

QUESTION ANSWERING TERJEMAH AL QUR'AN MENGGUNAKAN NAMED ENTITY RECOGNITION

Lukman Fakhid Lidimilah¹⁾

¹Manajemen Informatika, AMIK Ibrahimy Situbondo
email: luky.lukman7@gmail.com

Abstact

Al Quran is the guidance of all Muslims all over the world, including Indonesian society that most of the people are moslem. the purpose of this research is to help the society for seeking the quranic proposition in the free question forms or more often known Question Answering System (QAS). In this study the algorithm used to provide the answers is Rule Based. Rule Based on previous research by indexing documents using lucene with the same set of data and same category (Who, When, and Where) getting low accuracy result (60%, 60%, and 40%). In this research, the improvement of architecture by involving Named Entity Recognition (NER) and improvement of indexing and determination of candidate answer using Vector Space Model. The results of the improved architecture are proven increased accuracy with each accuracy of each category of 90%, 80%, and 50%.

Keywords: rule based, named entity recognition, vector space model

1. PENDAHULUAN

Al Qur'an adalah sumber hukum bagi umat muslim diseluruh dunia. Setiap muslim tentu sangat ingin mempelajari isi kandungan Al Qur'an, akan tetapi tidak semua orang mampu memahami Al Qur'an dengan bahasa aslinya yaitu bahasa Arab. Dengan adanya terjemah Al Qur'an maka isi kandungan Al Qur'an dapat tersampaikan meskipun tidak memahami bahasa Arab^[1].

Question Answering System (QAS) digunakan dalam banyak bahasa diantaranya bahasa Indonesia^[1], bahasa Arab seperti QARAB yang diusulkan oleh Hammo, dkk^[2], dan bahasa Turki yang diusulkan oleh Derici, dkk^[3], dan masih banyak lagi penelitian tentang QAS dengan bahasa lain. Tujuan QAS adalah untuk memecahkan problem dimana pengguna sering mempunyai pertanyaan-pertanyaan yang tentu ingin mendapatkan jawaban tepat dan akurat, sedangkan mereka selalu ingin mengungkapkan pertanyaan-pertanyaan mereka dengan bahasa alami^[2].

Ada beberapa metode yang pernah diusulkan untuk QAS dari berbagai bahasa seperti *N-Gram Passage*, *Rule Based*, *Vector Space Model (VSM)*. Dari penelitian-penelitian yang sudah dilakukan sebelumnya tentu metode-metode tersebut memiliki kelebihan dan kekurangan, seperti VSM dan *N-Gram* memiliki akurasi yang cukup baik akan tetapi waktu yang dibutuhkan untuk menjawab pertanyaan pengguna membutuhkan waktu

yang lama^{[4][5][6]}, sedangkan penelitian menggunakan metode Rule-based pada QAS berbahasa indonesia khususnya pada terjemah Al Qur'an sudah pernah dilakukan oleh Anggraeny, akurasi jawaban yang diperoleh cukup baik akan tetapi pada beberapa tipe pertanyaan tertentu seperti mengapa dan mana memiliki akurasi yang rendah, dengan masing-masing akurasi 68.03% untuk tipe mengapa dan 76.25% untuk tipe mana^[1].

Dalam penelitian lain juga dilakukan oleh Gusmita^[7] dengan mengusulkan indexing dokumen yang relevan menggunakan library *lucene*, akan tetapi penelitian tersebut menghasilkan akurasi kecil yaitu 40-60%, hal ini disebabkan keterbatasan *lucene* untuk memproses dokumen Indonesia.

