Pembangunan Synonym Set Kosa Kata Al-Quran dengan Pendekatan WordNet

Laras Gupitasari*[1], Moch. Arif Bijaksana[2], Arief Fatchul Huda[3]
[1,2]Universitas Telkom; Jl. Telekomunikasi No.01, Dayeuhkolot, Bandung, (022) 7565930
[3]UIN Sunan Gunung Djati; Jl. A.H. Nasution No.105, Cipadung, Bandung, (022) 7800525
[1,2]Program Studi Informatika, Bandung
[3]Fakultas Sains dan Teknologi, Bandung
e-mail: ***[1]**larasgupitasari@student.telkomuniversity.ac.id,
[2]arifbijaksana@telkomuniversity.ac.id, [3]afh@uinsgd.ac.id

Abstrak

Penelitian mengenai Al-Quran di bidang computational linguistics sangat menarik dan bermanfaat apabila dipandang dari seberapa pentingnya Al-Quran bagi umat Islam karena merupakan kitab suci Agama Islam. Penelitian ini bertujuan untuk membangun synonym set Al-Quran karena saat ini sumber daya untuk melakukan penelitian mengenai Al-Quran dapat dikatakan masih kurang. Dataset yang digunakan pada penelitian ini yaitu kata benda terjemahan Bahasa Inggris Al-Quran. Untuk menghasilkan synonym set, penelitian ini mengelompokkan kata menggunakan metode hierarchical clustering dan jarak antar kata dihitung menggunakan path similarity dari WordNet.Evaluasi pada penelitian ini menghasilkan nilai F-Measure sebesar 83% yang merupakan kesesuaian hasil synonym set oleh sistem dan hasil synonym set oleh ahli.

Kata kunci—Al-Quran, Synonym Set, Tesaurus, WordNet, Clustering

Abstract

Research about Quran in the field of computational linguistics is very interesting anduseful if it viewed from how important Quran is for the Muslims because it is the holy book ofIslam. This is research aims to establish a synonym set of Quran because nowadaysresources to conduct research on Al-Quran can be said still not enough. The dataset used inthis research is the english translated noun from the Quran. To produce synonym sets, thisresearch groups word using a hierarchical clustering method and the distance between words ismeasured using path similarity from WordNet. The evaluation of this research results gained F-Measure value 83% which is the suitability of the results of the synonym set by system and synonym setresults by linguist.

Keywords—Quran, Synonym Set, Thesaurus, WordNet, Clustering

1. PENDAHULUAN

Saat ini telah banyak penelitian linguistik dan *Natural Language Processing* (NLP) yang memanfaatkan Princeton WordNet (PWN)[1] sebagai bahan penelitian seperti pada pembangunan WordNet untuk Bahasa Turki [2], Bahasa Arab [3], Bahasa China [4], pembangunan Synonym Set untuk Bahasa Indonesia [5], dan masih banyak penelitian lainnya.

Penelitian ini memiliki tujuan untuk membangun *synonym set* atau *synset* Bahasa Arab pada Al-Quran dengan memanfaatkan *lexical semantic similarity* pada PWN. Pada penelitian sebelumnya mengenai pembangunan WordNet Bahasa Arab [3] telah berhasil membangun *lexical database*WordNet dengan fokus Bahasa Arab modern. Pada penelitian ini akan berfokus pada kosa kata dalam Al-Quran.

Al-Quran memiliki surah sebanyak 114 surah dan ayat berjumlah 6.236 ayat. Dari sekian banyak ayat yang ada, setiap ayat memiliki kaitan dengan sebagian ayat yang lainnya. Begitupun dengan kata dalam setiap ayat juga memiliki kaitan dengan sebagian kata yang lainnya, salah satu kaitannya yaitu kesamaan makna dari setiap kata. Bagi pembaca yang kurang memahami Bahasa Arab dan hanya membaca beberapa kata saja dalam Al-Quran, dapat menyebabkan kurangnya pemahaman mengenai arti sebenarnya dari setiap kata dalam Al-Quran.

