

PENERJEMAHAN KALIMAT BAHASA LAMPUNG-INDONESIA DENGAN PENDEKATAN *NEURAL MACHINE TRANSLATION* BERBASIS *ATTENTION*

TRANSLATION OF SENTENCE LAMPUNG-INDONESIAN LANGUAGES WITH NEURAL MACHINE TRANSLATION ATTENTION BASED APPROACH

Zaenal Abidin¹, Adi Sucipto², Arief Budiman³

Fakultas Teknik dan Ilmu Komputer^{1,2,3}

Universitas Teknokrat Indonesia

Email: zabin@teknokrat.ac.id

Dikirim 09 Juli 2018 Direvisi 26 Juli 2018 Disetujui 27 Juli 2018

Abstrak: Pada penelitian ini, penerjemahan secara otomatis pada bahasa Lampung ke bahasa Indonesia dilakukan dengan pendekatan *neural machine translation* (NMT) berbasis *attention*. NMT, sebuah pendekatan baru dalam teknologi mesin penerjemah, bekerja dengan memadukan *encoder* dan *decoder*. *Encoder* adalah komponen berupa *recurrent neural network* yang mengkodekan bahasa sumber menjadi vektor-vektor yang panjangnya tetap, dan *decoder* adalah sebuah komponen berupa *recurrent neural networks* yang membangkitkan hasil terjemahan secara komprehensif. Penelitian NMT diawali dengan pembuatan 3000 kalimat paralel bahasa Lampung (dialek *api*) – Indonesia kemudian dilanjutkan dengan penentuan parameter model NMT untuk proses *training data*, tahap selanjutnya adalah membangun model NMT dan menguji model NMT. Pengujian dengan pendekatan NMT menggunakan 25 kalimat tunggal tanpa *out-of-vocabulary* (OOV), 25 kalimat tunggal dengan OOV, 25 kalimat majemuk tanpa OOV dan 25 kalimat majemuk dengan OOV. Hasil pengujian penerjemahan kalimat bahasa Lampung ke bahasa Indonesia menunjukkan nilai rata-rata *bilingual evaluation understudy* (BLEU) yang diperoleh dengan pendekatan NMT adalah 51.96 %.

Kata kunci: *Decoder, Encoder, Neural machine translation, Out-of-Vocabulary, Recurrent neural networks.*

Abstract: In this research, automatically Lampung language translation into the Indonesian language was using *neural machine translation* (NMT) attention based approach. NMT, a new approach method in machine translation technology, that has worked by combining the *encoder* and *decoder*. The *encoder* in NMT is a *recurrent neural network* component that encrypts the source language to several length-stable vectors and the *decoder* is a *recurrent neural networks* component that generates translation result comprehensive. NMT Research has begun with creating a pair of 3000 parallel sentences of Lampung language (*api* dialect) and Indonesian language. Then it continues to decide the NMT parameter model for the data training process. The next step is building NMT model and evaluate it. The testing of this approach has used 25 single sentences without *out-of-vocabulary* (OOV), 25 single sentences with OOV, 25 plural sentences without OOV, and 25 plural sentences with OOV. The testing translation result using NMT attention shows the *bilingual evaluation understudy* (BLEU) an average value is 51, 96 %.

Keywords : *Decoder, Encoder, Neural machine translation, Out-of-Vocabulary, Recurrent neural networks.*

PENDAHULUAN

Kondisi saat ini di Indonesia, dari 617 bahasa daerah yang teridentifikasi terdapat 15 bahasa daerah yang dinyatakan punah dan 139 lainnya dalam status terancam punah. Informasi ancaman kepunahan bahasa daerah dinyatakan oleh Dadang Sunendar selaku Kepala Badan Pembinaan dan Pengembangan Bahasa Kementerian

Pendidikan dan Kebudayaan melalui hasil risetnya, usai menghadiri Kongres Bahasa Daerah Nusantara di gedung Merdeka Bandung pada tahun 2016 (Anonim, 2016).

Dalam konteks menghindari ancaman kepunahan tersebut, pelestarian bahasa daerah menjadi teramat penting karena bahasa merupakan salah satu identitas

penting suatu daerah. Bahasa daerah merupakan bentuk ekspresi kultural utama suatu etnis atau daerah, tidak terkecuali untuk bahasa Lampung. Bahasa Lampung merupakan identitas penting dari kebudayaan Lampung. Jika tidak ada upaya menjaga bahasa Lampung, maka akan mengakibatkan sirnanya kebudayaan masyarakat Lampung (Anonim, 2017).

Upaya menjaga dan melestarikan bahasa daerah tidak hanya harus dilakukan oleh pemerintah pusat, melainkan membutuhkan juga peran aktif pemerintah daerah, masyarakat, serta institusi pendidikan atau lembaga pendidikan. Setiap provinsi memiliki bahasa daerah, baik yang berbentuk tulisan maupun ucapan, termasuk salah satunya provinsi Lampung yang memiliki bahasa daerah dengan dua dialek utama, yaitu dialek *api* dan dialek *nyo*. Upaya pemerintah provinsi Lampung untuk melestarikan bahasa Lampung, salah satunya diberikan melalui pendidikan bahasa Lampung sebagai muatan lokal wajib dari mulai sekolah tingkat dasar sampai sekolah tingkat menengah atas. Hal ini sesuai peraturan gubernur nomor 39 tahun 2014 tentang mata pelajaran bahasa dan aksara Lampung sebagai muatan lokal wajib pada jenjang satuan pendidikan dasar sampai menengah atas dan didukung oleh ketersediaan buku ajar mulai dari SD, SMP dan SMA, berikut kamus bahasa Lampung.

Pemanfaatan alat berbasis komputer dapat digunakan sebagai upaya untuk melestarikan bahasa Lampung secara digital sesuai perkembangan teknologi saat ini, diantaranya dengan membangun kamus digital dan mesin penerjemah. Untuk mesin penerjemah, salah satu kakas mesin penerjemah yang ada saat ini adalah mesin penerjemah Google yang dapat diakses melalui laman <https://translate.google.com>. Mesin ini dapat diakses dengan mudah, akan tetapi pada mesin penerjemah Google hanya terdapat dua bahasa daerah yang tersedia

yaitu bahasa Jawa dan bahasa Sunda. Untuk bahasa Lampung tidak tersedia di mesin penerjemah Google.

Pendekatan mesin penerjemah dapat dilakukan dengan berbasis *data-driven* atau menggunakan korpus paralel (Bhattacharyya, 2015). *Neural machine translation* (NMT) adalah sebuah pendekatan baru dalam penerjemahan mesin yang menggunakan arsitektur *Recurrent neural networks* (RNNs) pada bagian *encoder* dan *decoder*-nya dan menggunakan korpus paralel sebagai bahan untuk *training data*. Performa NMT terbukti lebih baik dibanding *Statistical Machine Translation* (SMT) pada eksperimen 30 jenis penerjemahan (Junczyz-Dowmunt, dkk. 2016).