Dalam penelitian ini mengusulkan *Named Entity Recognition (NER)* pada data (dokumen terjemah Al Qur'an) untuk mengatasi problem akurasi pada penelitian sebelumnya ketika menggunakan *lucene*^[7], karena NER dapat menyaring string baik berupa kalimat atau kata yang mengandung jawaban atau tidak dengan memberi kelas pada kata atau kalimat dalam dokumen^[8]. Penggunaan NER ini sudah banyak dilakukan dalam penelitian yang terkait dengan ekstraksi dari dokumen karena NER membantu proses ekstraksi untuk meningkatkan akurasi jawaban, seperti penelitian yang dilakukan oleh Chen, dkk untuk ekstraksi bahasa China^[9], dan juga dipakai dalam penelitian untuk QAS Japanese yang

dilakukan oleh Liu, dkk^[10], bahasa Arab yang dilakukan oleh Akour, dkk^[11].

2. KAJIAN PUSTAKA

Penelitian dengan kasus *Question Answering System* telah banyak dilakukan oleh peneliti-peneliti sebelumnya baik itu tentang QAS bahasa Inggris, Jepang, hingga Arab. Oleh sebab itu perlu kita telaah kembali hasil dari peneliti-peneliti sebelumnya sebagai acuan untuk mengusulkan pendekatan-pendekatan yang terbaik.

Pada penelitian yang dilakukan oleh M. Akour yaitu tentang QArabPro, melakukan pengkajian terhadap bahasa Arab, karena menurutnya pada kasus QAS untuk bahasa arab masih belum matang karena beberapa problem, beberapa diantaranya adalah problem tentang analisis morfologi yang kompleks. Metode yang digunakan dalam penelitian ini adalah berbasis rule (*Rule Based*) dengan mengikuti trend *QA system* baru yaitu dengan menentukan kandidat jawaban sebelum dilakukan seleksi menggunakan metode *Rule Based*. Dalam QArabPro ini menggunakan tujuh tipe pertanyaan yaitu *من , ما , ماذا , اين , متى , لماذا , كم* dengan masing-masing akurasi yang diperoleh adalah 66.77%, 62.22%, 85.19%, 93.62%, 88.89%, 91.67%, 91.67%. Hasil yang diperoleh ini menyajikan semua tipe pertanyaan dimana pada penelitian sebelumnya hal itu tidak dilakukan. Hasil evaluasi yang dilakukan, akurasi keseluruhan mencapai 84%, akan tetapi pada pertanyaan *why* and *how(many/much)* masih relative rendah dengan masing-masing akurasi 62% untuk pertanyaan *why* dan 69% untuk pertanyaan *how(many/much)*^[11].

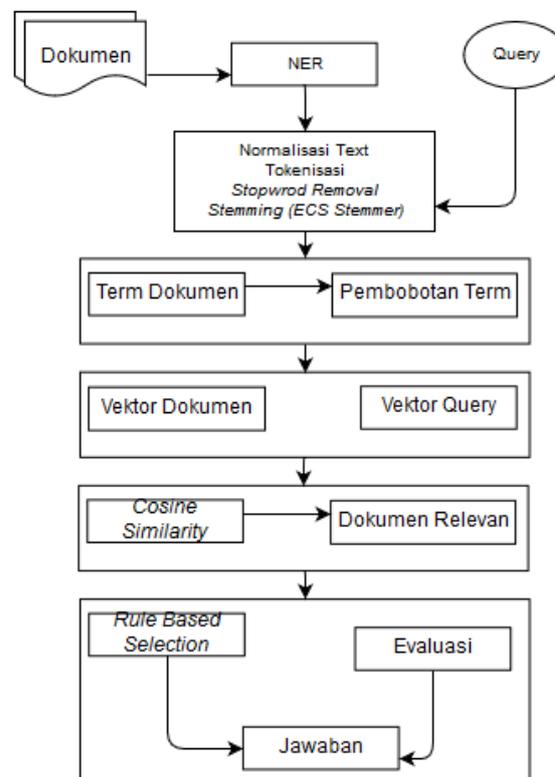
Pada penelitian lain yang dilakukan oleh Gusmita^[7] menyebutkan bahwa pada dua penelitian sebelumnya menurut Gusmita masih belum efisien, sebab pada peneliti pertama terdapat kekurangan pada pengambilan dokumen dan ekstraksi jawaban, karena tidak semua dokumen diambil yang relevan saja dan untuk algoritma ekstraksi jawaban masih menggunakan akal leksikal. Sedangkan pada peneliti kedua tidak efisien karena metode *Rule Based* untuk ekstraksi jawaban diterapkan pada seluruh dokumen. Dalam penelitian ini mengusulkan arsitektur baru dengan menggabungkan dua arsitektur pada penelitian sebelumnya dengan tidak melibatkan analisis dokumen karena digantikan oleh *Rule Based* pada dokumen yang relevan, dan dalam

