PEMBANGUNAN *BILINGUAL DICTIONARY* BAHASA JAWA DAN SUNDA MENGGUNAKAN *MONOLINGUAL CORPORA*

Laporan Tugas Akhir I

Disusun sebagai syarat kelulusan mata kuliah
IF4091/Tugas Akhir I dan Seminar

Oleh
SEKAR LARASATI MUSLIMAH
NIM : 13517114

PROGRAM STUDI TEKNIK INFORMATIKA
SEKOLAH TEKNIK ELEKTRO & INFORMATIKA
INSTITUT TEKNOLOGI BANDUNG
Desember 2020

PEMBANGUNAN *BILINGUAL DICTIONARY* BAHASA JAWA DAN SUNDA MENGGUNAKAN *MONOLINGUAL CORPORA*

Laporan Tugas Akhir I

Oleh

SEKAR LARASATI MUSLIMAH

NIM : 13517114

Program Studi Teknik Informatika

Sekolah Teknik Elektro dan Informatika

Institut Teknologi Bandung

Bandung, 8 Desember 2020

Mengetahui,

Pembimbing,

Dr. Eng. Ayu Purwarianti ST., MT.

NIP. 19770127 200801 2 011

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB I

PENDAHULUAN

Bab Pendahuluan menguraikan landasan kerja dan penulisan tugas akhir. Bab ini memuat latar belakang, rumusan masalah, tujuan, batasan masalah, metodologi, dan jadwal pelaksanaan tugas akhir.

I.1 Latar Belakang

Di era digital ini informasi dan teknologi berkembang begitu pesat sebagai akibat dari revolusi industri 4.0 yang terjadi di berbagai belahan dunia. Salah satu dampak yang dirasakan masyarakat akibat perkembangan tersebut adalah setiap orang dapat mengakses berbagai informasi tanpa dibatasi ruang dan waktu. Informasi dapat dengan mudah keluar dan masuk dari dan ke sebuah negara.

Mudahnya akses informasi juga menyebabkan mudahnya budaya asing masuk ke Indonesia. Hal yang perlu diwaspadai dari fenomena tersebut adalah eksistensi budaya Indonesia sendiri terutama budaya daerah sebagai aset negara. Budaya daerah harus mampu bersaing dengan budaya - budaya asing yang masuk ke Indonesia agar eksistensinya tetap terjaga. Salah satu aset budaya yang perlu dijaga adalah bahasa daerah.

Berdasarkan data yang tercatat oleh UNESCO, Indonesia menempati posisi kedua setelah Papua Nugini sebagai negara dengan jumlah bahasa daerah terbanyak. Dilansir melalui laman *Kompas.com*, Badan Pengembangan dan Pembinaan Bahasa serta Kementerian Pendidikan dan Kebudayaan, Indonesia menyatakan bahwa Indonesia memiliki 718 bahasa daerah per tahun 2020. Berdasarkan data kajian Badan Pengembangan dan Pembinaan Bahasa tercatat sejak tahun 2011 hingga 2019 terdapat 11 bahasa daerah Indonesia yang punah. 11 bahasa daerah yang punah adalah Bahasa Tandia (Papua Barat), Bahasa Mawes (Papua), Bahasa Kajeli/Kayeli (Maluku), Bahasa Piru (Maluku), Bahasa Moksela (Maluku),

Bahasa Palumata (Maluku), Bahasa Ternateno (Maluku Utara), Bahasa Hukumina (Maluku), Bahasa Hoti (Maluku), Bahasa Serua (Maluku), dan Bahasa Nila (Maluku). Suatu bahasa dikatakan punah karena tidak ada lagi penutur dari bahasa tersebut. Kaitannya dengan perkembangan teknologi, jumlah penutur bahasa daerah menurun karena tidak banyak aplikasi teknologi yang menggunakan bahasa daerah (Ulfa, 2019).

Beberapa tahun terakhir perkembangan teknologi *natural language processing (NLP)* juga sangat pesat. NLP merupakan cabang dari bidang studi *artificial intelligence* yang berkaitan dengan interaksi manusia dan komputer melalui bahasa alami baik dalam bentuk tulisan maupun lisan. Pemrosesan bahasa alami merupakan bidang yang tepat untuk pengembangan teknologi berbasis bahasa daerah. Berbagai aplikasi pemrosesan bahasa sudah banyak dikembangkan untuk bahasa *high-resource languages*, yaitu bahasa yang sumber datanya melimpah. Untuk bahasa *low-resource languages,* termasuk bahasa daerah di Indonesia, masih belum banyak pengembangannya karena keterbatasan sumber data.

Berdasarkan kondisi di atas, tugas akhir ini berupaya untuk melestarikan bahasa daerah dengan berpartisipasi dalam pengembangan teknologi pemrosesan bahasa alami berbasis bahasa daerah melalui pembangunan *bilingual dictionary* untuk bahasa daerah. *Bilingual dictionary* merupakan struktur yang menggambarkan hubungan setiap kata dalam dua bahasa yang tingkat perinciannya meliputi bentuk kata, entri leksikal, dan pembacaan (van der Eijk dkk., 1992).

Fokus dari tugas akhir ini adalah pembangunan *bilingual dictionary* untuk Bahasa Jawa dan Bahasa Sunda. Dipilih Bahasa Jawa dan Bahasa Sunda sebagai objek pembangunan *bilingual dictionary* karena kedua bahasa tersebut adalah bahasa yang paling banyak digunakan oleh masyarakat Indonesia *(most common spoken languages)*. Berdasarkan data Badan Pusat Statistik pada tahun 2015, tercatat jumlah penutur Bahasa Jawa sebanyak 84.300.000 jiwa dan untuk Bahasa Sunda sebanyak 42.000.000 jiwa.

Manfaat dari pembangunan *bilingual dictionary* ini adalah untuk menghasilkan *"lexical resource"* sebagai sumber pemrosesan *task* multilingual lainnya yang

memanfaatkan Bahasa Jawa dan Bahasa Sunda. Selain itu *bilingual dictionary* ini juga berguna untuk memudahkan pemahaman Bahasa Jawa seseorang jika orang tersebut sudah memiliki cukup pemahaman terhadap Bahasa Sunda dan sebaliknya.

Pembentukan bilingual dictionary pada tugas akhir ini dilakukan dengan memanfaatkan representasi cross-lingual dari sebuah kata. Representasi *cross-lingual* dari kata atau *cross-lingual embedding* memungkinkan penalaran arti kata pada konteks *multilingual* dan merupakan kunci untuk transfer *cross-lingual* pada pengembangan model pemrosesan bahasa alami untuk *low-resource languages* (Ruder, 2019). Secara umum terdapat 4 jenis pendekatan *cross-lingual embedding*, yaitu pendekatan *monolingual mapping* (Mikolov dkk., 2013; Faruqui dan Dyer, 2014; Xing dkk., 2015; Lazaridou dkk., 2015), pendekatan *pseudo-cross-lingual* (Gouws dan Sogaard, 2015; Duong dkk., 2016; Vulić dan Moens, 2016), pendekatan *cross-lingual training* (Chandar dkk., 2014; Sogaard dkk., 2015), dan pendekatan *joint optimization* (Shi dkk., 2015; Mogadala dan Rettinger, 2016). State-of-the-arts dari cross-lingual embedding tanpa paralel corpora ditunjukkan oleh Conneau dkk. (2018) dengan menggunakan pendekatan *mapping*.

Adapun perkembangan pembangunan *bilingual dictionary* untuk bahasa daerah di Indonesia telah diinisialisasi oleh Nasution dkk. (2016; 2017) melalui penelitiannya dalam pembentukan *bilingual dictionary* untuk Bahasa Melayu dan Bahasa Minangkabau. Pada penelitian tersebut digunakan dua bilingual dictionary, yaitu bilingual dictionary Bahasa Melayu dan Indonesia serta Bahasa Indonesia dan Minangkabau. Dalam penelitiannya digunakan pendekatan *constraint-based bilingual lexicon induction* yang memanfaatkan hubungan kekerabatan Bahasa Melayu dan Bahasa Minangkabau.