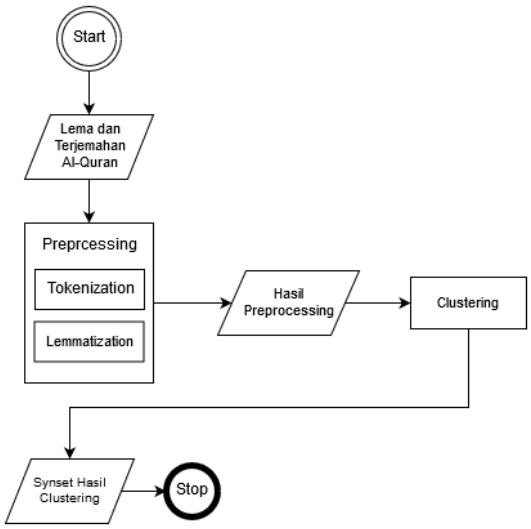
Setiap kata dalam Al-Quran dapat memiliki makna lebih dari satu, contohnya adalah kata ءَايَة (āyat) yang memiliki arti *sign* (QS. 2:39:4) dan *verse*(QS. 2:99:4). Lalu kata كِتَٰبِى (kitābī) yang memiliki arti *book* (QS. 2:2:2), *scripture* (QS. 2:23:8), dan *record* (QS. 11:6:15). Selain kata yang memiliki makna lebih dari satu, sebagian kata juga memiliki arti yang sama atau dekat, contohnya adalah kata رَبّ (rab) yang memiliki arti *Lord* (QS. 1:2:3) dan إِلَٰه (ilāh) yang memiliki arti *God* (QS. 2:133:17) dimana *Lord* dan *God* memiliki kedekatan makna yaitu Tuhan.

Sumber daya bahasa sangat penting untuk penelitian komputasi linguistik khususnya *Natural Language Processing* (NLP). Namun untuk korpus Bahasa Arab Al-Quran saat ini masih tergolong kurang walaupun beberapa tahun terakhir telah ada pengembangan korpus-korpus Bahasa Arab yang bebas akses. Korpus-korpus tersebut dapat dikatakan belum cukup mendorong para ahli bahasa untuk menerapkannya pada studi berbasis korpus mereka [6]. Oleh karena itu penelitian ini menjadi penting untuk membangun *synonym set* yang dapat digunakan sebagai *prototype* untuk membangun tesaurus Al-Quran.*Synonym set* yang terbentuk juga dapat digunakan untuk lebih memahami kosa kata – kosa kata pada Al-Quran.

Pendekatan utama dari penelitian ini adalah WordNet dan kamus monolingual yaitu tesaurus. Seperti yang kita tahu, WordNet merupakan *lexical database* Bahasa Inggris yang popular dan banyak digunakan oleh peneliti bidang komputasi linguistik dan NLP. Tesaurus juga merupakan kamus monolingual yang menyediakan hubungan sinonim antar kata dalam suatu bahasa [7] dan dapat diakses dengan bebas pada thesaurus.com atau www.lexico.com. Tersedia juga tesaurus dalam bentuk *offline*[7], [8]

2. METODE PENELITIAN

Data yang digunakan dalam penelitian ini adalah lema Al-Quran dan terjemahan Bahasa Inggris yang dapat diunduh di openburhan.net. Data tersebut berisikan total 77.794 baris data yang merupakan banyaknya kata yang ada dalam Al-Quran. Data tersebut akan diproses oleh sistem dengan memilah kata sehingga hanya kata benda saja yang akan digunakan. Proses dalam sistem dapat dilihat pada Gambar 1.



Gambar 1. Sistem Pembangunan Synonym Set

2.1 *Preprocessing*

Tujuan dari *preprocessing* ini adalah agar data yang digunakan siap untuk diproses oleh sistem. Karena data yang berkualitas tinggi akan menghasilkan hasil yang berkualitas tinggi juga dan dapat mengurangi biaya [9]. Data mentah dari openburhan.net berisikan daftar kata dasar Bahasa Arab atau dapat disebut lema, terjemahan setiap katanya dalam Bahasa Inggris dan juga letak setiap kata dalam Al-Quran. Karenatidak semua data mentah akan digunakan maka akan dilakukan proses *preprocessing* untuk memilah apa saja yang penting dan akan digunakan dalam penelitian ini. Proses *preprocessing* pada penelitian ini dilakukan dengan dua tahap yaitu tokenisasi dan lematisasi.