NMT yang banyak digunakan saat ini oleh para peneliti mesin penerjemah adalah NMT berbasis *attention*. NMT berbasis *attention* suatu pendekatan yang efisien dalam membaca keseluruhan kalimat atau paragraf, untuk mendapatkan konteks dan inti dari informasi yang sedang diekspresikan pada kalimat *input*, kemudian menghasilkan kata-kata yang diterjemahkan satu per satu. Setiap waktu berfokus pada bagian yang berbeda dari kalimat *input* untuk mengumpulkan rincian semantik yang dibutuhkan untuk menghasilkan kata *output* berikutnya (Goodfellow, dkk. 2016).

Penelitian SMT pada penerjemahan dua arah pada bahasa Inggris – Indonesia dengan melakukan perbandingan menggunakan RNNs dan N-gram sebagai model bahasanya (Hermanto, dkk. 2015). Selain itu, juga telah dilakukan penelitian penerjemahan bahasa Jepang – Indonesia dengan menggunakan SMT dan NMT *attention* (Adiputra & Arase, 2017).

Bahasa Lampung sendiri sebagai suatu bahasa daerah yang wajib dilestarikan. Oleh karena itu, penelitian ini bertujuan melakukan ujicoba penerjemahan secara otomatis dengan menggunakan pendekatan NMT berbasis *attention* dan bahasa

Lampung dialek *api* sebagai objek yang digunakan dalam penelitian ini. Alat yang digunakan adalah THUMT-theano (Zhang, J. dkk. 2017).

Tujuan penelitian ini adalah untuk mendapatkan model dan visualisasi penerjemahan bahasa Lampung-Indonesia dengan pendekatan NMT berbasis *attention*. Untuk kemudahan dalam penelitian maka dipilih salah satu dialek dalam bahasa Lampung yaitu dialek *api*, akan tetapi pada penelitian selanjutnya akan dilakukan pengamatan terhadap dialek *nyo*. Penelitian ini terkait pada dua domain utama yaitu pemrosesan bahasa alami dan bahasa daerah.

METODOLOGI

1. Bahan Penelitian

Bahan yang digunakan untuk melakukan uji coba penerjemahan kalimat bahasa Lampung dengan pendekatan NMT *attention* adalah korpus paralel bahasa Lampung - Indonesia. Faktanya tidak ditemukan korpus paralel bahasa Lampung-Indonesia. Ketiadaan korpus paralel bahasa Lampung - Indonesia mengharuskan peneliti untuk membuat korpus paralel secara manual berdasarkan buku bahasa Lampung sebagai referensi sekolah, buku percakapan sehari-hari bahasa Lampung dan buku afiksasi verba bahasa Lampung.

Ketiadaan korpus paralel bahasa Lampung-Indonesia diatasi dengan cara mengumpulkan buku acuan bahasa Lampung dialek *api* yang digunakan di provinsi Lampung. Setelah pengumpulan buku berhasil dilakukan, maka langkah selanjutnya adalah mengetikkan ulang bahasa Lampung dan terjemahannya dalam bahasa Indonesia dengan menggunakan aplikasi *gedit* yang ada di *Linux Ubuntu* 16.04, kemudian dilakukan penggalan informasi statistik data dari korpus paralel tersebut.

Korpus paralel bahasa Lampung-Indonesia adalah kalimat-kalimat bahasa

Lampung yang telah memiliki padanan artinya dalam bahasa Indonesia. Kalimat-kalimat yang ada dalam penelitian ini berasal dari buku acuan belajar bahasa Lampung, LKS belajar bahasa Lampung dan percakapan sehari-hari bahasa Lampung. Kalimat-kalimat tersebut telah diketik secara manual untuk membuat korpus paralel dan jumlahnya mencapai 3000 kalimat.

2. Alat yang Digunakan

Proses pembangunan, pengujian model untuk NMT *attention* digunakan perangkat keras dengan spesifikasi seperti berikut :

- Processor: Intel(R) Core(TM) i7 3540 M CPU @ 3.00Hz (4 CPUs)
- Jumlah core: 4 core
- Memory: 300 GB HDD dan 8 GB RAM

Perangkat lunak yang digunakan untuk pembangunan dan pengujian model NMT *attention*, dengan memanfaatkan korpus paralel, adalah *Linux Ubuntu* 16.04 LTS sebagai Sistem operasi, THUMT-theano sebagai kakas untuk *training data* membangun model translasi, pengujian translasi. Bahasa pemrograman Java sebagai alat memvisualisasi hasil penerjemahan. *gedit* sebagai kakas untuk *teks editor* (utf-8). Bahasa pemrograman Python untuk *text processing*.

3. Prosedur

Pendekatan NMT *attention* yang digunakan untuk penerjemahan kalimat bahasa Lampung ke bahasa Indonesia dilakukan dengan melalui beberapa tahapan solusi sebagai berikut:

- a. Mempersiapkan data. Adapun yang dimaksud data disini adalah (1) data *training* berupa korpus paralel, (2) data validasi yaitu sekumpulan kalimat yang digunakan untuk melakukan validasi terhadap hasil model yang diperoleh selama proses *training data*. Untuk data validasi, pada penelitian ini menggunakan

- skema *5-folds validation*, (3) data uji yaitu data / kalimat baru yang tidak terdapat pada paralel korpus.
- b. Melakukan *training data* bahasa Lampung-Indonesia dengan bantuan kakas THUMT-theano, dengan tujuan menghasilkan sebuah model penerjemahan bahasa Lampung-Indonesia. Sebelum *training data* dilakukan, peneliti harus menentukan nilai-nilai yang ada pada *file* konfigurasi.
 - c. Melakukan uji model atau disebut *decoding*, yaitu dengan cara memasukan ke dalam kakas THUMT-theano berupa data uji yaitu 25 kalimat tunggal tanpa *out-of-vocabulary* (OOV), 25 kalimat tunggal dengan OOV, 25 kalimat majemuk tanpa OOV dan 25 kalimat majemuk dengan OOV. Tujuan data uji ini adalah untuk mendapat hasil penerjemahan dari data uji berupa file berbahasa Indonesia dengan berekstensi *.trans*.
 - d. Melakukan pengukuran atau evaluasi hasil penerjemahan mesin NMT *attention* berekstensi *.trans* dengan kalimat referensi yang telah disediakan dengan bantuan kakas THUMT dan satuan pengukuran yang digunakan adalah *bilingual evaluation under study* (BLEU).
 - e. Menganalisis lebih dalam hasil penerjemahan mesin dan memberikan visualisasi hasil penerjemahan mesin dengan bantuan kakas THUMT. Bagian ini bersifat opsional.

Tahapan solusi diatas diilustrasikan pada Gambar 1 di bawah ini.