Untuk Bahasa Jawa dan Bahasa Sunda yang dikategorikan sebagai *low-resources languages*, sulit ditemukan korpus *parallel* dan *comparable*. Sehingga pada tugas akhir ini akan digunakan korpus *monolingual* dari masing-masing bahasa sebagai data latih model pembangunan *bilingual dictionary*. Selain itu akan digunakan

model *pre-trained word embedding* dari *mBERT* dan *XLM-RoBERTa* dalam pembentukan representasi vektor kata pada kedua bahasa tersebut. Pembentukan dataset uji untuk pengujian model disesuaikan dengan ketersedian data dari Bahasa Jawa dan Bahasa Sunda. Dataset uji digunakan untuk menguji dan menganalisis kinerja model pembangunan bilingual dictionary Bahasa Jawa dan Bahasa Sunda.

I.2 Rumusan Masalah

Berdasarkan uraian latar belakang tugas akhir, ditarik rumusan masalah sebagai berikut.

1. Bagaimana membangun data latih berupa korpus Bahasa Sunda dan Bahasa Jawa?
2. Bagaimana membangun model pembentukan *bilingual dictionary* untuk Bahasa Sunda dan Bahasa Jawa menggunakan *non-comparable* dan *non-parallel corpus*?
3. Bagaimana membangun data uji pembentukan *bilingual dictionary* untuk Bahasa Sunda dan Bahasa Jawa?

I.3 Tujuan

Tujuan yang ingin dicapai dari tugas akhir ini adalah membangun model pembentukan *bilingual dictionary* untuk Bahasa Jawa dan Bahasa Sunda menggunakan *monolingual corpora*.

I.4 Batasan Masalah

Berikut adalah batasan masalah dalam pengerjaan tugas akhir.

1. Pembangunan *bilingual dictionary* tidak memperhatikan hierarki bahasa yang terdapat pada Bahasa Jawa dan Bahasa Sunda.
2. Penelitian ini memanfaatkan model *pre-trained word embedding* yang sudah ada.

I.5 Metodologi

Berikut adalah metodologi yang digunakan dalam pengerjaan tugas akhir ini.

1. Analisis pembangunan *bilingual dictionary* untuk *low-resource language.*

   Pada tahap analisis awal ini dilakukan analisis terhadap model - model pembangunan *bilingual dictionary*, kakas NLP multilingual yang sudah ada, dan dataset berupa korpus Bahasa Jawa dan Bahasa Sunda. Kakas NLP multilingual yang dianalisis adalah *mBERT* dan *XLM Roberta*.

2. Analisis solusi pembangunan *bilingual dictionary* Bahasa Jawa dan Sunda.

   Setelah dilakukan eksplorasi terhadap model pembangunan, kakas, dan dataset pembangunan *bilingual dictionary*, pada tahap ini dilakukan analisis solusi pembangunan *bilingual dictionary* Bahasa Jawa dan Sunda. Hasil dari tahap ini adalah rancangan arsitektur umum model pembangunan *bilingual dictionary* Bahasa Jawa dan Sunda, struktur data latih dan data uji yang akan digunakan, serta metrik evaluasi yang sesuai untuk mengukur kinerja model solusi.

3. Pembangunan data latih dan data uji.

   Pada tahap ini dilakukan pembangunan data latih berupa korpus Bahasa Jawa dan Bahasa Sunda yang didapatkan melalui crawling artikel Wikipedia serta artikel berita Bahasa Jawa dan Bahasa Sunda. Pada tahap ini juga dilakukan pembangunan data uji yang sesuai dengan model solusi yang dihasilkan pada tahap sebelumnya.

4. Implementasi model solusi

   Pada tahap ini dilakukan implementasi model solusi yang didapatkan pada tahap sebelumnya. Implementasi akan dilakukan dengan menggunakan kakas NLP multilingual yang sesuai.

5. Pengujian dan Analisis Solusi

   Pada tahap ini dilakukan eksperimen dengan melakukan pembelajaran model pembangunan *bilingual dictionary* menggunakan data latih berupa

korpus Bahasa Jawa dan Sudsa yang telah disiapkan. Hasil pembelajaran model kemudian diuji dengan menggunakan data uji yang sesuai dengan model solusi. Dari hasil pengujian dilakukan analisis kinerja model dan perbaikan yang perlu dilakukan terhadap model solusi.

I.6 Jadwal Pelaksanaan Tugas Akhir

Jadwal pelaksanaan tugas akhir berikut dapat mengalami perubahan sesuai situasi dan kondisi selama pengerjaan tugas akhir. Jadwal berikut dibuat berdasarkan asumsi akan dilaksanakan wisuda pada bulan Juli tahun 2021.

Tabel I.1 Jadwal Pelaksanaan Tugas Akhir

Tanggal	Milestone	Deliverables
7 Agustus - 3 September 2020	Pemilihan topik dan calon pembimbing beserta preferensi area keilmuan oleh mahasiswa	Pilihan topik yang diminati
7-9 September 2020	Pemilihan prioritas calon bimbingan oleh dosen	Submisi pilihan calon pembimbing
18-19 September 2020	Penentuan alokasi pembimbing oleh tim TA	-
27 September 2020 - 17 Oktober 2020	Eksplorasi literatur penyusunan studi literatur	Referensi yang akan diacu dan laporan BAB II
18 Oktober 2020	Pengumpulan hasil studi literatur (BAB II Buku TA)	Laporan BAB II
25 Oktober 2020	Eksplorasi permasalahan dan penyusunan	Progress laporan BAB I

- 7 November 2020	BAB I	
8-9 November 2020	Pengumpulan rencana topik tugas akhir (BAB I Buku TA)	Laporan BAB I
15 November 2020 - 28 November 2020	Eksplorasi solusi permasalahan dan penyusunan rencana penyelesaian	Progress rancangan penyelesaian masalah
29 November 2020	Pengumpulan rencana penyelesaian masalah (BAB III Buku TA)	Laporan BAB III
7 Desember 2020	Pengumpulan Buku TA I	Buku TA I
14-23 Desember 2020	Pelaksanaan seminar TA I	-
8 Januari 2021 - 18 Februari 2021	Proses analisis penyelesaian masalah	Hasil analisis proses-proses yang dibutuhkan dalam rancangan solusi
19 Februari 2021 - 25 Februari 2021	Target penyelesaian tahap analisis penyelesaian masalah	Hasil analisis kebutuhan untuk rancangan solusi
26 Februari 2021 - 25 Maret 2021	Proses perancangan solusi	Hasil rancangan sistem

26 Maret 2021 - 1 April 2021	Target penyelesaian tahap perancangan solusi	
2 April 2021 - 29 April 2021	Proses implementasi rancangan solusi	Hasil implementasi sistem dan evaluasi
30 April 2021 - 2 Mei 2021	Target penyelesaian pengembangan solusi, penyerahan draft laporan TA yang telah ditandatangani oleh pembimbing melalui TU Akademik	Draft laporan TA
3-7 Mei 2021	Minggu Seminar TA-II	-
14 Mei 2021 - 3 Juni 2021	Revisi laporan TA	Perbaikan laporan TA
4 Juni 2021	Penyelesaian tahap validasi. Penyerahan laporan TA yang telah ditandatangani oleh pembimbing melalui TU Akademik	Hasil laporan TA final
7-18 Juni 2021	Minggu Sidang TA-II	-
19 Juni 2021 - 9 Juli 2021	Revisi laporan TA	Perbaikan laporan TA

16-17 Juli 2021	Wisuda ketiga T/A 2020/2021	-

BAB II

STUDI LITERATUR

Bab Studi Literatur berisi tentang teori dan metode yang berkaitan dengan topik tugas akhir. Studi Literatur mencakup penjelasan mengenai pembangunan *bilingual dictionary* menggunakan bahasa perantara, pembangunan *bilingual dictionary* dengan menggunakan *cross-lingual embedding*, kakas pemrosesan bahasa alami untuk *multilingual*, metrik evaluasi pembangunan *bilingual dictionary*, dan penelitian terkait pembangunan *bilingual dictionary*.

