yang digunakan pada mesin penerjemah sebanyak 3496 paralel korpus Bahasa Indonesia dan Bahasa Sunda. Pada pengujian dari mesin penerjemah dilakukan beberapa pengujian diantaranya : pengujian arsitektur RNN, pengujian model NMT dengan Attention dan tanpa Attention, pengujian model optimasi dan pengujian BLEU Score. Hasil yang didapatkan pada penelitian ini pada pengujian pertama arsitektur model variasi dari RNN didapatkan nilai optimal oleh GRU akurasi sebesar 99.17%. Pengujian kedua menghasilkan nilai optimal oleh model dengan menggunakan Attention dengan nilai akurasi 99.94%. Pengujian ketiga dalam perbandingan model optimasi dimana Adam mendapat hasil optimal dengan nilai akurasi 99.35%. Pengujian terakhir yaitu BLEU Score dimana menghasilkan nilai optimal bleu 92.63% dengan *brievity penalty* 0.929.

Penelitian ini menggunakan kategori Bahasa Sunda loma atau bahasa sunda yang digunakan untuk percakapan sebaya, mesin penterjemah menghasilkan prediksi pelatihan yang sesuai dimana data korpus menghasilkan keluaran yang sesuai

JURNAL TEKNOINFO
Volume 16, Nomor 2, Juli 2022, Page 313-322
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

saat memprediksi dari Bahasa Indonesia ke Bahasa Sunda apabila data masukan adalah kalimat yang terdapat pada korpus, namun hasil kurang sesuai ketika memasukan kalimat berbeda dari korpus yang ada dikarenakan pengaruh terbesar dalam mesin penterjemah membutuhkan korpus bahasa yang cukup banyak.

Adapun saran untuk penelitian selanjutnya yaitu melakukan mengembangan data yang lebih banyak untuk menghasilkan variasi predict yang lebih baik pada model NMT sehingga akan lebih mudah untuk mengenali kata-kata yang lebih banyak saat pelatihan dalam model bahasa dan dapat menerapkan penggunaan Transformer Neural Network untuk language model pada mesin penterjemah.

UCAPAN TERIMAKASIH

Terima kasih kepada Bu Arie yang telah memberi data bagi peneliti mencakup data Bahasa Indonesia dan Bahasa Sunda untuk penelitian ini.

DAFTAR PUSTAKA

[1] A. A. Suryani, D. H. Widyantoro, A. Purwarianti, and Y. Sudaryat, "Experiment on a phrase-based statistical machine translation using PoS Tag information for Sundanese into Indonesian," *2015 International Conference on Information Technology Systems and Innovation, ICITSI 2015 - Proceedings*, 2016.

[2] A. A. Suryani, I. Arieshanti, B. W. Yohanes, M. Subair, S. D. Budiwati, and B. S. Rintyarna, "Enriching English into Sundanese and Javanese translation list using pivot language," *2016 International Conference on Information, Communication Technology and System (ICTS)*, pp. 167–171, 2017.

[3] A. S. and R. K. Gaurav Tiwari, Arushi Sharma, "English - Hindi Neural Machine Translation - LSTM SeqToSeq and ConvS2S," *International Conference on Communication and Signal Processing*, pp. 871–875, 2020.

[4] H. Jiang, Y. He, M. Liao, Y. Jing, and C. Zhang, "English-Vietnamese machine translation model based on sequence to sequence algorithm," *Proceedings of 2020 IEEE 5th Information Technology and Mechatronics Engineering Conference, ITOEC 2020*, no. Itoec, pp. 1086–1091, 2020.

[5] P. Shah and V. Bakrola, "Neural machine translation system of indic languages - An attention based approach," *arXiv*, pp. 1–5, 2020.

[6] G. Klein, J. Senellart, and A. M. Rush, "OpenNMT : Neural Machine Translation Toolkit," *Proceedings of AMTA 2018, vol 1: MT Research Track*, vol. 1, pp. 177–184, 2018.

[7] Y. Wu, "A chinese-english machine translation model based on deep neural network," *Proceedings - 2020 International Conference on Intelligent Transportation, Big Data and Smart City, ICITBS 2020*, pp. 828–831, 2020.

[8] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in Neural Information Processing Systems*, vol. 4, no. January, pp. 3104–3112, 2014.

[9] B. Van Merri, "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," *arXiv*, 2013.

[10] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation By Jointly Learning to Align and Translate," *Published as a conference paper at ICLR*, pp. 1–15, 2015.

[11] D. Britz, A. Goldie, M. Luong, and Q. Le, "Massive Exploration of Neural Machine Translation Architectures," *arXiv*, 2017.

[12] Y. Liu, D. Zhang, L. Du, Z. Gu, J. Qiu, and Q. Tan, "A simple but effective way to improve the performance of RNN-Based encoder in neural machine translation task," *Proceedings - 2019 IEEE 4th International Conference on Data Science in Cyberspace, DSC 2019*, pp. 416–421, 2019.

[13] J. C. Heck and F. M. Salem, "Simplified minimal gated unit variations for recurrent neural networks," *Midwest Symposium on Circuits and Systems*, vol. 2017-Augus, pp. 1593–1596, 2017.

[14] A. M. Rush, "Structured Attention Networks," *Published as a conference paper at ICLR*, pp. 1–21, 2017.

[15] M. T. Luong, H. Pham, and C. D. Manning, "Effective Approaches to Attention-based Neural Machine Translation," *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421, 2015.