Pada sistem penerjemahan bahasa Lampung – Indonesia dengan pendekatan NMT *attention*, desain eksperimen yang dilakukan adalah:

1. Eksperimen pertama melakukan *training data*, *testing data* dan visualisasi hasil pada penerjemahan bahasa Lampung – Indonesia dengan pendekatan NMT model *Attention*

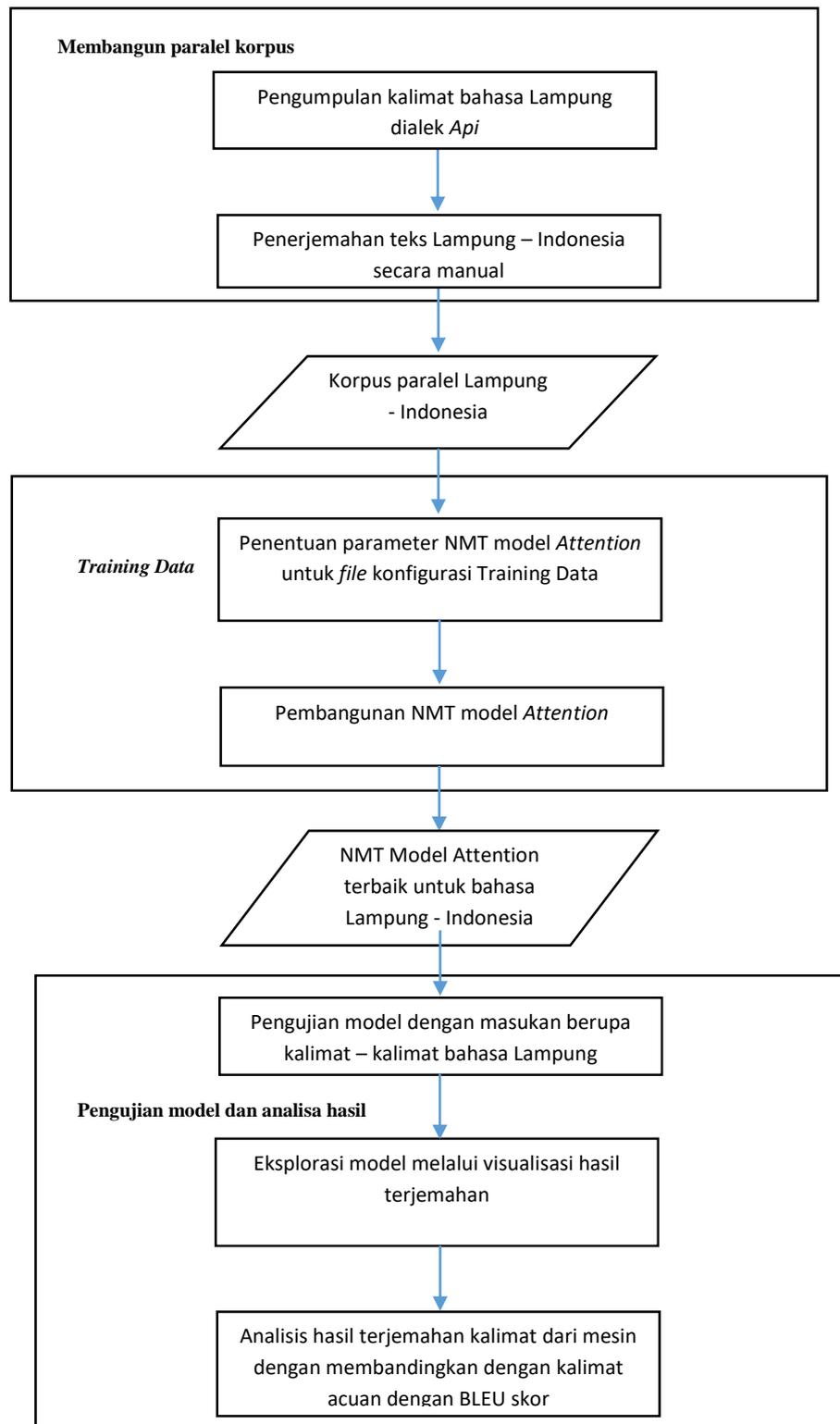
dengan ukuran *hidden layer n* adalah 500 dan ukuran dimensi vektor *word embedding m* adalah 310.

2. Eksperimen kedua melakukan *training data*, *testing data* dan visualisasi hasil pada penerjemahan bahasa Lampung – Indonesia dengan pendekatan NMT model *Attention* dengan ukuran *hidden layer n* adalah 1000 dan ukuran dimensi vektor *word embedding m* adalah 620.
3. Eksperimen ketiga melakukan *training data*, *testing data* dan visualisasi hasil pada penerjemahan bahasa Lampung – Indonesia dengan pendekatan NMT model *Attention* dengan ukuran *hidden layer n* adalah 1500 dan ukuran dimensi vektor *word embedding m* adalah 930.

HASIL DAN PEMBAHASAN

Eksperimen pertama, kedua dan ketiga dilakukan dengan menggunakan data latih 3000 kalimat, validasi eksperimen dilakukan dengan mekanisme *5-folds validation* pada 3000 kalimat tersebut kemudian data uji yang digunakan 100 kalimat terdiri dari 25 kalimat tunggal tanpa *out of vocabulary* (OOV), 25 kalimat tunggal dengan OOV, 25 kalimat majemuk tanpa OOV dan 25 kalimat majemuk dengan OOV.

Pengaturan *file* konfigurasi dilakukan sebelum masuk ke proses *training data*. Pengaturan *file* konfigurasi adalah penentuan parameter yang akan digunakan pada proses *training data*. Penetapan dilakukan dengan memberikan nilai-nilai yang berkesesuaian dan disimpan dalam sebuah *file* bernama THUMT.config.



Gambar 1. Tahapan solusi pembangunan model NMT *attention* bahasa Lampung – Indonesia

Tabel 1. Hasil pengujian dengan NMT *attention*

Nilai BLUE (%)	X1	X2	X3
25 Kalimat Tunggal tanpa OOV	37.95	62.46	57.95
25 Kalimat Tunggal dengan OOV	27.94	35.76	36.58
25 Kalimat Majemuk tanpa OOV	55.09	68.66	54.35
25 Kalimat Majemuk dengan OOV	30.55	40.96	36.8
Rata-rata	37.8825	51.96	46.42

X1 adalah eksperimen pertama, X2 adalah eksperimen kedua dan X3 adalah eksperimen ketiga. Secara eksplisit, informasi dari tabel 1 terlihat bahwa eksperimen kedua memberikan hasil evaluasi terbaik pada penerjemahan bahasa Lampung-Indonesia. Berdasarkan hasil pengamatan terhadap hasil penerjemahan bahasa Lampung – Indonesia pada 100 kalimat uji maka dapat dikelompokan / dikategorikan kesalahan penerjemahan yang terjadi pada eksperimen ini, yaitu :

1. *Extra words*, artinya mesin penerjemah memberikan hasil terjemahan mempunyai jumlah kata yang lebih banyak dari kalimat masukan.
2. *Unrelated words*, artinya mesin penerjemah memberikan hasil kata-kata yang tidak terkait dengan konteks kalimat.
3. *Negation Reversion*, artinya mesin penerjemah memberikan hasil penerjemahan yang berlawanan dengan makna yang sebenarnya atau bersifat kontradiksi.
4. *Word Omission*, artinya mesin penerjemah memberikan hasil penerjemahan yang jumlah katanya lebih sedikit dari kalimat masukan.