II.1 Pembentukan Bilingual Dictionary

Pembentukan bilingual dictionary akan dilakukan dengan menggunakan cross-lingual embedding. Representasi *cross-lingual* dari kata atau *cross-lingual embedding* memungkinkan penalaran arti kata pada konteks *multilingual* dan merupakan kunci untuk transfer *cross-lingual* ketika mengembangkan model pemrosesan bahasa alami untuk *low-resource languages*. Secara umum terdapat empat jenis pendekatan pembangunan *cross-lingual embedding*, yaitu pendekatan *monolingual mapping*, pendekatan *pseudo-cross-lingual*, pendekatan *cross-lingual training*, dan pendekatan *joint optimization* (Ruder, 2019).

1. Pendekatan Monolingual Mapping.

   Pada model pendekatan ini dilatih *word embedding* pada masing-masing *monolingual corpora* terlebih dahulu dan kemudian dilakukan pembelajaran *mapping* dari bahasa sumber ke bahasa target.

2. Pendekatan Pseudo-cross-lingual.

   Pada model pendekatan *pseudo-cross-lingual* dihasilkan korpus *pseudo-cross-lingual* dengan menggabungkan korpus-korpus dari bahasa yang berbeda berdasarkan konteks tertentu. Kemudian, pada korpus yang dihasilkan, dilatih *off-the-shelf* model *word embedding*.

3. Pendekatan *Cross-lingual Training.*

Pada pendekatan ini dilatih *word embedding* pada korpus paralel dan dilakukan optimasi batasan *cross-lingual* dari dua atau lebih bahasa sehingga vektor kata yang memiliki arti atau makna yang serupa akan berdekatan dalam ruang vektor bersama.

4. *Pendekatan Joint Optimization*.

   Pada pendekatan ini dilatih *word embedding* pada korpus paralel dan dilakukan optimasi batasan-batasan *monolingual* dan *cross-lingual*.

Seperti yang diuraikan pada latar belakang tugas akhir, keterbatasan data Bahasa Jawa dan Bahasa Sunda yang merupakan *low-resource languages* menyebabkan sulit ditemukan korpus paralel dalam jumlah yang besar. Oleh karena itu, pada subbab ini hanya akan diuraikan secara lebih rinci mengenai pendekatan-pendekatan yang sesuai dengan ketersediaan korpus Bahasa Jawa dan Bahasa Sunda, yaitu pendekatan *monolingual mapping* dan pendekatan *pseudo-cross-lingual*.

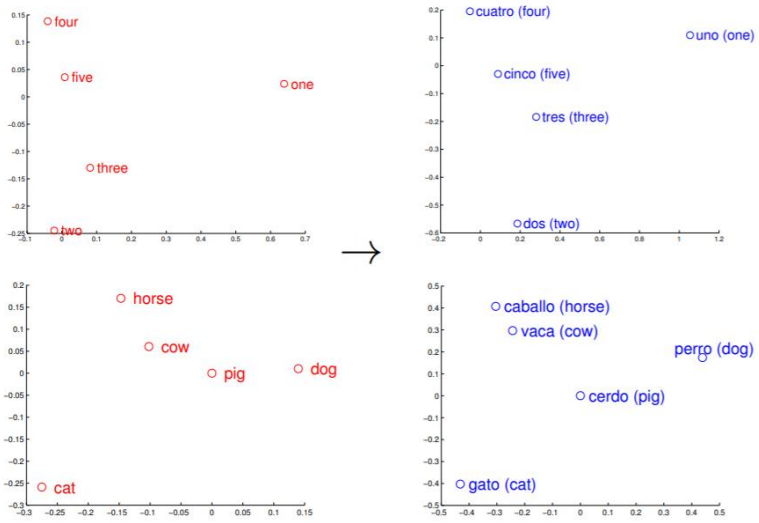
II.2.1 Pendekatan Monolingual Mapping

Pada model pendekatan *monolingual mapping* dilakukan training representasi kata *monolingual* pada masing - masing *monolingual corpora* terlebih dahulu. Kemudian dilakukan training terhadap matriks transformasi yang melakukan mapping dari representasi kata bahasa sumber ke representasi kata bahasa target. Model pendekatan ini juga memanfaatkan pasangan translasi kata dari bahasa sumber dan bahasa target sebagai *anchor* untuk pembelajaran mapping (Ruder, 2019). Berikut adalah metode - metode mapping yang digunakan pada penelitian - penelitian sebelumnya.

1. *Linear Projection*

   Dalam penelitiannya, Mikolov dkk. (2013) menunjukkan bahwa representasi kata dalam ruang vektor dapat menunjukkan relasi berarti antar kata. Selain itu terdapat kemiripan relasi geometrik dari setiap kata pada bahasa yang berbeda, misalnya relasi geometrik antar kata yang

menunjukkan angka dan juga hewan pada Bahasa Inggris mirip dengan relasi geometrik dari kata yang bersesuaian pada Bahasa Spanyol.



Gambar II.1 Contoh Relasi Geometrik Kata (Mikolov dkk., 2013)

Berdasarkan temuan tersebut, Mikolov dkk. (2013) menggunakan metode *linear mapping* untuk melakukan mapping vektor kata dari bahasa sumber ke bahasa target. Metode tersebut menggunakan matriks transformasi *W*. Pada penelitian tersebut dilakukan translasi terhadap 5000 kata dari bahasa sumber yang paling sering muncul di dalam korpus ke kata yang bersesuaian pada bahasa target. Pembelajaran matriks *W* dilakukan dengan menggunakan pendekatan *stochastic gradient descent* dengan meminimumkan jarak antara vektor hasil mapping suatu kata x_i dari bahasa sumber w_i menggunakan matriks *W* dan vektor kata z_i yang merupakan translasi dari kata tersebut. Jarak antara vektor tersebut dihitung dengan menggunakan *Mean Squared Error (MSE)*.

$$min \sum_{i=1}^{n} |Wx_i - z_i|^2 \quad \text{(II.1)}$$

2. *Canonical Correlation Analysis Projection*

Canonical Correlation Analysis (CCA) pertama kali digunakan oleh Haghighi dkk. (2008) untuk melakukan pembelajaran translasi *lexicon* kata. Faruqui dan Dyer (2014) menggunakan *CCA* untuk melakukan

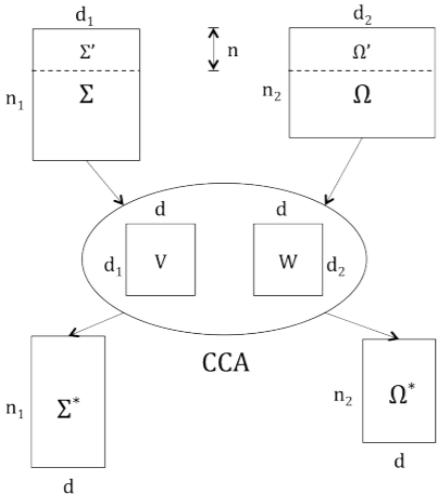
mapping vektor kata pada bahasa sumber dan bahasa target ke sebuah ruang *embedding* bersama *(shared embedding space).* Metode ini melakukan pembelajaran dua matriks transformasi yaitu, matriks transformasi untuk bahasa sumber $W^{s\rightarrow}$ dan matriks transformasi untuk bahasa target $W^{t\rightarrow}$. Sama halnya dengan linear projection, metode ini juga memerlukan pasangan translasi untuk masing-masing pembelajaran matriks transformasi. Korelasi antara hasil proyeksi bahasa sumber $W^{s\rightarrow}x_i^s$ dan hasil proyeksi bahasa target $W^{t\rightarrow}x_i^t$ dirumuskan sebagai berikut:

$$\rho(W^{s\rightarrow}x_i^s, W^{t\rightarrow}x_i^t) = \frac{cov(W^{s\rightarrow}x_i^s, W^{t\rightarrow}x_i^t)}{\sqrt{var(W^{s\rightarrow}x_i^s)var(W^{t\rightarrow}x_i^t)}} \quad \text{(II.2)}$$

Catatan : $cov(.,.)$ adalah kovarian dan $var(.)$ adalah variansi.