penelitian ini hanya menganalisa tiga tipe pertanyaan yaitu “siapa, kapan dan dimana”. Hasil yang diperoleh dalam penelitian ini, tidak lebih baik dari penelitian sebelumnya dengan masing-masing akurasi 60%, 60%, dan 40%, penyebab utama adalah karena keterbatasan *Lucene* dalam memproses bahasa Indonesia.

Selain dua peneliti diatas tentang QAS ada tiga peneliti lain yang menjadi rujukan dalam penelitian ini, yaitu Jovita dengan metode yang digunakan adalah *Vector Space Model* [6], Anggraeny dengan metode *Rule Based*^[1], dan Homaidi dengan metode *N-Gram Passage Retrieval*^[5].

3. METODE PENELITIAN

Dalam penelitian ini pendekatan metode atau metode yang diusulkan untuk meningkatkan akurasi dari *Rule Based* adalah dengan *Named Entity Recognition (NER)* dan perbaikan indexing dokumen menggunakan *ECS Stemmer (Enhanced Confix Stripping Stemmer)*. Arsitektur usulan dalam penelitian ini dapat dilihat pada Gambar 1.



Gambar 1 : Arsitektur *QA System*

3.1 Bahan Penelitian

Data adalah salah satu faktor penting yang akan digunakan sebagai bahan pengujian model, dengan adanya data maka model dapat di uji untuk mengetahui seberapa akuratkah

model yang diusulkan. Pada penelitian ini data yang digunakan dalam bentuk dokumen teks yaitu terjemahan Al Qur'an untuk surat Al Baqarah. Data di peroleh dengan cara menyetik dan menyalin dari dokumen PDF terjemah Al Qur'an digital yang mengacu pada Terjemah Al Qur'an DEPAG tahun 1989^[12].

3.2 Named Entity Recognition (NER)

Named Entity Recognition (NER) adalah bagian dari IE (*Information Extraction*) untuk mengidentifikasi kata atau kalimat untuk diberi tanda atau nama (seperti nama orang, tempat, jumlah uang dan lainnya), hal ini penting untuk banyak proses ekstraksi^{[13][14]}. Dalam penelitian ini ketentuan yang digunakan untuk mengidentifikasi waktu diwakili dengan tag <WT1>, tagging *name* diwakili dengan <NM1>, dan untuk lokasi diwakili dengan <LO1>.

3.3 Normalisasi Teks

Tahapan ini melakukan proses pengubahan teks menjadi huruf kecil (*lowercase*) dan menghapus simbol, dan karakter selain alphabet kecuali yang sudah terdefinisi.

3.4 Tokenisasi

Tokenisasi melakukan proses pemotongan karakter menjadi kata tunggal atau token. Pengambilan token tersebut dengan cara memisah semua kata dengan parameter spasi.

3.5 Penghapusan *Stopword* (*Stopword Removal*)

Dalam tahap ini akan dilakukan proses filtering, dimana kata yang sering dimunculkan dalam dokumen akan dihapus, seperti dan, yang atau tetapi dan sebagainya. Semua kata yang mempunyai frekuensi kemunculan yang tinggi akan di satukan dalam stoplist. Stoplist yang sudah dibuat akan dijadikan acuan untuk menghapus kata-kata yang ada dalam dokumen.