5. *Word Repetition*, artinya mesin penerjemah memberikan hasil penerjemahan yang mengandung unsur kata-kata yang terjadi secara berulang-ulang.
6. *No Problem*, lebih tepatnya ini bukanlah sebuah masalah melainkan fakta yang diperoleh peneliti yaitu mesin penerjemah memberikan hasil penerjemahan yang tepat sama sesuai acuan.

Tingkat kemunculan kata juga menjadi perhatian tersendiri guna digali. Hal tersebut akan membantu pada tahapan analisis selanjutnya. Sebuah kata yang sangat jarang muncul dalam korpus berpotensi terindikasi sebagai *unknown word* (UNK) pada model translasi sehingga sistem mengindikasikan sebagai OOV. Pengamatan terhadap korpus paralel bahasa Lampung-Indonesia yang terdiri dari 3000 kalimat bahasa Lampung – Indonesia diperoleh fakta bahwa lebih dari 89 % kata-kata yang terdapat pada *paralel copus* menunjukkan tingkat kemunculan yang sangat rendah yaitu kurang dari atau sama dengan 10 kali pemunculan. Informasi detail terikat kemunculan kata disajikan pada tabel 2 di bawah ini.

Tabel 2. Kemunculan kata pada korpus paralel

Tingkat Kemunculan	Jumlah Kata Muncul > 10 kali (%)	Jumlah Kata Muncul <= 10 kali (%)
Bahasa Lampung	7,81	92,19
Bahasa Indonesia	10,1	89,1

A. *Analisa Hasil Penerjemahan pada Kalimat Tunggal tanpa OOV (TTOOV)*

TTOOV diujicobakan pada eksperimen pertama, kedua dan ketiga. Hasil pengamatan kesalahan penerjemahan yang terjadi pada TTOOV disajikan secara rinci pada tabel 3 di bawah ini.

Pada masing-masing kategori kesalahan yang terjadi, pada pengujian 25 kalimat tunggal tanpa OOV diperoleh 3 utama fakta yaitu :

- Pada bagian *no problem*, kalimat yang berhasil diterjemahkan yaitu 12 kalimat.
- Pada bagian *Unrelated words*, hal ini disebabkan kondisi kosa kata yang digunakan pada penelitian ini, kosa kata termasuk pada tingkat kemunculan yang sangat rendah sehingga sistem NMT *attention* mendeteksi sebagai *unknown word*.
- Pada bagian jenis *error* lainnya, tetap ada walau dengan angka yang kecil.

Tabel 3. Hasil pengujian pada TTOOV

TTOOV	Eksperimen Pertama	Eksperimen Kedua	Eksperimen Ketiga
<i>Extra words</i>	7	2	4
<i>Unrelated words</i>	13	11	13
<i>No Problem</i>	3	12	11
<i>Negation Reversion</i>	4	2	1
<i>Word Omission</i>	4	3	1
<i>Word Repetition</i>	2	1	0

Selanjutnya, diberikan contoh hasil eksperimen pada 25 kalimat tunggal tanpa OOV.

Tabel 4. Contoh hasil penerjemahan pada TTOOV

Bahasa Lampung	nyak mak mepoh kawai di wai
Bahasa Indonesia	saya tidak mencuci baju di sungai
Eksperimen Pertama TTOOV	saya tidak akan segelas pada di sungai
Eksperimen Kedua TT OOV	saya sedang mandi di sungai
Eksperimen Ketiga TT OOV	saya tidak mencuci baju di sungai

Pada hasil eksperimen pertama diketahui bahwa ada dua kata yaitu *mepoh* dan *kawai*. Berdasarkan informasi dari kemunculan kata, diketahui kata *mepoh* muncul 7 kali dan kata *kawai* muncul 23 kali. Hal utama yang terjadi pada eksperimen pertama ini, konteks kalimat belum berhasil terekam dengan baik pada eksperimen pertama sehingga diperoleh hasil penerjemahan yang tidak berhubungan secara konteks kalimat dan terdapat extra word, dari yang seharusnya hanya 2 kata *mepoh* dan *kawai* yang terdeteksi salah menghasilkan tiga kata terjemahan yang tidak berhubungan (*unrelated words*). Pada eksperimen kedua, terlihat hasilnya terdapat satu kata yang belum diterjemahkan yaitu *mepoh*. Kondisi jauh lebih baik dari eksperimen pertama karena di eksperimen kedua tidak terjadi *extra word*. Faktor utama pada eksperimen kedua adalah tingkat kemunculan kata *mepoh* yang masih rendah yaitu 7 kali. Pada eksperimen ketiga diperoleh hasil sesuai yang diinginkan.

Visualisasi hasil penerjemahan hanya diambil dari eksperimen pertama dan ketiga pada TTOOV diberikan pada Gambar 2 di bawah ini.

B. Analisa Hasil Penerjemahan pada Kalimat Tunggal dengan OOV (TDOOV)

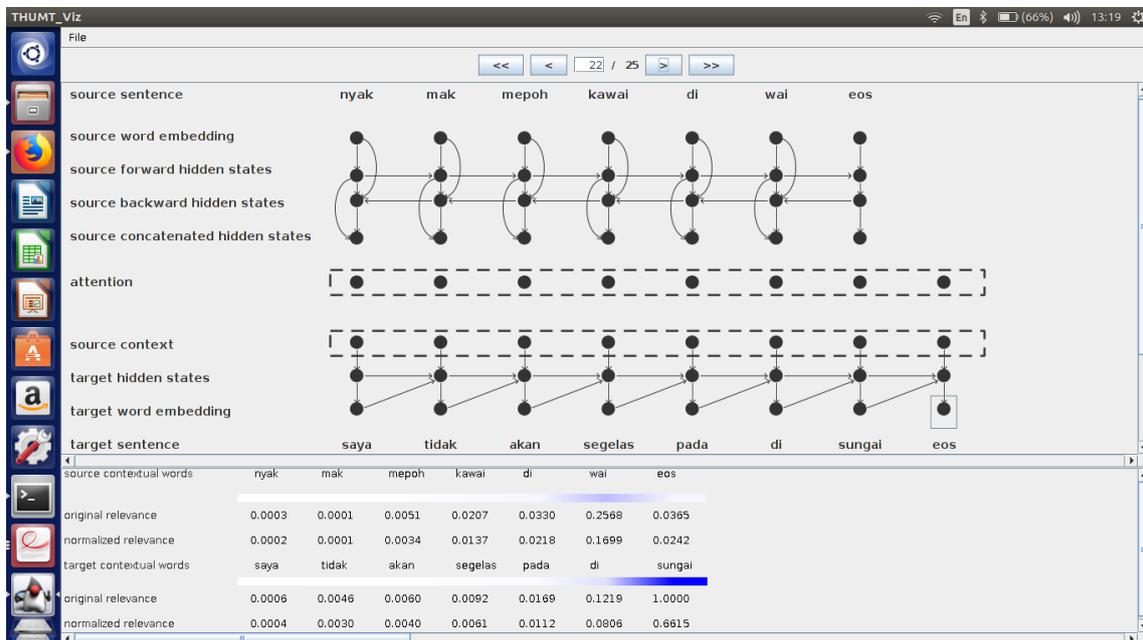
Dengan membandingkan perolehan nilai dari ketiga eksperimen di tabel 5 di bawah ini. Pada masing-masing kategori kesalahan yang terjadi, pada pengujian 25 kalimat tunggal dengan OOV diperoleh 3 utama fakta yaitu :