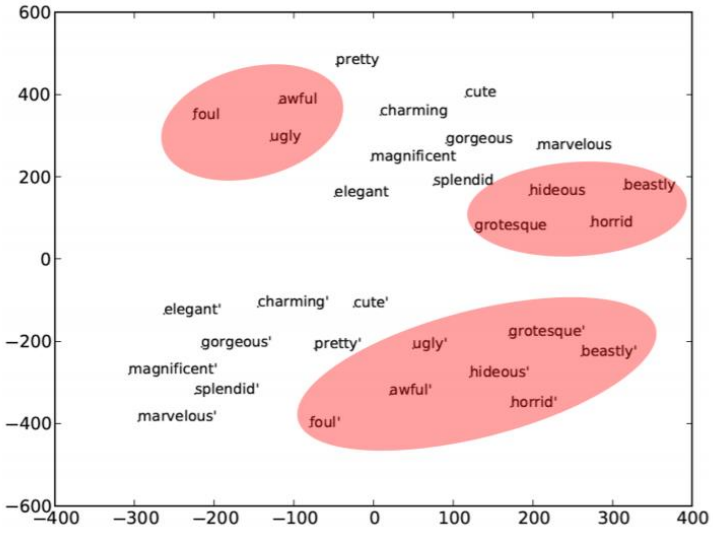
CCA berusaha memaksimalkan nilai korelasi atau yang juga berarti meminimalkan nilai korelasi negatif antara $W^{s\rightarrow}x_i^s$ dan $W^{t\rightarrow}x_i^t$, seperti yang dirumuskan berikut.

$$\Omega_{CCA} = -\sum_{i=1}^{n} \quad \rho(W^{s\rightarrow}x_i^s, W^{t\rightarrow}x_i^t) \text{ (II.3)}$$



Gambar II.2 Pendekatan *CCA* (Faruqui dan Dyer, 2014)

Metode ini dapat membedakan antara sinonim dan antonim dari kata yang diproyeksikan.



Gambar II.3 Sinonim dan Antonim (Merah) dari Kata *'Beautiful'* (Faruqui dan Dyer, 2014)

Lu dkk. (2015) menggunakan *deep neural network* untuk mengoptimasi korelasi dari hasil proyeksi bahasa sumber dan bahasa target. Ammar dkk. (2015) memperluas penggunaan metode ini untuk multilingual CCA dengan Bahasa Inggris sebagai ruang embedding utama (shared embedding space) dari bahasa-bahasa lainnya.

3. *Orthogonal Transformation*

Metode ini merupakan improvisasi dari metode *linear mapping* dengan membuat matriks transformasi W menjadi orthogonal, yaitu $W^\top W = I$. Hal tersebut dapat dipenuhi melalui persamaan $W = VU^\top$ dan dapat diperoleh menggunakan *Singular Value Decomposition (SVD)*, yaitu $X^{t\top} X^s = U\Sigma V^\top$ dengan X adalah matriks *word embedding*, V adalah *vocabulary*. Metode ini digunakan Artetxe dkk. (2016) untuk menangani *monolingual invariance*. Metode ini juga digunakan untuk regularisasi *mapping* oleh Zhang dan Liu (2017).

4. Metode *Max-margin* Menggunakan *Intruder*

Lazaridou (2015) menggunakan *max-margin based ranking loss* untuk mengurangi *hubness* seperti yang ditimbulkan pada pendekatan yang

digunakan Mikolov dkk. (2013). *Hubness* adalah kondisi kecenderungan beberapa kata muncul sebagai *nearest neighbor* dari banyak kata lainnya. *Max-margin based ranking loss* sama dengan *Max-margin loss (MML)* yang digunakan oleh Collobert dan Weston (2008) untuk melakukan pembelajaran model *monolingual word embedding* untuk menghasilkan nilai yang lebih tinggi untuk *correct word sequence* dari *incorrect word sequence*. Pada metode ini, *MLL* digunakan untuk menghasilkan nilai *cosine similarity* yang lebih tinggi dari pasangan translasi kata (x_i^s, x_i^t) dan pasangan kata acak (x_i^s, x_j^t) yang dirumuskan sebagai berikut.

$$\Omega_{MML} = \sum_{i=1}^{n} \sum_{j \neq i}^{k} max\{0, \gamma - cos(W \quad x_i^s, y_i^t) + cos(W \quad x_i^s, y_j^t)\}$$

(II.4)

Dinu dkk. (2015) melakukan optimasi pendekatan ini dengan memilih sebuah *intruder* untuk pasangan acak dari kata yang dilakukan *mapping*. *Intruder* adalah kata y_j^t yang jaraknya dekat dengan $W \quad x_i^s$ tetapi jaraknya jauh dari x_i^t. Hal tersebut berguna untuk mengidentifikasi kasus kegagalan mendekati fungsi target sehingga perlu diperbaiki.

II.2.2 Pendekatan Pseudo-Cross-Lingual

Model pendekatan *pseudo-cross-lingual* bertujuan untuk membangun korpus *pseudo-cross-lingual* sehingga didapatkan relasi antara vektor kata dalam bahasa sumber dan vektor bahasa target melaluinya (Ruder, 2019). Berikut adalah metode - metode yang menggunakan pendekatan *pseudo-cross-lingual*.

1. *Random Translation Replacement*

   Pada metode ini digunakan pasangan translasi kata dari bahasa sumber dengan bahasa target. Dilakukan penggabungan korpus bahasa sumber dengan korpus bahasa target. Kemudian untuk setiap kata yang terdapat

pada pasangan translasi kata diganti dengan kata yang merupakan translasi dari kata tersebut dengan probabilitas 50%.

Metode ini digunakan oleh Gouws dan Søgaard (2015) untuk menghasilkan korpus *pseudo-cross-lingual* secara eksplisit serta digunakan pendekatan *Continuous Bag of Word (CBOW)* untuk mengubah suatu kata menjadi representasi dalam bentuk vektor. Dalam penelitiannya Gouws dan Søgaard (2015) menggunakan pasangan translasi yang diperoleh melalui *Google Translate*. Selain itu, dalam penelitian tersebut juga dilakukan eksperimen dimana penggantian kata pada dua korpus tidak berdasarkan pasangan translasi melainkan menggunakan *part-of-speech equivalent*. Kata yang memiliki label *part-of-speech* yang sama dalam bahasa yang berbeda akan saling digantikan. Pada penelitian tersebut didapatkan bahwa penggantian kata berdasarkan pasangan translasi kata hasilnya lebih baik daripada penggantian kata berdasarkan label *part-of-speech*.

2. *On-the-fly Replacement and Polysemy Handling*

   Daripada melakukan penggantian kata secara acak selama tahap *pre-processing* (Gouws dan Søgaard, 2015), metode ini melakukan penggantian pada kata yang merupakan *center of word* ketika mengubah kata menjadi representasinya dalam bentuk vektor dengan menggunakan pendekatan *CBOW* (Duong dkk., 2016).

