OPEN ACCESS
Indo-JC
Indonesia Journal of Computing
ISSN 2460-9056
socj.telkomuniversity.ac.id/indojc

Ind. Journal on Computing
Vol. 4, Issue. 2, Sept 2019. pp. 127-136
doi:10.21108/indojc.2019.4.2.331

Pembangunan Pensejajaran Kata *Monolingual* (*Monolingual Word Alignment*) pada Terjemahan Al-Quran Bahasa Indonesia

Kurnia Sari Sopi Lingga [#1], Moch Arif Bijaksana [*2], Arie Ardiyanti [#3]

[#] *School of Computing, Telkom University Bandung Indonesia*
[1] kurniasarisopilingga@student.telkomuniversity.ac.id
[2] arifbijaksana@.telkomuniversity.ac.id
[3] ardiyanti@telkomuniversity.ac.id

Abstract

This paper discusses the development of monolingual word alignment in the translation of the Indonesian Qur'an. This topic was taken because alignment is the main component of some natural language processing, which is the contextual recognition, textual integration, identification, paraphrase detection, question answering and text summarization. Besides this the translation of the Qur'an is very much in its version so that it requires interpretation to interpret different translations but has the same meaning. With this technique a number of different Qur'anic translation words can be aligned, so that the words will be grouped according to their semantic similarities. This technique can also be used further for the construction of synonyms and WordNet. Input from the system is a pair of translations for the same verse. The evaluation in this study resulted in a correlation score of 0.82 with a value of error tolerance in the system of 0.18. Correlation between verses that have semantic similarities insofar as this research can be said to be adequate but if you want to improve again, you need a more complete feature and knowledge base.

Keywords: *alignment*, holy quran, *monolingua*l, semantik, translation

Abstrak

Paper ini membahas tentang pembangunan pensejajaran kata *monolingual* pada terjemahan Al-Quran Bahasa Indonesia. Topik ini diambil karena *alignment* merupakan komponen utama dari beberapa pemrosesan bahasa alami yaitu *textual entailment recognition, textual similarity identification, paraphrase detection, question answering* dan *text summarization*. Selain itu terjemahan Al-Quran ini sangat banyak versinya sehingga membutuhkan penafsiran untuk mengartikan terjemahan yang berbeda namun memiliki makna yang sama. Dengan adanya teknik ini beberapa kata terjemahan Al-Quran yang berbeda dapat disejajarkan, sehingga kata-kata tersebut akan terkelompokkan berdasarkan kemiripan semantiknya. Teknik ini juga dapat digunakan lebih lanjut untuk pembangunan *synonim set* dan WordNet. *Input*-an dari sistem berupa pasangan terjemahan untuk sebuah ayat yang sama. Evaluasi pada penelitian ini menghasilkan skor korelasi 0.82 dengan nilai toleransi kesalahan pada sistem sebesar 0.18. Korelasi antar ayat yang memiliki kemiripan semantik sejauh penelitian ini sudah dapat dikatakan memadai namun jika ingin ditingkatkan kembali maka diperlukannya fitur dan basis pengetahuan yang lebih lengkap lagi.

Kata Kunci: *alignment*, alquran, *monolingual*, semantik, terjemahan

Received on August 2019. Accepted on Sept 2019

I. PENDAHULUAN

AL-QURAN diturunkan Allah kepada ummat manusia dijadikan sebagai pedoman dan merupakan kitab suci ummat Islam [1]. Al-Quran diterjemahkan dalam sejumlah bahasa. Dalam terjemahan itu sendiri terdiri dari beberapa versi yaitu terjemahan versi Universitas Islam Indonesia (UII), Departemen Agama, Al-Azhar dan Quraish Shihab merupakan versi terjemahan yang populer pada saat ini. Satu ayat Al-Quran yang diterjemahkan ke dalam beberapa versi terjemahan memiliki satu makna semantik. Oleh karena itu, kita dapat menyusun kata-kata yang memiliki keterkaitan semantik dari beberapa versi terjemahan. Proses ini dapat dilakukan dengan menerapkan *word alignment*. *Word alignment* bertujuan untuk menemukan kombinasi terbaik dari pasangan unit semantik yang serupa dalam konteks yang sama [2] [3]. Artinya, bahwa *word wlignment* akan mengukur kemiripan kata atau *similarity* [4] [5].