2.1.1 Tokenisasi

Terjemahan satu kata Bahasa Arab dalam Al-Quran tidak selalu memiliki satu arti saja, namun bergantung pada makna setiap ayatnya.Misalnya kata هُدًى (hudan) yang memiliki arti *'for [the] guidance'*. Karena penelitian ini berfokus pada kata benda saja maka untuk mendapatkan kata *guidance* perlu dilakukan proses tokenisasi untuk memisahkan token yang ada pada terjemahan sekaligus juga menghilangkan karakter-karakter yang tidak diperlukan. Tabel 1berisikanbeberapa contoh proses tokenisasi.

Tabel 1. Proses Tokenisasi

Sebelum proses tokenisasi	Setelah proses tokenisasi
['the lord', 'their lord', 'your lord', 'his lord', 'your lord']	['lord']
['the earth', 'and the earth', 'the earth', 'the earth', 'the earth']	['earth']
['to his people', 'o my people', 'for his people', 'for people', 'a group']	['people'], ['group']
['our signs', 'my signs (for)', 'in (the) signs', 'his signs']	['signs'], ['verses'], ['sign']
['with [the] messengers', 'a messenger', 'and his messengers', 'a messenger']	['messengers'], ['messenger']

2.1.2 Lematisasi

Setelah setiap kata pada terjemahan dipisahkan, selanjutnya adalah mengubah setiap kata menjadi bentuk kata dasarnya karena akan digunakan sebagai entri pada WordNet. Tabel 2berisikan hasil dari proses lematisasi dan entri yang akan digunakan pada sistem.

Tabel 2. Proses Lematisasi

Hasil proses lematisasi	Entri pada sistem
['lord']	['رَبّ', 'lord']
['earth']	['أرْض', 'earth']
['people','group']	['قَـوْم', 'people'], [`قَـوْم', 'group']
['sign', 'verse']	['ءَايَـة', 'sign'], ['ءَايَـة', 'verse']
['messenger']	['رَسُول', 'messenger']

2.1.3 Pengelompokan dengan Teknik Clustering

Pada penelitian ini, teknik yang digunakanuntuk menggabungkan lema-lema yang memiliki kedekatan makna adalah menggunakan metode *hierarchical clustering*. Alasan memilih *hierarchical clustering* karena melihat dari tujuan penelitian ini yaitu untuk membangun *synonym set* yang tidak diketahui jumlahnya pada awal proses. Sehingga dengan menggunakan *hierarchical clustering* inisialisasi jumlah *cluster* tidak perlu dilakukan pada awal proses dan jumlah *cluster* tidak dapat diprediksi sebelum proses *clustering* selesai [10]. Fleksibilitas yang dimiliki oleh *hierarchical clustering*juga berguna pada biaya perhitungan yang lebih kecil [11].

Algoritma 1 adalah algoritma *clustering* yang digunakan pada penelitian ini.

Algorithm 1 Teknik Klasterisasi

1: **Input:** Himpunan lema dan terjemahan
2: **for** i **to** semua $lema_i$
3: **for** j **to** semua $lema_j$
4: Hitung jarak antara lema i dengan lema j
4: **if** similarity ≥ threshold **then**
6: Letakkan kedua lema ke dalam satu klaster
7: **Else**
8: Letakkan lema pada klaster yang berbeda
9: **Output:** Himpunan *synonym set*

Perhitungan jarak dihitung menggunakan PATH *similarity*pada WordNet yaitu mengukur kedekatan antar kata dengan menghitung jumlah node pada jalur terpendek antar kata dan hubungan '*is-a*' antar kata pada WordNet [12]. Untuk mendapatkan jarak tersebut menggunakan Persamaan 1.

$$PATH(s1, s2) = \frac{1}{path\_length(s1,s2)} \quad (1)$$

$s1$merupakan kata pertama yang akan dibandingkan dengan kata kedua yaitu$s2$. Nilai minimum yang akan dihasilkan adalah 0 dan nilai maksimumnya adalah 1.