3.6 Stemming

Stemming adalah salah satu proses dari *Information Retrieval* (IR) untuk mencari kata dasar. Banyak algoritma yang dapat dipakai untuk *stemming* salah satunya adalah algoritma *stemming* Nazief dan Adriani yang digunakan untuk *stemming* bahasa indonesia.

Algoritma *stemming* Nazief dan Adriani atau dikenal CS Stemmer (*Confix Stripping Stemmer*) didasari oleh aturan morfologi Bahasa Indonesia secara komprehensif, yaitu mengelompokkan dan mengenkapsulasi

imbuan-imbuan baik yang diijinkan maupun yang dilarang, meliputi awalan (*prefix*), akhiran (*suffix*), dan gabungan awalan dan akhiran (*confix*), disebut juga dengan istilah *circumfix*^[15].

Setelah dilakukan beberapa analisa dan percobaan kecil terhadap algoritma CS Stemmer, ditemukan beberapa contoh kata yang tidak dapat di-stemming, dari percobaan tersebut maka muncul algoritma baru untuk memperbaiki kegagalan dari CS Stemmer yang di berinama ECS Stemmer dengan merevisi beberapa aturan-aturan dari CS Stemmer^[16].

3.7 Vector Space Model

Vector Space Model (VSM) adalah metode untuk melihat tingkat kedekatan atau kesamaan (*similarity*) term dengan cara pembobotan term. Dokumen dipandang sebagai sebuah vektor yang memiliki magnitudo (jarak) dan direction (arah)^[17]. Pada Model VSM ini setiap dokumen di dalam database dan query pengguna direpresentasikan oleh suatu vektor multi-dimensi.

Dalam model ruang vektor, database dari semua dokumen direpresentasikan oleh term-document matrix (atau *term-frequency matrix*). Setiap sel pada matriks menunjukkan bobot dari sebuah term dalam suatu dokumen. Jika nilainya nol maka term tersebut tidak muncul pada dokumen itu. Proses penghitungan bobot melibatkan dua elemen yang saling melengkapi yaitu frekuensi term *i* dalam dokumen *j*, dan *inverse document frequency* dari term *i*. Term yang lebih sering muncul pada dokumen menjadi lebih penting karena dapat mengindikasikan topik dari dokumen. Frekuensi term *i* dalam dokumen *j* didefinisikan sebagai berikut:

$$tf_{ij} = \frac{f_{ij}}{\max_i(f_{ij})}$$

f_{ij} adalah jumlah kemunculan term *i* pada dokumen . Frekuensi tersebut dinormalisasi dengan frekuensi dari term yang sering muncul pada dokumen. *Inverse document frequency* digunakan untuk menunjukkan *discriminative power* dari term *i*. Secara umum term yang muncul di berbagai dokumen kurang mengindikasikan untuk topik tertentu. Rumus dari *inverse document frequency* didefinisikan sebagai berikut:

$$idf_i = \log_2\left(\frac{n}{df_i}\right)$$

df_i adalah frekuensi dokumen dari term i dan dapat diartikan juga sebagai jumlah dokumen yang mengandung term i . \log_2 digunakan untuk meredam efek relatif terhadap tf_{ij} . *Weight* (bobot) W_{ij} dihitung menggunakan pengukuran TF-IDF yang didefinisikan sebagai berikut:

$$W_{ij} = tf_{ij} \times idf_i$$

Bobot tertinggi diberikan kepada term yang muncul sering kali dalam dokumen j tetapi jarang dalam dokumen lain. Setelah penentuan bobot maka akan dilakukan pengukuran kemiripan teks, pengukuran teks yang populer adalah (*cosine similarity*). Metode ini mengukur nilai cosinus sudut antara dua vektor. Cosinus dari dua vector dapat diturunkan dengan menggunakan rumus *Euclidean dot product*.