Salah satu metode *word alignment* adalah *monolingual word alignment*. *Monolingual word alignment* merupakan metode *alignment* yang diimplementasikan untuk mencari kesamaan dan kemiripan semantik antar teks dalam satu bahasa atau dalam bahasa yang berbeda dengan mensejajarkan terjemahan yang dijadikan studi kasus [6]. *Monolingual word alignment* terdiri dari beberapa properti konsep. Ini memiliki definisi identik antar kata, set sinonim, dan hubungan dengan konsep lain seperti hiponim dan hypernim. *Monolingual word alignment* sederhana dan mudah ditiru tetapi masih menunjukkan kinerja yang baik dalam melakukan kinerja pensejajaran kata demi kata dalam kalimat [6]. Hal itu adalah salah satu alasan mengapa *monolingual word alignment* mendapat banyak perhatian di antara para peneliti. Dalam menerapkan algoritma ini, kami menggunakan cara yang dilakukan oleh Md Arafat sultan, Steven Bethard and Tamara Sumner karena terbukti berfungsi baik dan menghasilkan akurasi yang tinggi [6].

Metode *monolingual word alignment* membutuhkan sebuah *dataset monolingual* yang berfungsi sebagai kumpulan objek yang memiliki keterkaitan semantik [6]. Saat ini terdapat korpus *monolingual word alignment* dari penelitian MSR RTE 2007 dalam Bahasa Inggris [7]. Namun korpus *monolingual word alignment* versi bahasa Indonesia khususnya untuk terjemahan Al-Quran masih sangat terbatas. Dalam *paper* ini kami menerapkan algoritma *word alignment* pada teks terjemahan Al-Quran dari beberapa versi yaitu terjemahan versi Departemen Agama dan UII.

Keluaran yang dihasilkan berupa pasangan-pasangan kata yang memiliki kemiripan semantik. Evaluasi dari sistem dilakukan dengan menghitung jumlah kata yang dipadankan dengan tepat dibandingkan dengan jumlah kata dalam kalimat. Diharapkan penelitian ini dapat meniru intuisi Muslim dalam mengukur kesamaan konsep Al-Quran pada tingkat semantik. Kemudian di masa depan, penelitian ini dapat dibawa ke masalah yang lebih menarik terkait pemrosesan data yang menghasikan akurasi kemiripan semantik yang lebih baik dengan data set yang lebih banyak dan pembangunan *synset* dalam WordNet.

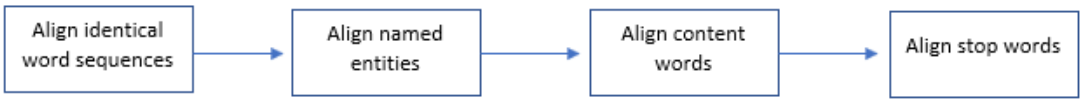
II. KAJIAN LITERATUR

Penelitian yang dilakukan oleh Md Arafat sultan, Steven Bethard and Tamara Sumner menghadirkan *Monolingual Alignment* sederhana dan mudah ditiru yang menunjukkan kinerja canggih dengan mengandalkan hampir tidak ada pengawasan dan sejumlah kecil sumber daya eksternal. Kumpulan data yang digunakan berformat korpus MSR yang selaras secara manual dengan pelatihan set test dari Microsoft Research [8] dalam hal ini juga penulis menggunakan *dataset* berformat MSR. Sehingga dalam pembacaan *dataset* memiliki kualitas dan kecepatan yang lebih baik. Berdasarkan hipotesis menunjukkan bahwa kata-kata dengan makna yang sama mewakili pasangan potensial untuk pensejajaran jika terletak dalam konteks yang sama, sehingga dalam penelitian tersebut diusulkan penggunaan *monolingual alignment* [6]. Dua Asumsi yang sering dibuat dalam penelitian adalah :

1) Unit semantik terkait dalam dalam dua kalimat memiliki arti yang identik atau makna dari kata yang mirip
2) Kesamaan dalam konteks semantik pada masing masing kalimat memiliki keterkaitan makna