Setelah perhitungan jarak antar kata, jarak tersebut dibandingkan dengan *threshold* sebesar 0.5. Nilai tersebut didapatkan atas hasil proses percobaan pada beberapa nilai dan dengan nilai 0.5 dapat menghasilkan *synset* yang cukup baik. Jika *similarity* sama dengan atau lebih besar dari *threshold* artinya antar kata memiliki kedekatan yang cukup besar dan jika *similarity*lebih kecil dari *threshold* artinya kedekatan antar kata adalah kecil.

Lema-lema yang memiliki kedekatan besar akan digabungkan menjadi satu *cluster* atau himpunan kata yang memiliki kedekatan makna. Sehingga lema yang sama dapat berada lebih dari satu himpunan. Dan dalam satu himpunan dapat berisikan lebih dari satu lema yang memiliki kedekatan makna atau dapat disebut sinonim.

3. HASIL DAN PEMBAHASAN

Total lema kata benda yang digunakan pada penelitian ini adalah sebanyak 5.615 kata benda yang unik dalam Al-Quran. Dari 5.615 kata tersebut terdapat 427 kata atau sama dengan 11% keseluruhan data tidak ditemukan dalam *database* WordNet. Diantaranya kata *sulaiman, firaun, safa, marwah* dan lainnya yang merupakan kata dalam bahasa islami dan hanya ada pada Al-Quran. Tidak adanya kata-kata tersebut dalam *database* WordNet adalah karena WordNet merupakan *lexical database* gabungan dari kamus dan tesaurus Bahasa Inggris [13], sehingga untuk kata-kata tersebut bukan merupakan kata yang ada dalam kamus maupun tesaurus Bahasa Inggris.Kata lain yang juga tidak ada pada WordNet adalah kata *I, my, they, him,* dan lainnya karena merupakan kata dalam *closed-class words* pada WordNet terminologi [14].

Hasil *synset* oleh sistem dapat dilihat pada Tabel 3.

Tabel 3. Jumlah Hasil Synset yang Terbentuk

Synset berisi single member	Synset berisi multiple member	Total
12	4552	4564

Dari keseluruhan dataset sistem menghasilkan total *synset* sebanyak 4.564 dengan 12*synset* berisikan satu lema saja dan 4.552 *synset* berisikan lebih dari satu lema. Beberapa contoh hasil *synset* beserta terjemahannya dapat dilihat pada pada Tabel 4.

Tabel 4. Beberapa Hasil Synonym Set Beserta Terjemahannya

Entri	**Synonym Set**
ٱسْم	['ٱسْم', 'name', 'سَمِيّ', 'name', 'تَسْمِيَة', 'name']
ٱللَّه	['ٱللَّه', 'allah', 'إِلَٰه', 'god', 'ٱللَّهُمَّ', 'allah']
حَمْد	['حَمْد', 'praise', 'أَوَّاب', 'praise']
	['حَمْد', 'thanks', 'شَكَرَ', 'thanks', 'شُكُور', 'thanks']
رَبّ	['رَبّ', 'lord', 'بَارِئ', 'creator', 'إِلَٰه', 'god', 'فَاطِر', 'creator', 'خَٰلِق', 'creator', 'خَلَّٰق', 'creator', 'إِلَٰه', 'lord']
عَٰلَمِين	['عَٰلَمِين', 'universe', 'عَٰلَمِين', 'world', 'دُنْيَا', 'world']
مَٰلِك	['مَٰلِك', 'master', 'رَبّ', 'master', 'مَوْلَىٰ', 'master']