- OOV yang sengaja diberikan pada kalimat tunggal memberikan efek penurunan nilai pada bagian kategori *no problem* sehingga tidak ada satupun mesin NMT *attention* memberikan hasil penerjemahan yang sesuai dengan kalimat acuan.
- Fakta kosa kata yang jarang muncul mendominasi faktor penyebab banyaknya mesin NMT memberikan hasil terjemahan tetapi masuk dalam kategori *unrelated words*, ditambah hal ini dipengaruhi OOV yang sengaja diberikan di eksperimen 1, 2 dan 3.

- Pada bagian *word omission* dan *word repetition* terjadi penurunan nilai, sedangkan pada bagian *extra words* sedikit naik nilainya.

Tabel 5. Hasil pengujian pada TDOOV

TDOOV	Eksperimen Pertama	Eksperimen Kedua	Eksperimen Ketiga
<i>Extra words</i>	2	3	3
<i>Unrelated words</i>	19	22	23
<i>No Problem</i>	0	0	0
<i>Negation Reversion</i>	1	3	2
<i>Word Omission</i>	7	5	5
<i>Word Repetition</i>	1	1	0



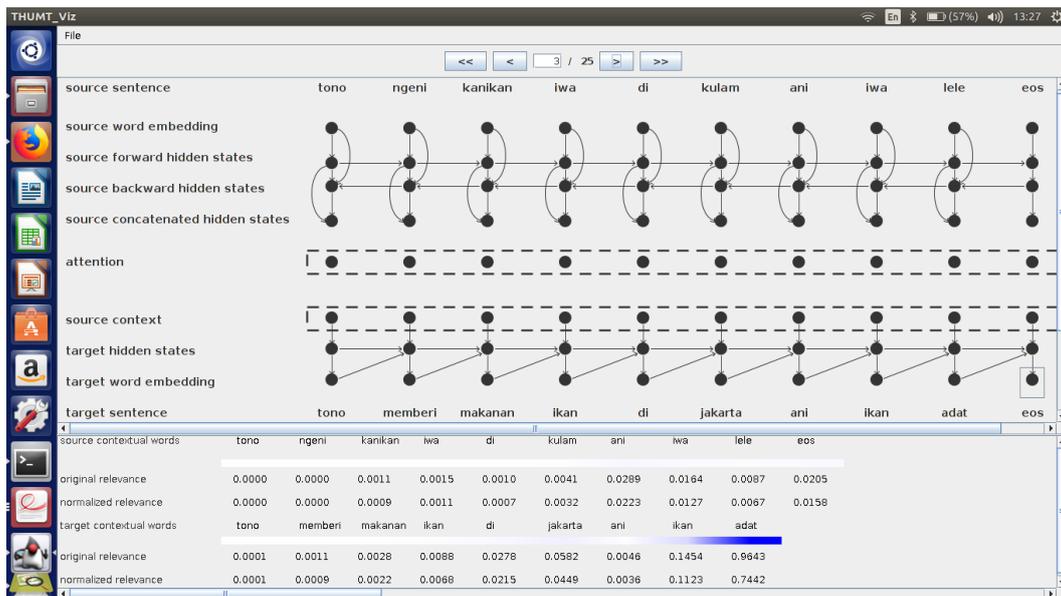
Gambar 2. Visualisasi hasil eksperimen pertama pada TDOOV

Selanjutnya, diberikan contoh hasil eksperimen pada 25 kalimat tunggal dengan OOV. Kalimat “*tono ngeni kanikan iwa di kulam ani iwa lele*” adalah contoh kalimat yang sengajar diberikan sebuah kata OOV yaitu *lele*. Efek kata *lele* pada eksperimen 1, 2 dan 3 mengacaukan konteks kata-kata yang dekat dengan kata “*lele*”, hal ini karena di penelitian ini tidak menyinggung solusi atas keberadaan OOV. Hal ini akan menjadi kajian selanjutnya. Hasil eksperimen pertama menunjukkan terjadi kekeliruan penerjemahan kata *toni* menjadi *hasan*, hal ini disebabkan sistem NMT membaca secara konteks lebih dekat ke kata “*hasan*”. PoS Tag menjadi alternatif pilihan untuk penelitian selanjutnya guna meningkatkan kualitas penerjemahan.

Tabel 6. Contoh hasil penerjemahan pada TDOOV

Bahasa Lampung	tono ngeni kanikan iwa di kulam ani iwa lele
Bahasa Indonesia	tono memberi makanan ikan di kolam ikan lele
Eksperimen Pertama TD OOV	tono memberi makanan ikan di kolam lukman
Eksperimen Kedua TD OOV	tono memberi makanan ikan di jakarta ani ikan adat
Eksperimen Ketiga TD OOV	tono memberi makanan ikan di kolam pertemuan ikan

Visualisasi hasil penerjemahan eksperimen kedua pada TDOOV diberikan pada Gambar 3 di bawah ini.



Gambar 3. Visualisasi hasil eksperimen kedua pada TDOOV

C. Analisa Hasil Penerjemahan pada Kalimat Majemuk tanpa OOV (MTOOV)

Hasil dari tabel 7 di bawah ini menunjukkan pengujian pada kalimat majemuk menunjukkan tren nilai BLEU yang lebih baik dari kalimat tunggal karena kalimat majemuk mempunyai

jumlah kata yang lebih banyak dari kalimat tunggal. Dengan keberadaan kata-kata ini akan memberi kontribusi terhadap pembentukan konteks kalimat sesuai yang diinginkan. Dengan membandingkan perolehan nilai dari ketiga eksperimen di tabel 7 di bawah ini. Pada masing-masing kategori kesalahan yang terjadi, pada

pengujian 25 kalimat majemuk tanpa OOV diperoleh 2 utama fakta yaitu :

- Faktor tingkat kemunculan kata-kata yang sangat jarang sesuai info tabel V.3 masih mendominasi kondisi eksperimen kalimat majemuk walaupun tanpa OOV sehingga kemunculan *unrelated words* tetap tinggi di semua eksperimen dan kemunculan *extra words*.
- Pada kalimat majemuk, potensi terjadinya *negation reversion*, *word omission* dan *word repetition* menjadi berkurang akibat kontribusi dari kata-kata dari kalimat majemuk tersebut memberikan efek pembentukan konteks kalimat yang lebih baik.