Artinya bahwa *alignment* hanya berdasarkan pada dua asumsi tersebut untuk menemukan pasangan-pasangan kata terbaik yang memiliki kemiripan semantik dalam konteks yang sama [2] [3] [9] [10]. Penelitian yang pernah juga dilakukan dengan menggantikan *decoding* berbasis simulasi *annealing* yang disebut MANLI dengan pemrograman linier bilangan bulat, dan mencapai kecepatan yang tinggi. Dalam penelitian

ini menemukan bukti kontekstual dalam bentuk kendala sintaksis yang berguna dalam menyelaraskan kata-kata dan memperluas model dengan menambahkan fitur yang menandai nilai ketergantungan, secara efektif membawa pengaruh yang lebih kuat dari kesamaan kontekstual ke dalam keputusan penyelarasan [2]. Selain itu pada penelitian lain salah satu dari beberapa *aligner*, *aligner* yang paling baik hingga saat ini baik dalam hal akurasi dan kecepatan, yang disebut JacanaAlign, dikembangkan oleh Yao et al [9]. Dalam sistem JacanaAlign juga menggunakan metode *word alignment* namun memiliki sedikit perbedaan, JacanaAlign adalah *word alignment* yang merumuskan perataan sebagai masalah pelabelan urutan. Setiap kata dalam kalimat sumber dilabeli dengan indeks kata target yang sesuai jika ditemukan keselarasan. Meskipun JacanaAlign mengungguli peningkatan MANLI namun memiliki fitur yang kurang kontekstual, sulit untuk membandingkan peran konteks dalam dua model karena perbedaan paradigmatik yang besar.



Gambar 1. *Overview* Sistem [6]

Sehingga, dapat disimpulkan kemiripan kata semantik telah menjadi komponen utama dari sebagian besar *aligners* dengan berbagai ukuran kemiripan kata telah digunakan, termasuk kemiripan *string*, kemiripan berbasis sumber daya (berasal dari satu atau lebih sumber daya *leksikal* seperti WordNet) dan kemiripan distribusi (dihitung dari statistik kemunculan kata dalam korporasi besar) dengan menggunakan Parafrase *Database* (PPDB) [3], yang merupakan sumber besar parafrase leksikal dan frasa yang dibangun menggunakan pivoting *dwi*bahasa di atas korporasi paralel besar [3] [6]. Sistem yang dibangun beroperasi sebagai modul *alignment* yang berbeda dalam jenis pasangan kata yang disejajarkan. Gambar 1 adalah representasi secara umum dengan bentuk sederhana dari *overview* sistem *monolingual word alignment* [6]. Setiap modul menggunakan bukti kontekstual untuk membuat keputusan penyelarasan.

Pada penelitian ini sistem yang akan dibangun mengadaptasi pemanfaatan Parafrase *Database* (PPDB) dengan menggunakan metode *monolingual word alignment* dari beberapa paper yang telah dijelaskan. Dalam pengembangannya berdasarkan *paper* yang menjadi acuan tersebut, sistem yang dibangun menghasilkan pasangan-pasangan kata yang memiliki kemiripan semantik. Adanya pendekatan algoritma kemiripan semantik kata. Karena sifat phrasa yang dimiliki, dengan memperlakukan entitas yang disebutkan secara terpisah dari kata-kata konten lainnya. Dengan menggunakan data set terjemahan Al-Quran versi bahasa Indonesia karena ketersediaannya masih terbatas.

III. METODE PENELITIAN

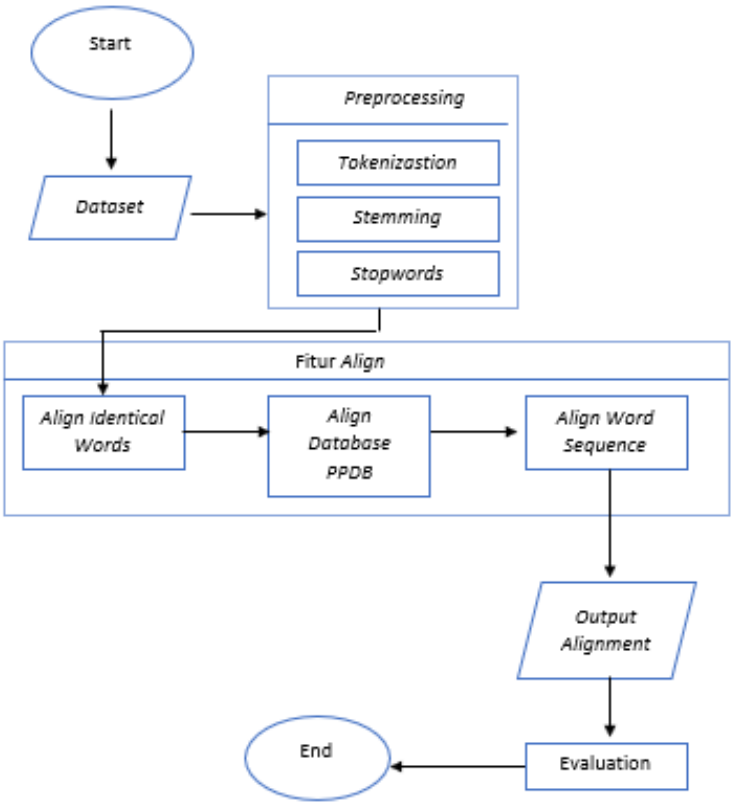
Adapun *input*-an dari sistem dalam melakukan analisis pensejajaran terjemahan ayat-ayat Al-Quran versi Bahasa Indonesia berupa pasangan atau himpunan yang berbeda untuk sebuah ayat yang sama. Sehingga menghasilkan pasangan-pasangan kata yang memiliki kemiripan semantik.