   Penanganan terhadap polisemi, yaitu suatu kata yang memiliki makna lebih dari satu dilakukan melalui metode empiris yang memilih kata translasi untuk menggantikan suatu kata. Metode tersebut memilih kata translasi $\underline{w_i}$ yang merupakan kata yang paling mirip *(most similar)* terhadap kombinasi vektor kata dari kata bahasa sumber v_{w_i} dan *context vector* h_i.

$$\underline{w_i} = argmax_{w \in dict(w_i)} cos(v_{w_i} + h_i, v_w) \quad \text{(II.5)}$$

Hasil translasi dengan menggunakan metode ini memiliki nilai *coverage* yang tinggi tetapi seringkali hasilnya *noisy*.

3. *Multilingual Cluster*

   Pada metode *Multilingual Cluster* digunakan *bilingual dictionaries* untuk melakukan clustering terhadap kata dalam bahasa yang berbeda (Ammar dkk., 2016). Pada metode ini dilakukan penggabungan korpus *monolingual* dari bahasa sumber dan bahasa target. Kemudian dilakukan penggantian token kata yang memiliki cluster yang sama dengan menggunakan indeks cluster.

4. *Document Merge and Shuffle*

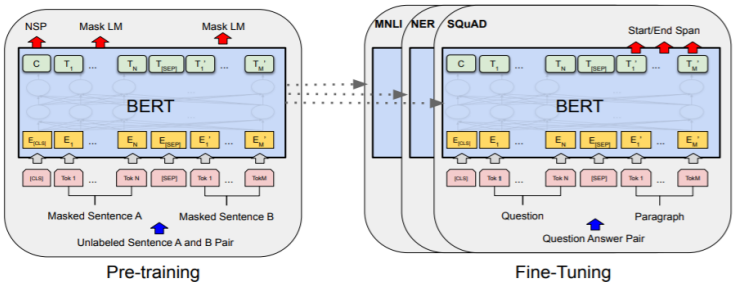
   Pada pendekatan *Document Merge and Shuffle* digunakan *aligned-document*. Dilakukan penggabungan dua *aligned-document* dari bahasa yang berbeda. Kemudian dilakukan *shuffle* dokumen dengan melakukan permutasi kata secara acak. Dengan cara tersebut diharapkan akan dihasilkan ruang *embedding* yang *robust* berdasarkan *bilingual context* untuk kedua bahasa (Vulić dan Moens, 2016). Strategi penggabungan dokumen yang dianggap lebih baik adalah dengan memasukkan kata dari setiap bahasa ke sebuah dokumen *pseudo-bilingual* berdasarkan urutan kata-kata tersebut muncul di dokumen *monolingual* dan panjang dari dokumen *monolingual* (Ruder, 2019).

II.2 Kakas Natural Language Processing untuk Pemrosesan Multilingual

Berdasarkan pencapaian perkembangan teknologi dan hasil penelitian-penelitian yang dilakukan dalam bidang *Natural Language Processing* dan *Natural Language Understanding* terdapat beberapa model representasi bahasa alami yang dapat digunakan dalam pemrosesan *task* multilingual. Model representasi bahasa alami ini dapat digunakan untuk mendukung pembuatan berbagai kakas multilingual lainnya.

II.3.1 BERT

Bidirectional Encoder Representations from Transformer (BERT) adalah *framework* representasi model bahasa alami yang terdiri dari dua langkah terpisah untuk merepresentasikan model bahasa alami, yaitu *pre-training* dan *fine-tuning*. Selama proses *pre-training*, model representasi bahasa dilatih dengan menggunakan data *unlabeled* pada beberapa task *pre-training*. Untuk keperluan berbagai *downstreams* task, dilakukan *fine-tuning* sesuai dengan jenis *downstream task* dengan input berupa *pre-train* model BERT yang parameternya diinisialisasi melalui proses *pre-training* terlebih dahulu. Pada model bahasa BERT, untuk setiap *downstream task* memiliki proses *fine-tuning* yang berbeda tetapi diinisialisasi dengan model *pre-trained* yang sama (Devlin dkk., 2019).



Gambar II.4 Arsitektur *Pre-training* dan *Fine-tuning*

Multilingual BERT adalah model *BERT* yang dilatih pada teks dari berbagai bahasa. Model ini dilatih menggunakan data teks yang bersumber dari konten *Wikipedia* dengan *sharing vocabulary* antar bahasa. Permasalahan umum yang muncul ketika melatih data teks yang berisi teks dari banyak bahasa adalah *imbalance-data* untuk *high-resource language* dan *low-resource language*. Multilingual BERT menangani permasalahan tersebut dengan melakukan *oversampling* pada data *low-resource language* dan *undersampling* pada data *high-resource language*.

Multilingual BERT terdiri dari beberapa jenis berdasarkan cakupan bahasa, yaitu :

1. *BERT-Base, Multilingual Cased*

Model BERT ini terdiri dari 104 bahasa yang tersebar di seluruh penjuru dunia. Arsitektur model BERT ini menggunakan 12 *output layer*, 768 *hidden layer*, 12 *head*, dan 110 juta parameter.

2. *BERT-Base, Multilingual Uncased*

   Model BERT ini terdiri dari 102 bahasa yang tersebar di seluruh penjuru dunia. Arsitektur model BERT ini menggunakan 12 *output layer*, 768 *hidden layer*, 12 *head*, dan 110 juta parameter.

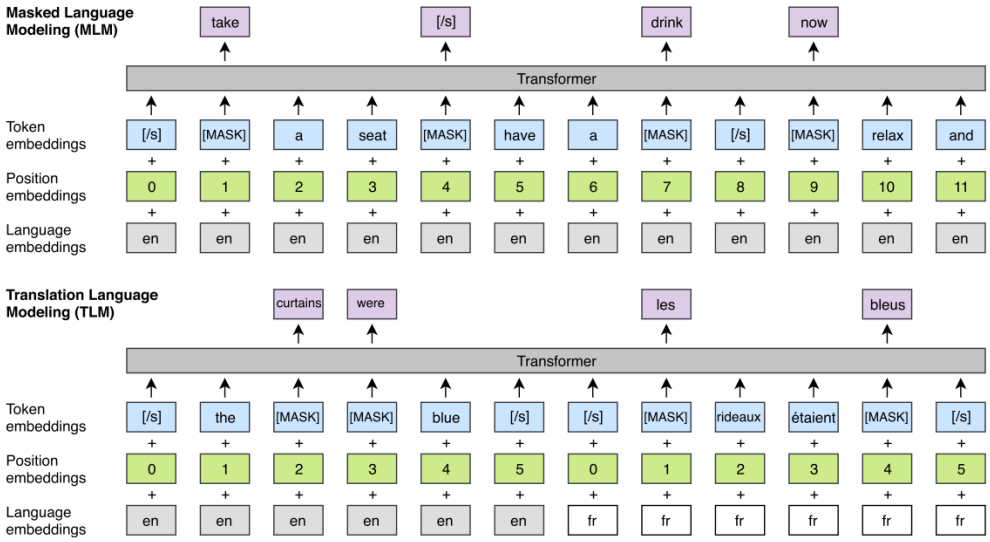
3. *BERT-Base, Chinese*

   Model BERT ini terdiri dari bahasa China yang disederhanakan dan bahasa China tradisional. Arsitektur model BERT ini menggunakan 12 *output layer*, 768 *hidden layer*, 12 *head*, dan 110 juta parameter.

Penggunaan model *Multilingual Cased* lebih direkomendasikan dari pada *Multilingual Uncased* karena sudah dilakukan perbaikan isu normalisasi pada banyak bahasa. model *pre-trained Multilingual BERT* terdiri dari 102 bahasa termasuk Bahasa Indonesia, Bahasa Jawa, dan Bahasa Sunda di dalamnya.