Tabel 7. Hasil eksperimen pada kondisi MTOOV

MTOOV	Eksperimen Pertama	Eksperimen Kedua	Eksperimen Ketiga
<i>Extra words</i>	0	1	2
<i>Unrelated words</i>	19	17	18
<i>No Problem</i>	4	8	6
<i>Negation Reversion</i>	0	0	1
<i>Word Omission</i>	8	1	2
<i>Word Repetition</i>	2	0	1

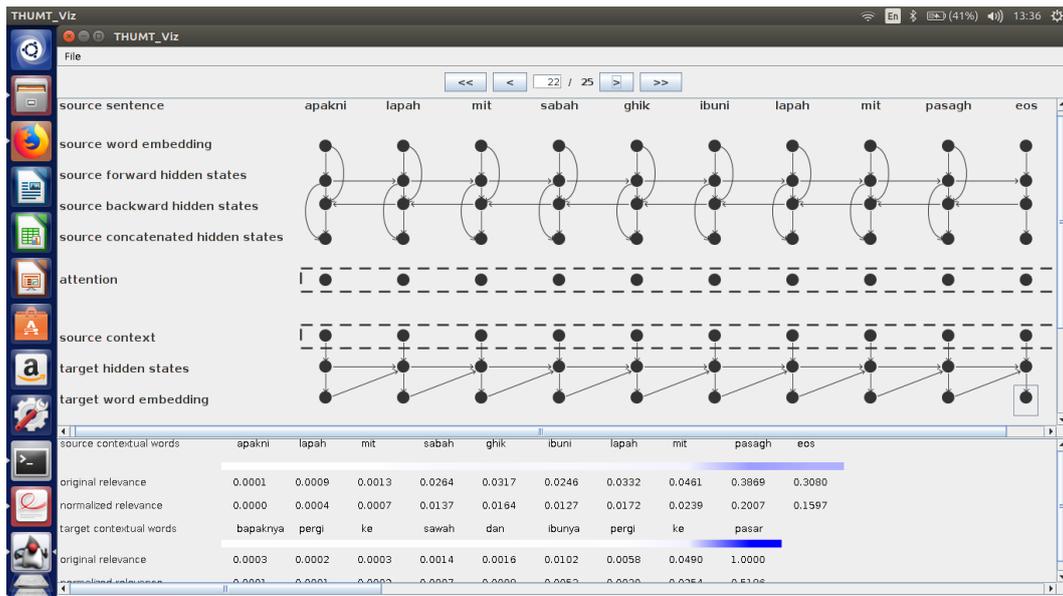
Selanjutnya, diberikan contoh hasil eksperimen pada 25 kalimat majemuk tanpa OOV.

Tabel 8. Contoh hasil penerjemahan pada MTOOV

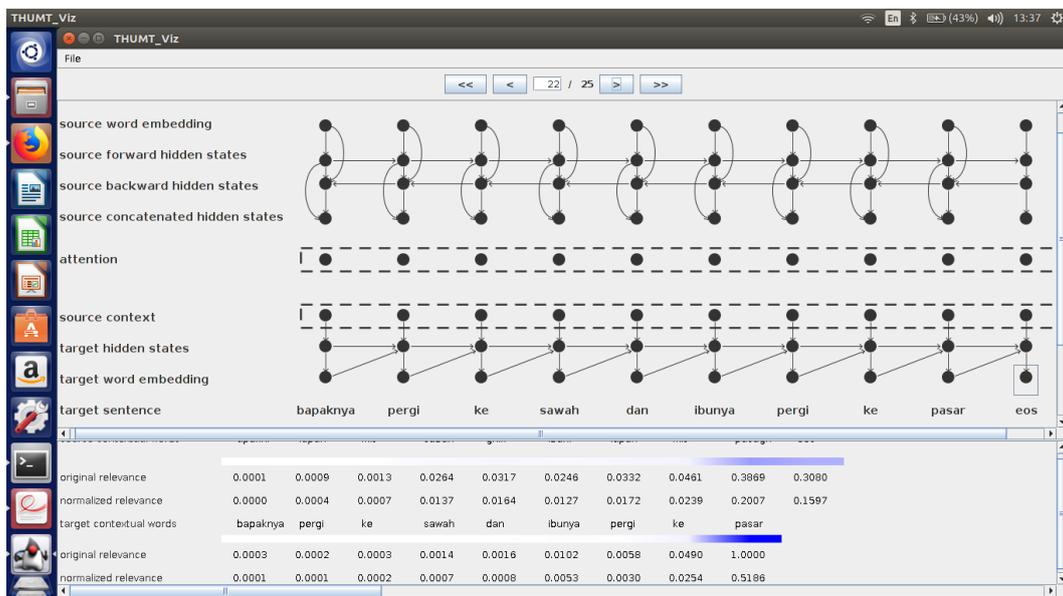
Bahasa Lampung	apakni lapah mit sabah ghik ibuni lapah mit pasagh
Bahasa Indonesia	bapaknya pergi ke sawah dan ibunya pergi ke pasar
Eksperimen Pertama MT OOV	bapaknya pergi ke sawah dan ibunya pergi ke pasar
Eksperimen Kedua MT OOV	bapaknya pergi ke sawah dan ibunya pergi ke pasar
Eksperimen Ketiga MT OOV	ayahnya pergi ke sawah dan ibunya pergi ke pasar

Tabel 8 memberikan sebuah contoh yang sempurna dari sisi hasil decoding dan hampir sempurna dari sisi evaluasi. Dikatakan sempurna karena hasil eksperimen 1, 2 dan 3 pada sebuah contoh di 25 kalimat majemuk tanpa OOV ini memberikan hasil yang benar secara makna akan tetapi dikatakan hampir sempurna dari sisi evaluasi karena di hasil eksperimen yang ketiga muncul kata “ayahnya”. Kata “ayahnya” akan dianggap berbeda dengan kata “bapaknya” ketika dilakukan pengecekan menggunakan nilai BLEU, sehingga untuk penelitian selanjutnya sebaiknya dibuatkan kalimat target/tujuan itu terdiri dari dua jenis untuk mengakomodasi beberapa kata yang mempunyai makna sama seperti kata “ayahnya” dan “bapaknya”.

Untuk tabel 8, berikut di bawah ini disajikan visualisasi hasil penerjemahan eksperimen pertama pada MT OOV diberikan pada gambar di bawah ini, akan tetapi hasil visualisasi belum utuh karena keterbatasan ruang maka disajikan dalam dua Gambar 4 dan 5 di bawah ini :



Gambar 4. Visualisasi hasil eksperimen pertama pada MTOOV (Bagian pertama)



Gambar 5. Visualisasi hasil eksperimen pertama pada MTOOV (Bagian kedua)

D. Analisa Hasil Penerjemahan pada Kalimat Majemuk dengan OOV (MDOOV)

dipaparkan pada bagian Tabel 9 di bawah ini.