A. Gambaran Umum Sistem

Pada sistem ini ada beberapa proses yang akan dijalankan yaitu tahap *preprocessing*, pembacaan *dataset* format MSR, ekstraksi fitur dan *implementasi* fitur.

Berikut merupakan penjelasan dari Gambar 2 terkait langkah-langkah dari sistem yang akan dibangun dalam penelitian ini :

1) Sistem akan membaca *dataset* format MSR yang sebelumnya telah dilakukan *preprocessing*. Pada tahap *preprocessing* ini *dataset* akan menjalani proses modifikasi diantaranya *tokenisasi, stopwords* dan *stemming*.



Gambar 2. Rancangan Sistem

2) Pada *dataset* yang digunakan terdiri dari dua kalimat yaitu kalimat pertama merupakan kalimat dijadikan acuan atau pembanding dengan kalimat yang kedua untuk menentukan kemiripan semantik pada terjemahan yang di-*input*-kan. Namun dalam hal ini Kalimat kedua pada *dataset* berformat MSR. Kemudian pada kalimat kedua akan dilakukan proses pembersihan dan pemisahan data yang nantinya hanya akan diambil katanya saja untuk proses yang akan dilakukan di dalam sistem. Dalam hal ini memerlukan sebuah *regular expression* untuk melakukan pemisahan kata dengan karakter (/ /), terdapat angka yang menunjukkan anotasi yang menunjukkan indeks dari kata yang memiliki kemiripan semantik, serta atribut (NULL) [8]. Setelah data telah terpisah dengan atribut yang ada sebelumnya kemudian dilakukan proses *tokenization* atau pemisahan antar kata dan dilakukan proses *lower case* atau merubah semua kata menjadi huruf kecil agar memudahkan sistem melakukan *alignment*. Adapun contoh dari *dataset* yang digunakan dapat dilihat pada Gambar 3.

```
# sentence pair 3
Maha Pemurah lagi Maha Penyayang.
NULL ( / / ) Mahamurah ( 2 / / ) , ( / / ) Maha ( 1 4 / / )  Penyayang ( 5 / / ) . ( 6 / / )
# sentence pair 4
Yang menguasai di Hari Pembalasan.
NULL ( / / ) Yang ( 1 / / ) Mempunyai ( / / ) Hari ( 4 / / ) Pembalasan ( 5 / / ) . ( 6 / / )
```

Gambar 3. Contoh *Dataset* [8]

3) Kemudian sepasang terjemahan yang sudah diproses akan dilakukan *align* dengan mengimplementasikan fitur *alignment*. Fitur *alignment* merupakan tahapan *alignment* yang mengekstrak atau mendapatkan kata dari pasangan teks yang mempengaruhi proses *similarity*. Dalam penelitian ini,

diimplementasikan pendekatan *word similarity* yang mengidentifikasi kata yang serupa dan memiliki makna yang sama. Adapun fitur dari *alignment* yang diterapkan sebagai berikut :

a) Sistem akan melakukan proses *align identical word* dengan mengidentifikasi kemiripan semantik berdasarkan kemiripan yang identik baik makna dan *string* pada kata.

b) Sistem akan melakukan proses *align* PPDB dimana sistem akan mengidentifikasi kata yang mempunyai makna yang sama dari *database* PPDB.

c) Sistem akan melakukan proses *align word sequences* dimana sistem akan memproses kalimat yang mempunyai keterurutan kata yang sama.