II.3.2 XLM-RoBERTa

XLM-RoBERTa merupakan model multilingual pre-trained yang terdiri dari 100 bahasa termasuk Bahasa Jawa dan Bahasa Sunda dengan data berasal dari *CommonCrawl*. *XLM-RoBERTa* mirip model multilingual pre-trained *XLM* tetapi mengalami improvisasi pada *task Masked Language Modeling* (Conneau, Khandelwal, dkk., 2020). XLM telah mencapai hasil *state-of-the-art* pada klasifikasi *cross-lingual*, *unsupervised* dan *supervised machine translation* (Conneau & Lample, 2019). XLM menggunakan *shared sub-word vocabulary* yang dibentuk melalui *Byte Pair Encoding (BPE)* dalam pembentukan model *pre-train cross-lingual*. Tokenisasi kata pada *XLM-RoBERTa* menggunakan *SentencePiece.*



Gambar II.5 Model *Pre-Training Cross-Lingual*

Berdasarkan ukuran *layer*-nya, XLM-RoBERTa dibagi menjadi dua jenis, yaitu :

1. *XLM-RoBERTa-Base* yang terdiri dari 12 *layer*, 768 *hidden state*, 12 *self-attention head*, dan 270 juta parameter.
2. *XLM-RoBERTa-Large* yang terdiri dari 24 *layer*, 1024 *hidden state*, 16 *self-attention head*, dan 550 juta parameter.

II.3 Metrik Evaluasi

Berdasarkan penelitian yang dilakukan oleh Conneau dkk. (2019), *cross-lingual embedding* dievaluasi melalui beberapa *task*, yaitu :

1. *Word Translation*

   Task word translation digunakan untuk mengukur hasil translasi kata dari kata sumber yang diberikan dengan metric evaluasi berupa *precision*.

   $$precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad \text{(II.6)}$$

   Hasil translasi yang benar sesuai dengan *ground truth* dihitung sebagai *True Positive*, sedangkan translasi yang salah dihitung sebagai *False Positive*. Pada task ini dilaporkan nilai $precision@k$ dengan nilai $k = 1, 5, 10$. Nilai *k* berarti jumlah berapa kali sebuah translasi yang benar terjadi.
2. *Cross-lingual Semantic Word Similarity*

Task cross-lingual semantic word similarity digunakan untuk mengukur kualitas ruang *cross-lingual word embedding*. Tujuan dari task ini adalah menunjukkan korelasi *cosine similarity* antara dua kata berbeda bahasa berkorelasi dengan *human-labeled score*. Untuk menunjukkan korelasi tersebut digunakan *Pearson correlation*.

$$r = \frac{\sum_{i=1}^{n} \cos(X_i,\bar{X})\cos(Y_i,\bar{Y})}{\sqrt{\cos(X_i,\bar{X})^2}\sqrt{\cos(Y_i,\bar{Y})^2}} \quad \text{(II.7)}$$

X_i adalah hasil translasi dengan menggunakan model *cross-lingual* word embedding dan $\bar{X}$ adalah kata dari bahasa sumber. Y_i adalah kata translasi berdasarkan data *human-labeled* dan $\bar{Y}$ adalah kata dari bahasa sumber pada data *human-labeled.*

3. *Sentence Translation Retrieval*

   Task sentence translation retrieval serupa dengan *task word translation*, yaitu menggunakan metrik *presicion* pada level kalimat. Pada task ini dilaporkan nilai $precision@k$ dengan nilai $k = 1, 5, 10$. Nilai k berarti *fraction of pairs* dari translasi yang benar dari kata sumber pada jangkauan *k-nearest neighbors.*

Berdasarkan penelitian yang dilakukan oleh Nasutio dkk. (2017), pembentukan bilingual dictionary dievaluasi dengan menggunakan metric *precision*, *recall*, dan *F-measure*. Persamaan untuk precisiondapat dilihat melalui persamaan 2.1. Berikut adalah persamaan untuk recall dan *F-measure.*

$$recall = \frac{True\ Positive}{True\ Positive + False\ Negatice} \quad \text{(II.8)}$$

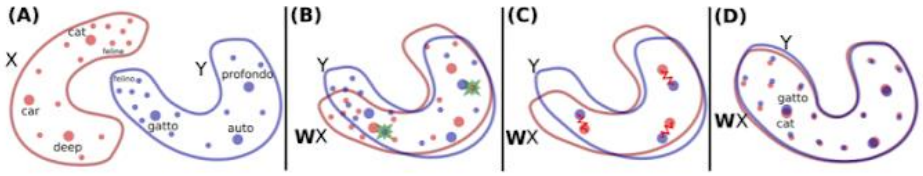
$$F - measure = \frac{precision \times recall}{precision + recall} \quad \text{(II.9)}$$

Hasil translasi yang benar sesuai dengan *ground truth* dihitung sebagai *True Positive*, sedangkan translasi yang salah dihitung sebagai *False Positive*.

II.4 Penelitian Terkait Pembangunan Bilingual Dictionary

II.4.1. MUSE

MUSE atau *Multilingual Unsupervised and Supervised Embedding* merupakan *library* bahasa pemrograman *python* dari *Facebook* untuk *multilingual embedding*. MUSE berusaha menghasilkan *multilingual embedding* hanya dengan menggunakan dua *monolingual corpora*. MUSE menggunakan metode *adversarial training* untuk melakukan pembelajaran proses *mapping* dari *source* ke *target space* (Conneau dkk., 2018).



Gambar II.6 Ilustrasi Metode *Mapping MUSE*

Berikut adalah langkah-langkah mapping dari *source space* ke *target space* yang digunakan MUSE sesuai dengan gambar ilustrasi diatas.

A. Dua distribusi *word embedding*, Bahasa Inggris yang ditandai dengan warna merah dan huruf X, Bahasa Italia yang ditandai dengan warna biru dan huruf Y.
B. *Adversarial learning*, melakukan training matriks rotasi W untuk *alignment* kedua distribusi bahasa. Dua titik hijau adalah kata random yang di *fed* ke *discriminator* untuk menentukan apakah dua kata tersebut berasal dari distribusi yang sama.
C. *Refine* mapping menggunakan matriks W dengan *Procrustest*.
D. Translasi dengan *mapping* W dan *distance matrix*.

Metode mapping yang digunakan oleh MUSE adlaah metode *linear mapping* yang diuraikan pada subbab II.2.1. MUSE juga melakukan refinement pada metode *mapping*-nya dengan membuat *synthetic parallel dictionary* dan model pembelajaran *mapping* hanya mempertimbangkan *frequent words* serta hanya mempertahankan *nearest neighbors*. Selain itu diimplementasikan algoritma