Untuk mengetahui akurasi hasil *synonym set* yang terbentuk perlu dilakukan evaluasi. Pada penelitian ini metode yang digunakan untuk evaluasi adalah metode F-Measure karena F-Measure dapat digunakan untuk mengukur akurasi metode *clustering*[15] dengan melibatkan dua factor yaitu *recall* dan *precision*. Selain itu, proses evaluasi ini juga melibatkan ahli bahasa yaitu seseorang yang ahli dalam pemahaman Al-Quran untuk membuat *gold standard* yaitu *synset* yang benar menurut ahli. Persamaan 2 merupakan persamaan untuk menghitung *recall* dan Persamaan 3 adalah persamaan untuk menghitung *precision*.

$$Recall = \frac{N_{ij}}{N_i} \quad (2)$$

$$Precision = \frac{N_{ij}}{N_j} \quad (3)$$

Recall merupakan seberapa banyak elemen benar dalam *synset* yang terpilih, sedangkan *precision* adalah seberapa banyak elemen dalam *synset* terpilih merupakan elemenyang benar.$N_i$adalah jumlah elemen pada *synset* oleh sistem danN_j adalah jumlah elemen pada*gold standard*. Sedangkan N_{ij}merupakan jumlah elemen *synset* oleh sistem yang ada pada *gold standard*. Kemudian persamaan untuk menghitung F-Measure terdapat pada Persamaan 4 dengan simbol F.

$$F = \frac{2*Recall*Precision}{Recall+Precision} \quad (4)$$

Tabel 5 berisikan hasil perbandingan *synset* oleh sistem dan *synset* oleh ahli atau *gold standard*menggunakan 7 data uji.

Tabel 5.Perbandingan Synset

Entri Kata	Synset oleh Sistem	Synset oleh Ahli (gold standard)
رَبّ	[خَلَّق , خَلِق , فَاطِر , إِلَٰه , بَارِئ , رَبّ]	[خَلَّق , خَلِق , فَاطِر , إِلَٰه , بَارِئ , رَبّ]
عَذَاب	[رِجْس , بَأْس , جَزَآء , عِقَاب , نَكَل , رِجْز , عَذَاب , صَعُود , نَكِير]	[بَأْس , جَزَآء , عِقَاب , نَكَل , رِجْز , عَذَاب]
نَّاس	[أُمَّة , طَآئِفَة , فِئَة , فَرِيق , إِنسَٰن , قَوْم , ءَال , نَّاس , عُصْبَة , قَرْن , قَوْم]	[قَوْم , أُمَّة , فَرِيق , إِنسَٰن , قَوْم , ءَال , نَّاس , بَشَر , أَنَام , قَرْن]
حَذَر	[رَّوْع , خَشِىَ , رُعْب , خَوْف , اتَّقَى , حَذَر]	[رَّوْع , خَشِىَ , رُعْب , خَوْف , حَذَر]
ظَالِم	[مُسِىٓء , غَرُور , خَٰطِئِين , أَثِيم , مُعْتَدِين , عَاد , ظَالِم]	[مُسِىٓء , غَرُور , أَثِيم , مُعْتَدِين , عَاد , ظَالِم]
وَالِد	[أُمّ , آبَاء ,وَالِد , أَبَوَان , مَوْلُود , وُلِدَت , وَالِد]	[أُمّ , آبَاء , وَالِد , أَبَوَان , مَوْلُود , وُلِدَت , وَالِد]
شَيْطَٰن	[سُوٓء , بِئْسَ , فُسُوق , سَيِّئَة , طُغْيَٰن , شَيْطَٰن]	[خَطِيَٓئَة , جُنَاح , بِئْسَ , فُسُوق , طُغْيَٰن , شَيْطَٰن , وَازِرَة , سُوٓء , ذَنب]

Menggunakan F-Measure sebagai metode evaluasi sistem mendapatkan nilai *recall* sebesar 81%, nilai *precision* sebesar 86%, dan F-measure sebesar 83%. Faktor yang mempengaruhi nilai *recall* dan *precision* adalah karena terdapat selisih antara jumlah *synset* oleh sistem dan jumlah *synset* oleh *gold standard* sehingga mempengaruhi ketepatan elemen yang relevan.