4) Setelah dilakukan semua fitur maka sistem akan menampilkan *output* berupa nomor indeks yang ditampilkan menunjukkan hasil dari proses penyelarasan dalam bentuk koleksi pasang ayat-ayat Indonesia dari ayat-ayat Alquran yang memiliki kemiripan semantik setelah adanya proses pensejajaran. Kemudian sistem akan melakukan proses evaluasi dalam bentuk perhitungan *prescision*, *recall* dan F1-skor. Proses evaluasi ini dilakukan untuk mengaetahui seberapa baik hasil yang dihasilkan oleh sistem dibanding dengan pasangan-pasangan terjemahan yang dilakukan oleh manusia secara manual.

B. Dataset Building

Dataset yang digunakan dalam penelitian ini adalah berupa pasangan terjemahan Al-Quran bahasa Indonesia versi terjemahan Departemen Agama dan UII karena merupakan terjemahan yang sudah populer di kalangan masyarakat dari beberapa terjemahan Al-Quran lainnya. *Dataset* versi cetak dilakukan dengan cara proses digitalisasi. Jumlah *dataset* yang digunakan dalam penelitian ini sejumlah 200 pasangan terjemahan ayat Al-Quran yang sama.

C. Data MSR

Data MSR adalah data yang berasal dari Microsoft Research yang dibuat pada tahun 2006 dengan tujuan agar data dapat digunakan dalam berbagai jenis penelitian mulai dari pencarian informasi, semantik kesamaan [8]. Sehingga pada *paper* ini *dataset* yang digunakan berformat MSR. Maksud dari *dataset* berformat MSR artinya Format data file yang beranotasi dikumpulkan dalam dua set data yang sesuai dengan RTE 2006 set pengembangan dan tes [8]. Setiap file berisi daftar pasangan *teks-hipotesis* dalam format berikut:

1) Garis yang menandai awal pasangan *teks-hipotesis*
2) Garis yang barisnya berisi kalimat teks, gaya *Penn Treebank tokenized*
3) Garis yang berisi kalimat hipotesis, masing-masing gaya *Penn Treebank*, kata diikuti oleh daftar *indeks* yang menunjuk ke kata yang sesuai atau kata-kata dalam kalimat teks

Data MSR adalah data yang dimiliki informasi tentang penyelarasan antar kata. Dapat dicontohkan sebagai berikut :
untuk penjelasan "sentence pair" adalah penanda pasangan kalimat, baris setelah pasangan kalimat adalah

```
#sentence pair 1
ECB spokeswoman , Regina Schueller , declined to com- ment on a report
in Italy s La Repubblica newspaper that the ECB council will discuss Mr.  Fazio s
role in the takeover fight at its Sept .  15 meeting .
NULL ( / / ) Regina ( 4 p1 p2 / / ) {Shueller ( 5 p1 p2 / / ) works ( / / )
for ( / / ) Italy ( 14 / / ) s (15 / / ) La ( 16 / / ) Repubblica ( 17 / / )
newspaper( 18 / / ) .  ( 38 / / )
```

Gambar 4. Contoh Format MSR [8]

kalimat pertama dan kata NULL adalah penanda untuk kalimat kedua bersama dengan *gold standar*, angka

dalam tanda kurung adalah *gold standar*, dalam hal ini terdapat dua kondisi yaitu PASTI dan MUNGKIN jika *gold standar* yang pasti maka harus disejajarkan disimbolkan dengan huruf p merupakan *gold standar*. Kondisi kedua adalah mungkin atau dapat disejajarkan. Untuk membaca perintah di atas adalah, Regina akan sejajar dengan kata Regina di indeks keempat di kalimat pertama dan p1, p2 yang berarti Regina dapat dikatakan selaras dengan ECB dan juru bicara [8].

D. Preprocessing

Dalam penelitian ini tahap-tahap *prepocessing* yang digunakan adalah sebagai berikut :

1) *Tokenisasi* adalah proses untuk membagi atau memotong teks yang berupa kalimat, paragraf atau dokumen menjadi token-token/ bagian bagian tertentu. Contohnya jika ada 1 kalimat "Dengan menyebut nama allah yang maha pemurah lagi maha penyayang" maka untuk proses tokenisasi kalimat tersebut akan dipotong menjadi per-kata sebagai berikut "dengan", "menyebut", "nama", "allah", "yang", "maha", "pemurah", "lagi", "maha", "penyayang".