Efek penambahan satu atau dua kata yang bersifat OOV pada kalimat majemuk ini sangat signifikan akibatnya, hal ini tercermin dari perubahan yang terjadi pada bagian *no problem*, tidak ada satupun kalimat yang sesuai dengan kalimat acuan dan secara umum sama seperti yang

Tabel 9. Hasil Eksperimen Pada Kondisi MDOOV

MD OOV	Eksperimen Pertama	Eksperimen Kedua	Eksperimen Ketiga
<i>Extra words</i>	4	4	4
<i>Unrelated words</i>	23	23	24
<i>No Problem</i>	0	0	0
<i>Negation Reversion</i>	1	2	2
<i>Word Omission</i>	14	3	8
<i>Word Repetition</i>	3	2	4

Selanjutnya, diberikan contoh hasil eksperimen pada 25 kalimat majemuk dengan OOV.

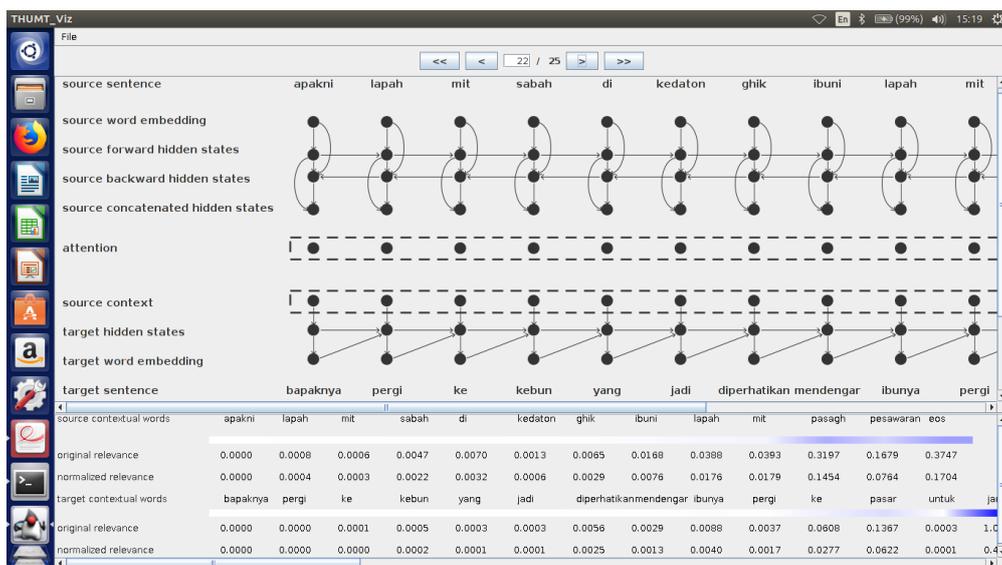
Tabel 10. Contoh Hasil Penerjemahan pada MDOOV

Bahasa Lampung	apakni lapah mit sabah di kedaton ghik ibuni lapah mit pasagh pesawaran
Bahasa Indonesia	bapaknya pergi ke sawah di kedaton dan ibunya pergi ke pasar pesawaran
Eksperimen Pertama MT OOV	bapaknya pergi ke kebun yang jadi diperhatikan mendengar ibunya pergi ke pasar untuk jari
Eksperimen	belanda pergi ke sawah di

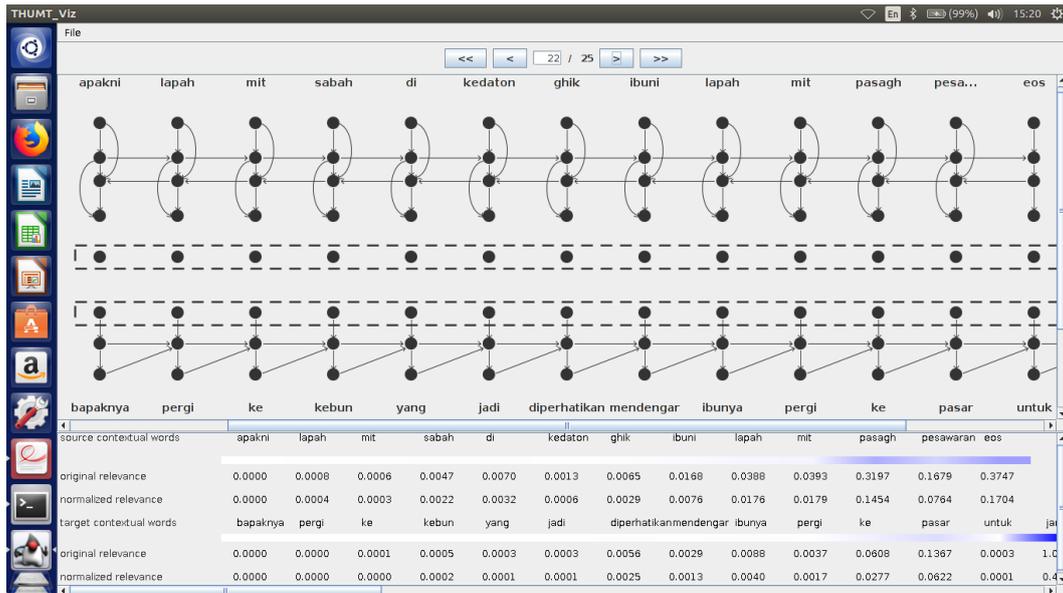
Kedua MT OOV	berdua besar-besar dan ibunya pergi ke pasar guna
Eksperimen Ketiga MT OOV	ayahnya pergi ke sawah di kalender hidup ibunya pergi ke pasar untuk

Tabel 10 memberikan sebuah contoh hasil terjemahan NMT attention tidak ada yang sesuai dengan kalimat acuan dalam bahasa Indonesia, akan tetapi perlu diamati satu per satu hasil terjemahan dari eksperimen 1, 2 dan 3 tersebut di atas. Hasil eksperimen 1 menunjukkan kacaunya konteks kalimat yang dihasilkan ditandai oleh adanya kesalahan kategori *extra words*, *unrelated words* dan diperparah oleh efek dua OOV yaitu kata “kedaton” dan “pesawaran”. Hasil eksperimen 2 jauh lebih baik karena *extra words*-nya berkurang tetapi masih kacau konteks kalimatnya dan eksperimen 3 walau hasilnya masih kacau konteks kalimatnya tetapi tidak ada *extra words*.