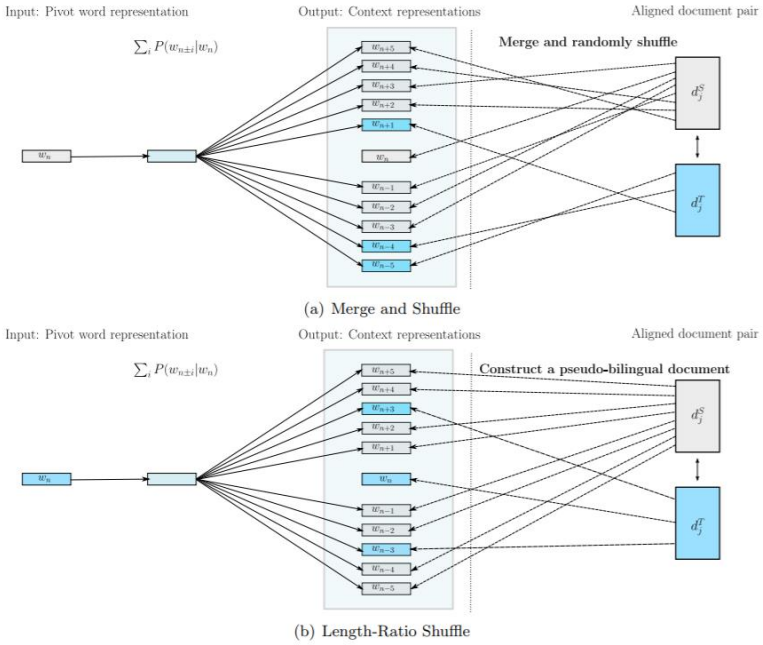
Procrustest. Model *mapping* ini menggunakan *bi-partite neighborhood graph* dimana untuk setiap kata di *dictionary* akan terhubung ke K *nearest neighbor* dari *target language*-nya. Untuk menangani permasalahan hubness, MUSE menggunakan *pairing rule* berdasarkan *reverse rank* dan *inverted soft-max (ISF).*

Tabel II.1 Evaluasi *MUSE English - Italian*

	English to Italian			Italian to English		
	P@1	P@5	P@10	P@1	P@5	P@10
With Cross-lingual Supervision						
Procrustes - CSLS	63.7	78.6	81.1	56.3	76.2	80.6
Without Cross-lingual Supervision						
Adv _ Refine - CSLS	66.2	80.4	83.4	58.7	76.5	80.9

II.4.2. Bilingual Distributed Word Representations from Document-Aligned Comparable Data

Penelitian ini memanfaatkan korpus bahasa berupa data document-aligned. Pendekatan yang digunakan pada penelitian ini adalah pendekatan *pseudo-cross-lingual* dengan metode *merge and shuffle* dan *length-ratio shuffle* (Vulić dan Moens, 2016). Dalam penelitian ini proses pembentukan cross-lingual word embedding melalui tiga tahapan, yaitu tahap merging, shuffling, dan pembelajaran skirepresentasi vektor kata. Pertama-tama kedua korpus bahasa akan digabungkan menjadi satu membentuk korpus *pseudo-bilingual*. Kemudian dari korpus *pseudo-bilingual* tersebut dilakukan *shuffle* menggunakan metode *merge and shuffle* atau *length-ratio shuffle*. Uraian mengenai metode tersebut dapat dilihat pada subbab II.2.2 poin 4. Pada hasil korpus pseudo-bilingual yang sudah silakukan shuffle kata, dilakukan pembelajaran *skip-gram negatives sampling* untuk menghasilkan representasi vektor dari setiap kata.



Gambar II.7 Arsitektur *Merge and Shuffle (a)* dan *Length-Ratio Shuffle (b)*

Metrik evaluasi yang digunakan pada penelitian ini adalah metrik akurasi terhadap task pembentukan *bilingual lexicon one-to-one translation pairs*. Berikut adalah hasil evaluasi untuk pasangan Bahasa Spanyol dan Bahasa Inggris.

Tabel II.2 Evaluasi *Merge and Shuffle* dan *Length-Ratio Shuffle*

	d = 100	***d = 200***	***d = 300***
Merge and Shuffle			
cs:16, MIN	*0.607*	*0.600*	*0.577*
cs:16, MAX	*0.617*	*0.613*	*0.596*
cs:16, AVG	*0.625*	*0.630*	*0.613*
cs:48, MIN	*0.658*	*0.676*	*0.672*
cs:48, MAX	*0.665*	*0.685*	*0.688*
cs:48, AVG	*0.675*	*0.694*	***0.705***
Length-Ratio			
cs:16	*0.627*	*0.610*	*0.602*
cs:48	***0.678***	***0.701***	*0.703*

BAB III

ANALISIS DAN RANCANGAN SOLUSI

Bab Analisis dan Rancangan Solusi berisi uraian analisis permasalahan tugas akhir dan rancangan solusi berdasarkan teori dan metode yang telah dipelajari melalui penelitian – penelitian yang dilakukan sebelumnya. Bab ini dibagi menjadi analisis permasalahan dan rancangan solusi model pembangunan *bilingual dictionary* untuk Bahasa Jawa dan Bahasa Sunda.

III.1 Analisis Permasalahan

Analisis permasalahan model pembangunan *bilingual dictionary* untuk tugas akhir ini meliputi analisis terhadap dataset yang tersedia dan model pembelajaran pembangunan *bilingual dictionary* untuk Bahasa Jawa dan Sunda.

III.1.1 Analisis Dataset

Pembangunan *bilingual dictionary* sangat bergantung pada ketersediaan *corpus* dari bahasa sumber dan bahasa target. Berdasarkan eksplorasi sumber data yang dilakukan, untuk Bahasa Jawa dan Sunda tidak tersedia korpus *parallel* maupun *comparable*. Pada tugas akhir ini akan digunakan korpus *monolingual* yang berisi kumpulan kalimat yang diperoleh dari artikel – artikel Wikipedia dan artikel berita yang berbahasa Jawa dan Sunda yang diperoleh melalui proses *crawling*. Untuk korpus Bahasa Jawa, artikel berita diperoleh melalui *crawling* terhadap website berita, seperti *Djakalodang.com* dan *solopos.com*. Untuk korpus Bahasa Sunda, artikel berita diperoleh melalui crawling terhadap website berita, seperti *koransunda.com* dan *bewarajabar.com*. Berikut adalah contoh kalimat dalam korpus Bahasa Jawa dan Bahasa Sunda yang belum dilakukan *pre-processing*.

```
Jawa = [“Usul Tentara Nasional Indonésia nepaki dina ambal warsa
TNI.Pras pirembagan 5 Oktober 2015 14.42 (UTC)”, “Ananging tibake
wis tau dadi AP ing taun 2010 ora apa-apa mas?”, “Apa ana usulan
liya?”, “Pras pirembagan 5 Oktober 2015 14.49 (UTC)Sarujuk Gapura
Bradenburg mas.”, “Panganggo ora olèh ngusulaké ping pidho
```

sawijining artikel nganti usulan kapisan artikel kasebut wis nampa dhukungan utawa kritikan wis didandani”]

Sunda = [“BudayaAkulturasi nyaéta hiji prosés nu aya patalina jeung masarakat nu mucunghul lamun hiji bubuhan manusa nu mibanda kabudayaan tinangtu disanghareupkeun kana unsur tina kabudayaan séjén.”, “Kabudayaan séjén éta laun-laun bakal ditarima jeung diolah kana jeroeun kabudayaanana sorangan kalayan henteu nyababkeun leungitna unsur kabudayaan bubuhan éta sorangan.”, “Conto akulturasi: budaya rap ti nagara deungeun dihijikeun jeung basa Jawa, mucunghul rap maké basa Jawa.”]

Dalam proses pembelajaran model pembentukan *bilingual dictionary* kumpulan kalimat tersebut akan diubah menjadi representasi vektor dengan memanfaatkan model *pre-trained word embedding mBERT* untuk Bahasa Jawa dan Sunda.

Data uji yang akan digunakan pada tugas akhir ini berupa kumpulan pasangan translasi dari kedua bahasa. Pasangan translasi diperoleh melalui translasi dengan menggunakan *Google Translate* dan *Wiktionary Wikipedia*.

translation_pairs = [(x_1,y_1), (x_2,y_2), …, (x_n,y_n)]

Dengan x_i adalah kata dalam Bahasa Jawa dan y_i adalah kata dalam Sunda yang merupakan translasi dari x_i.

III.1.2 Analisis Model Pembelajaran

Berdasarkan ketersediaan data untuk Bahasa Jawa dan Sunda dan studi literatur terhadap model pembentukan *bilingual dictionary* pada subbab II.1, berikut adalah pendekatan - pendekatan pembentukan model pembangunan *bilingual dictionary* yang akan digunakan pada tugas akhir ini.