2) Setelah dilakukan proses *tokenisasi*, *preprocessing* yang dilakukan selanjutnya adalah proses *stemming* atau memotong imbuhan kata dalam kalimat menjadi kata dasar dengan bantuan *library stemming* bahasa Indonesia yaitu Sastrawi *Stemmer* [11]. Contohnya adalah imbuhan " meng-, me-, kan-, di- dan sebagainya. Dapat diilustrasikan pada kalimat berikut :
versi 1 : Segala puji bagi Allah, Tuhan semesta alam.
versi 2 : Segala puji-pujian untuk Allah, Pemelihara semesta alam.
maka dari kalimat diatas kata "puji" dengan kata "puji-pujian" pada terjemahan versi 2 karena kata "puji-pujian" untuk kata dasarnya adalah "puji" sehingga memiliki kemiripan semantik.

3) *Stopwords Removal* adalah kata umum (*common words*) yang biasanya muncul dalam jumlah besar dan dianggap tidak memiliki makna sehingga akan dihilangkan. *Stopwords* biasanya dimanfaatkan dalam *task information retrieval*. Pada penelitian adanya penerapan *stopwords removal* dengan menggunakan *Phython* Sastrawi sehingga mengacu pada *list* kata-kata yang terdapat pada *library* sastrawi. Kita bisa menambah atau mengurangi *stopwords* tersebut [11]. Contohnya adalah "yang","ke","para","namun","sambil","sesudah","sebelum".

Proses *preprocessing* diperjelas dalam Tabel I berdasarkan tahap tahap *preprocessing* yang akan dilakukan dalam penelitian ini.
Kalimat 1
mereka itulah yang tetap mendapat petunjuk dari tuhan mereka dan merekalah orang-orang yang beruntung

Tabel I
HASIL DARI PROSES *Preprocessing*

Tokenisasi	**Stopwords**	**Stemming**
mereka	mereka	mereka
itulah	-	itu
yang	-	yang
tetap	-	tetap
mendapat	mendapat	dapat
petunjuk	petunjuk	petunjuk
dari	-	dari
tuhan	tuhan	tuhan
mereka	mereka	mereka
merekalah	merekalah	mereka
orang	orang	orang
yang	-	yang
orang	orang	orang
yang	-	yang
beruntung	beruntung	untung

E. Ekstraksi Fitur

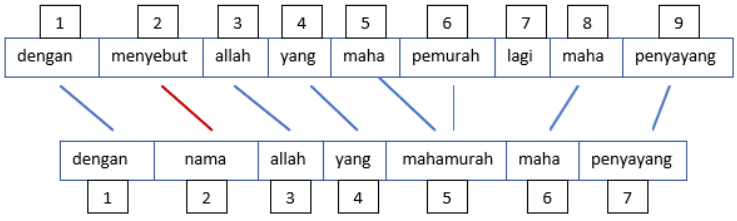
Dalam *paper* ini ada 3 fitur *alignment* yang akan di uji coba untuk menghasilkan informasi yang berharga mengenai bagaimana dan seberapa besar dua buah kalimat saling berkaitan. Sebelumnya fitur ini juga sudah dijelaskan oleh penulis namun dalam bagian ini akan lebih dijelaskan dengan lengkap.

1) *Align Identical Word*
   Fitur atau modul dalam *alignment* yang dapat digunakan ialah dengan urutan kata-kata yang identik. Terdapat dua indikator dalam menentukan kesamaan kata-kata, yang pertama identik sama secara *string* dan yang kedua sama secara kontekstual. Fitur ini adalah fitur *alignment* paling sederahana karena tidak memeriksa arti kata berdasarkan pada pengaturan huruf dalam kata [6] Contohnya sebagai berikut :
   versi 1 : Segala puji bagi Allah, Tuhan semesta alam
   versi 2 : Segala puji-pujian untuk Allah, Pemelihara semesta alam
   dalam kalimat di atas pasangan kata yang akan dipadankan yaitu kata "Segala", "allah", "semesta", "alam" dengan kata "Segala", "allah", "semesta", "alam" pada terjemahan versi 2 karena memiliki *string* kata yang sama persis.
2) *Align Word Sequence*
   Fitur *align word sequences* digunakan untuk mengidentifikasi pasangan kata yang mempunyai urutan kata yang sama dengan minimal 2 kesamaan kata [12]. Adapun *output* yang akan dihasilkan terlihat pada Gambar 5.