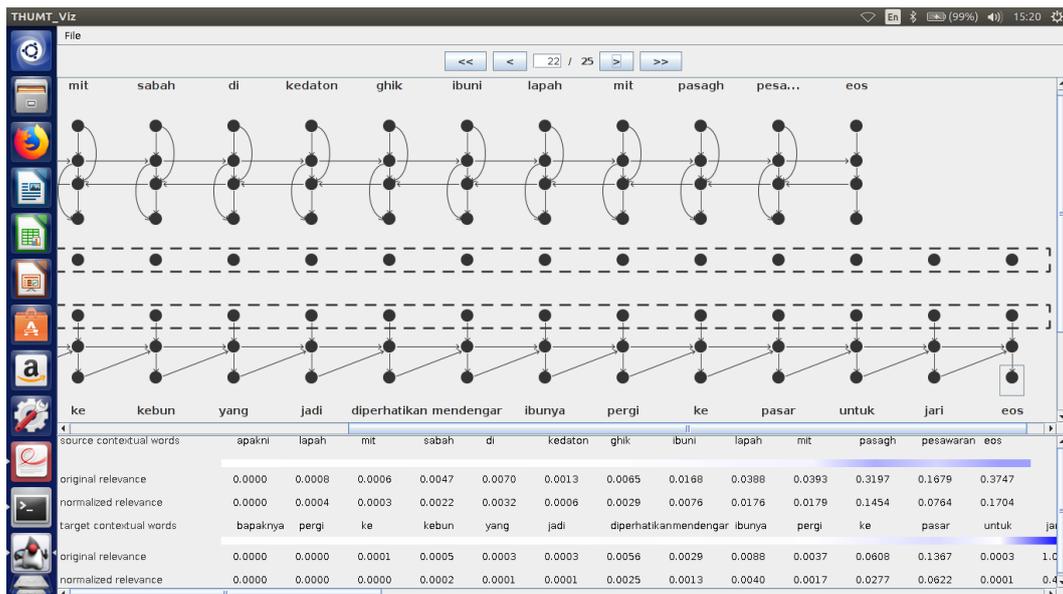
Untuk tabel 10 berikut di bawah ini disajikan visualisasi hasil penerjemahan eksperimen pertama pada MDOOV diberikan pada gambar di bawah ini, akan tetapi hasil visualisasi belum utuh karena keterbatasan ruang maka disajikan dalam Gambar 6, 7, 8 dan 9 berikut ini :



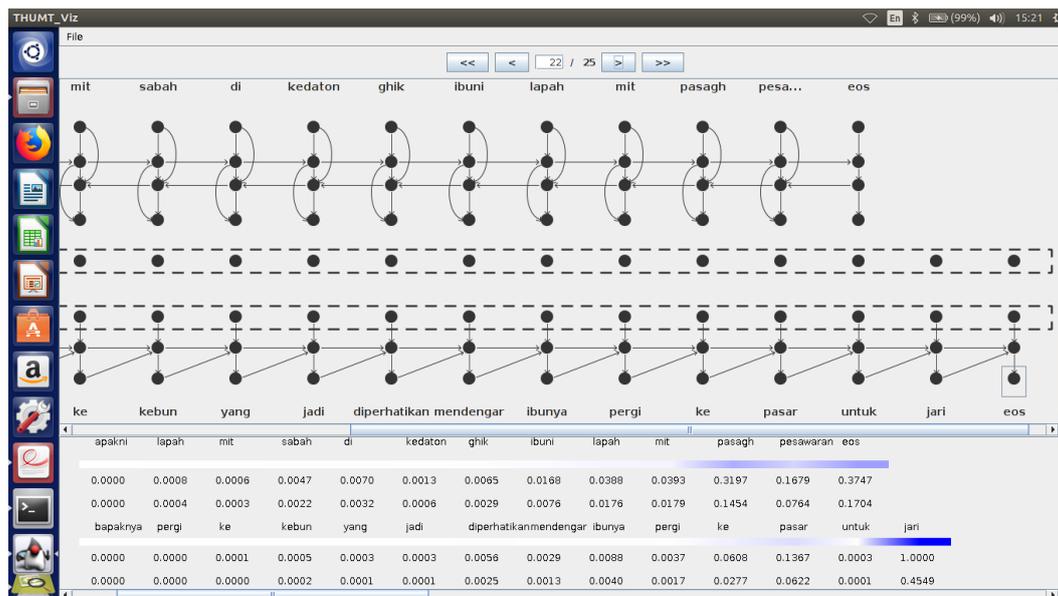
Gambar 6. Visualisasi hasil eksperimen pertama pada MDOOV (Bagian pertama)



Gambar 7. Visualisasi hasil eksperimen pertama pada MDOOV (Bagian kedua)



Gambar 8. Visualisasi hasil eksperimen pertama pada MDOOV (Bagian ketiga)



Gambar 9. Visualisasi hasil eksperimen pertama pada MDOOV (Bagian keempat)

KESIMPULAN DAN SARAN

Berdasarkan pemaparan hasil dan pembahasan diperoleh kesimpulan yaitu :

1. Penerjemahan bahasa Lampung-Indonesia dengan pendekatan NMT *attention* dengan perolehan rata-rata nilai akurasi BLEU sebesar 51.96 %. Pendekatan ini menggunakan 3000 kalimat pada korpus paralel dan konfigurasi dimensi yang terbaik diperoleh pada menggunakan dimensi *word embedding* sebesar 620 dan dimensi *hidden layer* pada bagian *ecoder* dan *decoder*-nya sebesar 1000.
2. Masalah utama yang dihadapi pada penerjemahan dengan pendekatan NMT *attention* pada bahasa Lampung-Indonesia adalah tingkat kemunculan kosa-kata yang rendah dibawah 19 akan menyebabkan sistem penerjemah membacanya sebagai *unknown word* (UNK), jumlah korpus paralel yang masih sedikit yaitu 3000 kalimat. Selain itu, perlu penanganan khusus untuk kosa kata yang memang belum ada di korpus atau sebagai OOV.

Penelitian ini adalah titik awal untuk melakukan penelitian pada bahasa

Lampung dari aspek pemrosesan bahasa alami khususnya pada mesin penerjemah, sehingga masih terbuka lebar untuk dilanjutkan. Saran yang dapat diambil dari penelitian ini untuk penelitian selanjutnya yaitu untuk penelitian NMT *attention* guna penerjemahan bahasa Lampung selanjutnya, disarankan menggunakan teknik yang dapat mereduksi kosa kata menjadi bentuk lain seperti sub kata seperti yang telah dilakukan oleh Rico Senrich dan kawan-kawan dari *University of Edinburgh*.

DAFTAR PUSTAKA

- Adiputra, C.K. dan Arase, Y. (2017) : *Performance of Japanese-to-Indonesian Machine Translation on Different Models*. Retrieved from http://www.anlp.jp/proceedings/annual_meeting/2017/pdf_dir/A5-5.pdf.
- Bhattacharyya, P. (2015). *Machine Translation*. CRC Press, Mumbai, India, 25-27.
- Goodfellow, I. Bengio, Y. dan Courville, (2016): *Deep Learning*, MIT Press, London, 462-464.

- Hermanto, A., Adji, T. B., & Setiawan, N. (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*. Retrieved from https://workshop2016.iwslt.org/downloads/IWSLT_2016_paper_4.pdf
- Zhang, J., Ding, Y., Shen, S., Cheng, Y., Sun, M., Luan, H., dan Liu, Y. (2017): *THUMT-theano: An Open Source Toolkit for Neural Machine Translation*. Retrieved from <https://arxiv.org/abs/1706.06415>.
- Anonim, 2017. *Nestapa Guru Bahasa Lampung*. Retrieved from <http://www.lampost.co/berita-nestapa-guru-bahasa-lampung>.
- Anonim, 2016. *139 Bahasa Daerah di Indonesia Terancam Punah*. Retrieved from <http://nasional.republika.co.id/berita/nasional/umum/16/08/02/ob9t2h383-139-bahasa-daerah-di-indonesia-terancam-punah>.



Halaman Kosong