4. KESIMPULAN

Pada penelitian ini hasil penggabungan kata menggunakan metode*hierarchical clustering*dan pendekatan WordNet menghasilkan *synonym set* kosa kata Al-Quran sebanyak 4.564. Hasil *synonym set* dari penelitian ini dapat digunakan sebagai *prototype* untuk pembuatan korpus Al-Quran guna menambah sumber daya penelitian Al-Quran dan juga dapat digunakan oleh umat Islam untuk mempelajari kitabnya dengan lebih baik lagi.

5. SARAN

Untuk penelitian selanjutnya diharapkan dapat menggunakan kelas kata yang lebih banyak dan tidak hanya kata benda saja. Pengukuran kedekatan antar kata menggunakan WordNet tidak dapat menangani kata yang tidak ada dalam kamus dan tesaurus Bahasa Inggris hal tersebut dapat mengurangi elemen dari *synonym set*. Sehingga, untuk menentukan jarak antar kata dan pengelompokan kata, metode lain layak dicoba demi meningkatkan akurasi dari pembangunan *synonym set* kosa kata Al-Quran.

DAFTAR PUSTAKA

[1] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller 1990, *"Introduction to WordNet: An on-line Lexical Database*," *Int. J. Lexicogr.*

[2] O. Bilgin, Ö. Çetinoğlu, O. Cetinoglu, and K. Oflazer 2004, "*Building a WordNet for Turkish," Rom. J. Inf. Sci. Technol., Vol. 7, No. 1–2, pp. 163–172,*

[3] S. Elkateb *et al.* 2006, *"Building a WordNet for Arabic," in Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006, pp. 29–34*.

[4] S. Wang and F. Bond 2013, "Building The Chinese Open Wordnet (COW): Starting from Core Synsets," in *Proceedings of The 11th Workshop on Asian Language Resources*, pp. 10–18.

[5] A. Saputra and Others 2010, *"Building Synsets for Indonesian Wordnet with Monolingual Lexical Resources," in 2010 International Conference on Asian Language Processing, pp. 297–300*.

[6] A. Al-Thubaity, M. Khan, M. Al-Mazrua, and M. Al-Mousa 2013, *"New Language Resources for Arabic: Corpus Containing More Than Two Million Words and A Corpus Processing Tool," in 2013 International Conference on Asian Language Processing, pp. 67–70.*

[7] T. Redaksi, "Tesaurus Bahasa Indonesia Pusat Bahasa, 2009, *" Pus. Bahasa, Dep. Pendidik. Nas., 2008.*

[8] M. Waite, *Oxford Thesaurus of English*, Oxford University Press,

[9] S. A. N. Alexandropoulos, S. B. Kotsiantis, and M. N. Vrahatis 2019, *Data Preprocessing in Predictive Data Mining*, Vol. 34. Springer.

[10] K. N. PUTRI, 2019, *"Clustering Ekstraksi Synonym Set Bahasa Indonesia Menggunakan*
Agglomerative Hierarchical Clustering,", Skripsi, Program Studi Informatika, Universitas
Telkom, Bandung.

[11] Y. Rani and H. Rohil, "*A Study of Hierarchical Clustering Algorithm," Int. J. Inf.*

Comput. Technol., p. 113, 2013.

[12] T. Pedersen, S. Patwardhan, and J. Michelizzi 2004,"*WordNet:: Similarity: Measuring The Relatedness of Concepts," in Demonstration Papers at HLT-NAACL 2004, , pp. 38–41.*

[13] K. Samhith, S. A. Tilak, and G. Panda 2016, *"Word Sense Disambiguation Using Wordnet Lexical Categories," in 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES), pp. 1664–1666.*

[14] D. A. Wiranata, M. A. Bijaksana, and M. S. Mubarok 2018, *"Quranic Concepts Similarity Based on Lexical Database," in 2018 6th International Conference on Information and Communication Technology (ICoICT), pp. 264–268.*

[15] S. Chormunge and S. Jena, 2015, *"Efficiency and Effectiveness of Clustering Algorithms for High Dimensional Data," Int. J. Comput. Appl.*