Gambar 5. *Contoh Align Word Sequences* Sistem [12]

3) *Align* PPDB
   Align PPDB yaitu *aligner* yang mengandalkan *paraphrase database* dalam menentukan *alignment*, diadaptasi dari *paper* Sultan et al yang di-*submit* ke SemEval [6], kata yang akan di-*align* akan diperiksa dalam PPDB, jika kata yang akan di-*align* terdapat dalam PPDB maka pasangan kata tersebut akan dikategorikan sebagai pasangan kata yang *align*. Penulis melakukan list kata secara manual melalui kamus thesaurus untuk didigitalisasikan sehingga menggantikan PPDB secara otomatis karena PPDB tidak tersedia di Bahasa Indonesia.

Berikut adalah simulasi dari sistem yang akan dilakukan. Pada Gambar 6 diberikan 1 pasang kalimat terjemahan dengan hasil *alignment* dipetakan pada Tabel II.



Gambar 6. Proses *Alignment*

Tabel II merupakan hasil proses dari penggunaan fitur *monolingual word alignment* dengan menampilkan *output* berupa nomor indeks pada pasangan ayat-ayat Al-Quran yang memiliki kemiripan semantik.

Tabel II
HASIL DARI FITUR *Alignment*

Indeks	Kalimat per-Kata		Keterangan
[1,1]	dengan	dengan	*Identical dan Sequence*
[2,2]	menyebut	nama	PPDB
[3,3]	allah	allah	*Identical dan Sequence*
[4,4]	yangt	yang	*Identical dan Sequence*
[5,5]	maha	mahamurah	*Identical dan Sequence*
[5,6]	pemurah	mahamurah	*Identical*
[8,6]	maha	maha	*Identical*
[9,7]	penyayang	penyayang	*Identical*

F. Matrik Evaluasi

Pada Perhitungan evaluasi digunakan proses perhitungan *Confusion Matrix* dengan evaluasi *instrinsik* untuk mendapatkan nilai *prescision*, *recall* dan F1-skor terhadap hasil *alignment* yang telah diproses oleh sistem.

- Recall merupakan perhitungan rasio dari jumlah prediksi yang benar terhadap total prediksi yang diharapkan [13].

$$recall = \frac{TP}{TP + FN} \tag{1}$$

- Precision merupakan perhitungan pengukuran rasio dari jumlah prediksi yang benar terhadap total prediksi [13].

$$precision = \frac{TP}{TP + FP} \tag{2}$$

- Dengan Skor F1 merupakan pengukuran tingkat akurasi pada sistem yang telah dibuat [13].

$$F_1 = \frac{2(precision * recall)}{precision + recall} \tag{3}$$

IV. HASIL DAN DISKUSI

A. Skenario Pengujian

Tabel III menunjukkan beberapa skenario pengujian yang dilakukan dengan melakukan simulasi dan kombinasi dari fitur *monolingual word alignment*. Adapun skenario pengujian dari 5 pengujian yang dilakukan adalah sebagai berikut.

Tabel III
CONTOH DARI KOMBINASI FITUR

Nomor	Skenario Pengujian
1.	Identical Word + Stopword
2.	Identical Word + Alig word Sequence
3.	Identical Word + Align word Sequence + PPDB
4.	Align word Sequence + PPDB
5.	Identical Word + PPDB + Stopword

B. Hasil Pengujian

Tabel IV menunjukkan skor korelasi dari kombinasi fitur pada Tabel III.

Tabel IV
HASIL DARI SKENARIO PENGUJIAN

Skenario Pengujian	Precision	Recall	F1
Identical Word + Align word Sequence	0.68	0.70	0.68
Identical Word + Stopword	**0.84**	0.80	**0.82**
Identical Word + PPDB + Stopword	**0.83**	**0.82**	**0.82**
Identical Word + Align word Sequence + PPDB	0.75	**0.82**	0.78
Align word Sequence + PPDB	0.57	0.63	0.59

C. Analisis dari Hasil Pengujian

Tabel V merupakan hasil dari pengujian dengan "**the best accuracy**" untuk skor F1 0.82 dengan kombinasi fitur menggunakan *identical word*, PPDB dan *stopword*. Untuk hasil *prescision* yang terbaik dihasilkan oleh kombinasi fitur *identical word* dan *stopword* menghasilkan 0.84. Dalam kombinasi *alignment word sequence* dan PPDB menghasilkan F1 0,59, *precision* 0,57 dan *recall* 0,63 ini karena penyelarasan *alignment word sequence* akan digunakan berdasarkan minimal 2 urutan kata yang memiliki kesamaan semantik sehingga jika kata tersebut tidak memiliki Kesamaan semantik tidak akan dihitung dengan benar sehingga memberikan pengaruh ambigu.