III.1.2.1. Pendekatan *Monolingual-Mapping*

Pendekatan *mapping-based* bertujuan untuk mempelajari pemetaan dari ruang *embedding monolingual* ke ruang *cross-lingual embedding* bersama. Pada tugas akhir ini akan digunakan pendekatan *mapping-based* dengan metode linear projection seperti yang diadaptasi dari penelitian Mikolov dkk. (2013) dengan memperhatikan ortogonalitas matriks transformasi mapping seperti yang

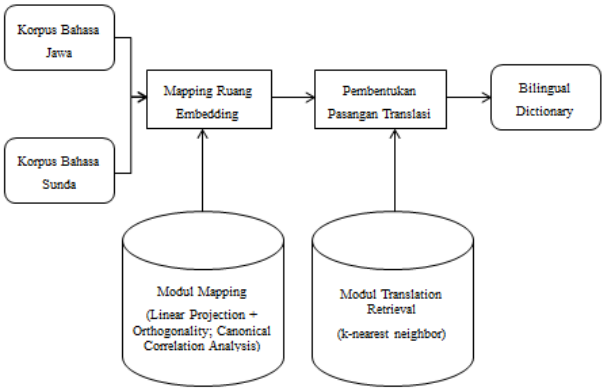
dilakukan pada penelitian Ziang dan Liu (2017). Selain itu juga akan dilakukan eksperimen menggunakan metode *Canonincal Correlation Analysis* yang diadaptasi dari penelitian Faruqui dan Dyer (2014). Masing – masing metode tersebut diuraikan pada subbab II.1 poin 1, 2, dan 3.

III.1.2.2. Pendekatan *Pseudo-Cross-Lingual*

Pendekatan ini memerlukan data berupa *document-topic aligned* dari kedua bahasa yang diperoleh melalui *crawling* sebagaimana dijelaskan pada subbab III.1.1. Pada pembentukan *cross-lingual embedding* untuk model pembanguan bilingual dictionary pada tugas akhir ini akan digunakan penggabungan dua korpus Bahasa Jawa dan Bahasa Sunda dengan menggunakan teknik *Merge and Shuffle* (Vulić & Moens, 2016) sebagaiman yang diuraikan pada subbab II.2 poin 4.

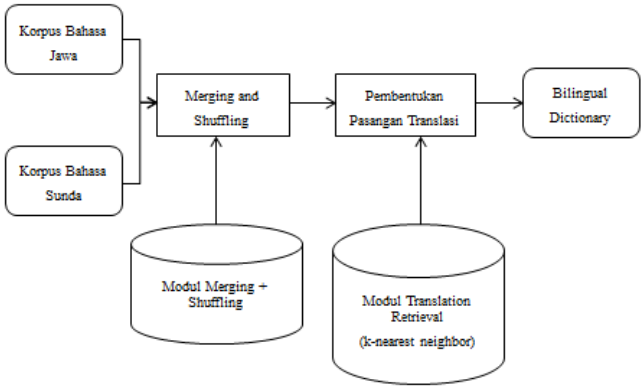
III.2 Rancangan Solusi

Berdasarkan studi literatur pada bab II, secara umum model pembangunan *bilingual dictionary* terdiri dari dua tahapan, yaitu pemetaan atau *alignment* antar bahasa dan pemilihan kandidat translasi. Arsitektur model pembangunan *bilingual dictionary* Bahasa Jawa dan Bahasa Sunda untuk pendekatan *monolingual-mapping* diadaptasi dari arsitektur *MUSE* (*Multilingual Unsupervised and Supervised Embedding)* yang diuraikan dalam penelitian Conneau dkk. (2017). Hanya saja dalam tugas akhir ini dilakukan eksperimen terhadap dua metode *mapping* seperti yang diuraikan pada subbab III.1.2.1. Berikut adalah arsitektur umum dari model pembangunan *bilingual dictionary* dengan menggunakan pendekatan *monolingual-mapping*.



Gambar III.1 Arsitektur Umum Pendekatan *Monolingual-Mapping*

Berikut adalah arsitektur umum model pembangunan *bilingual dictionary* dengan menggunakan pendekatan *pseudo-cross-lingual*.



Gambar III.2 Arsitektur Umum Pendekatan *Pseudo-Cross-Lingual*

Hasil eksperimen dari masing – masing pendekatan akan dievaluasi melalui *task word translation* dengan menggunakan metrik *precision* yang diuraikan pada subbab II.3 poin 1 dan *task cross-lingual semantic similarity* dengan menggunakan metrik *pearson correlation* yang diuraikan pada subbab II.3 poin 2.

DAFTAR PUSTAKA

Mikolov, T., Le, Q. V., & Sutskever, I. (2013). *Exploiting Similarities among Languages for Machine Translation*. Retrieved from http://arxiv.org/abs/1309.4168

Faruqui, M., & Dyer, C. (2014). Improving Vector Space Word Representations Using Multilingual Correlation. Proceedings of *the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 462 – 471. Retrieved from http://repository.cmu.edu/lti/31

Xing, C., Liu, C., Wang, D., & Lin, Y. (2015). Normalized Word Embedding and Orthogonal Transform for Bilingual Word Translation. *NAACL-2015*, 1005–1010.

Lazaridou, A., Dinu, G., & Baroni, M. (2015). Hubness and Pollution: Delving into Cross-Space Mapping for Zero-Shot Learning. Proceedings of *the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 270–280.

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing. Proceedings of *the 25th International Conference on Machine Learning - ICML '08, 20(1)*, 160–167.

Guo, J., Che, W., Yarowsky, D., Wang, H., & Liu, T. (2015). Cross-lingual Dependency Parsing Based on Distributed Representations. Proceedings of *the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1234–1244. Retrieved from http://www.aclweb.org/anthology/P15-1119

Ammar, W., Mulcaire, G., Tsvetkov, Y., Lample, G., Dyer, C., & Smith, N. A. (2016). *Massively Multilingual Word Embeddings*. Retrieved from http://arxiv.org/abs/1602.01925

Artetxe, M., Labaka, G., & Agirre, E. (2016). Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. Proceedings of *the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP-16)*, 2289–2294.

Xiao, M., & Guo, Y. (2014). Distributed Word Representation Learning for Cross-Lingual Dependency Parsing. *CoNLL*.

Gouws, S., & Søgaard, A. (2015). Simple task-specific bilingual word embeddings. *NAACL*, 1302–1306.

Duong, L., Kanayama, H., Ma, T., Bird, S., & Cohn, T. (2016). Learning Crosslingual Word Embeddings without Bilingual Corpora. Proceedings of *the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP-16)*.

Vulić, I., & Moens, M.-F. (2016). *Bilingual Distributed Word Representations from Document-Aligned Comparable Data*. Journal of Artificial Intelligence Research, 55, 953–994. Retrieved from http://arxiv.org/abs/1509.07308

Ulfa, Mariam. (2019). *Eksistensi Bahasa Daerah di Era Disrupsi*. Stilistika: Jurnal Pendidikan Bahasa dan Sastra, 197-207.

Nasution, A. H., Murakami, Y., & Ishida, T. (2017). A Generalized Constraint Approach to Bilingual Dictionary Induction for Low-Resource Language Families. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *17*(2). https://doi.org/10.1145/3138815

Conneau, A., Lample, G., Ranzato, A., Denoyer, L., & Jégou, H. (n.d.). *WORD TRANSLATION WITHOUT PARALLEL DATA*. Retrieved October 18, 2020, from https://github.com/facebookresearch/MUSE

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*(Mlm), 4171–4186.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (n.d.). *Unsupervised Cross-lingual Representation Learning at Scale*. Retrieved October 18, 2020, from https://github.com/facebookresearch/cc