Tabel V
Prescision, *recall*, SKOR F1 PADA "THE BEST CASE"

	This work (Full Feature)	**This Work (Best Case)**
Precision	0.75	**0.83**
Recall	0.82	**0.82**
Skor F1	0.78	**0.82**

V. KESIMPULAN

Pada penelitian ini telah dibangun sebuah sistem untuk membangun data latih *monolingual word alignment* pada terjemahan Al-Quran versi bahasa Indonesia berupa data terjemahan yang telah dilakukan proses *align* oleh sistem dengan format data MSR dan divisualisasikan dalam bentuk gambar agar lebih mudah dalam proses pembacaannya.

Dalam penelitian pembangunan *monolingual word alignment* pada terjemahan Al-Quran versi bahasa Indonesia menghasilkan skor F1 terbaik didapat dari kombinasi dari 3 fitur *alignment* yang diimplementasikan yaitu *align identical words*, *stopwords removal* dan *align database* PBBD dengan hasil F1 tertinggi sebesar 0.82. Penggunaan *stopwords removal* terbukti dapat meningkatkan hasil akurasi sistem dibandingkan tanpa penggunaan *stopwords*. Penelitian lebih lanjut dapat dilakukan dengan menambahkan fitur *contextual similarity* seperti *align named entity* bahasa Indonesia khusus untuk data terjemahan Al-Quran agar POStag yang dihasilkan sesuai dengan nama yang ada di Al-Quran dan fitur *align content words* yang meliputi fitur *align dependency* dan *align neighbourhood*.

PUSTAKA

[1] H. A. H. Sanaky, *Metode Tafsir [Perkembangan Metode Tafsir Mengikuti Warna atau Corak Mufassirin*. Al-Mawarid Edisi XVIII Tahun 2008, 2008.

[2] K. M. Kapil Thadani, *Optimal and Syntactically-Informed Decoding for Monolingual Phrase-Based Alignment*. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2011.

[3] C. D. Bill MacCartney, Michel Galley, *A Phrase-Based Alignment Model for Natural Language Inference*. Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2008.
[4] M. I. Danushka Bollegala, Yutaka Matsuo, *Measuring Semantic Similarity between Words Using Web Search Engines*. Information Systems, Information Search and Retrieval, 2017.
[5] T. Brychcin and L. Svoboda, *UWB at SemEval-2016 Task 1: Semantic textual similarity using lexical, syntactic, and semantic information*. Proceedings of the 10th InternationalWorkshop on Semantic Evaluation (SemEval 2016), San Diego, CA, 2016.
[6] S. M. A. Steven Bethard and T. Sumner, *Back to Basics for Monolingual Alignment: Exploiting Word Similarity and Contextual Evidence*. aclweb.org, 2014.
[7] C. Fujita, Inuiárez, *A Class-oriented Approach to Building a Paraphrase Corpus*, 2003, vol. 1.
[8] C. Brockett, *Aligning The RTE 2006 Corpus*. Natural Language Processing Group, Microsoft Research Technical Report MSR-TR-2007-77, 2007.
[9] C. C.-B. Xuchen Yao, Benjamin Van Durme and P. Clark, *A Lightweight and High Performance Monolingual Word Aligner*. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics*, 2013a.
[10] C. C.-B. Xuchen Yao, Benjamin Van Durme and P. Clark., *Semi-Markov Phrase-based Monolingual Alignment.* In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics*, 2013b.
[11] K. V. Ghag and K. Shah, "Comparative analysis of effect of stopwords removal on sentiment classification," *2015 International Conference on Computer, Communication and Control (IC4)*, pp. 1–6, 2015.
[12] L. T. C. O. R. M. Hanna Bechara, Rohit Gupta and J. van Genabith, *Replicating the success of monolingual word alignment and neural embeddings for semantic textual similarity*. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), 2016.
[13] J. Euzenat, "Semantic precision and recall for ontology alignment evaluation." 2007, pp. 348–353.