Diberikan dua vector dengan atribut-atribut, A dan B, nilai *cosine similarity*, $\cos(\theta)$, dinyatakan menggunakan dot product dan magnitude sebagai:

Hasil dari kemiripan berkisar dari -1 bermakna berlawanan, 1 bermakna persis sama (mirip), 0 yang mengindikasikan *orthogonality* (*decorrelation*) dan diantara nilai 0 sampai 1 yang mengindikasikan antara mirip dengan tidak mirip. Untuk pencocokan teks (*text matching*), atribut dari vector A dan B umumnya term frequency vectors dari dokumen. *Cosine Similarity* dapat dilihat sebagai metode untuk menormalisasi panjang dokumen selama perbandingan.

Pada kasus *information retrieval*, *cosine similarity* dari dua dokumen memiliki rentang 0 sampai 1, karena *Term Frequencies* (TF-IDF Weight) tidak dapat bernilai negatif. Sudut antara dua *term frequency vector* tidak dapat lebih besar dari 90° . Diberikan vektor yang merepresentasikan query q , dokumen dandan t term yang diekstrak dari database, nilai cosinus antara q dan d didefinisikan sebagai berikut^[18]:

$$Similarity \left(\vec{d}_j, \vec{q} \right) = \frac{\vec{d}_j \cdot \vec{q}}{|\vec{d}_j| |\vec{q}|} = \frac{\sum_{i=1}^t (w_{ij} \cdot w_{iq})}{\sqrt{\sum_{i=0}^t w_{ij}^2 \cdot \sum_{i=0}^t w_{iq}^2}}$$

3.8 Rule Based

Rule based adalah metode yang digunakan untuk *Quesiton Answering System* untuk *scoring* dan membuktikan bawah kalimat berisi jawaban. Dalam algoritma berbasis aturan ini kalimat-kalimat akan diberi skor atau poin, di

mana kalimat dengan skor paling tinggi akan ditandai sebagai jawabannya.

Semua jenis pertanyaan akan dikelompokkan pada fungsi kata umum pencocokan (*word matching*) dimana akan menghitung jumlah kata dalam pertanyaan dan kalimat berdasarkan pertimbangan-pertimbangan. Menurut Rilof dan Thellen berikut kategori tingkat penilaian, *clue* (+3), *good_clue* (+4), *confident* (+6), *slam_dunk* (+20)^[19].

3.9 Evaluasi

Evaluasi adalah proses untuk melihat kesuksesan dari model yang diusulkan dalam hal ini adalah untuk mengetahui keakuratan (*Accuracy*) dari jawaban yang diberikan pada pengguna terhadap pertanyaan yang diberikan. Penghitungan akurasi menggunakan rumus relevansi yaitu :

$$akurasi = \frac{jawaban\ benar}{total\ pertanyaan} \times 100$$

4. HASIL DAN PEMBAHASAN

4.1 Similarity

Setelah semua proses dilakukan maka akan ditemukan *similarity* (kemiripan) antar dokumen dengan query pertanyaan seperti hasil berikut ini adalah hasil dokumen relevan dari salah satu query pertanyaan yang diujikan pada penelitian ini dari 10 pertanyaan yaitu “Dimana rumah Allah?”. Dari query tersebut maka akan diperoleh 5 kandidat dokumen (dokumen relevan) dari seluruh dokumen (286 dokumen ayat). Kandidat dokumen dapat dilihat pada tabel 1.

4.2 Rule Based Scoring

Kandidat dokumen pada tabel 1 akan diproses kembali dengan *Rule Based* untuk mencari dokumen terbaik (*top document*) dari 5 dokumen terpilih seperti yang terlihat pada tabel 1. Hasil dari *scoring Rule Based* dapat dilihat pada tabel 2.

4.3 Evaluasi

Proses terakhir dalam penelitian ini adalah evaluasi, yang mana metode yang diusulkan akan dievaluasi tingkat keakuratannya. Berikut hasil dari beberapa testing yang dilakukan menggunakan 10 pertanyaan pada setiap tipe pertanyaan seperti terlihat pada tabel 3.

Untuk menghitung akurasi dari metode yang diusulkan menggunakan rumus (5). Berikut hasil rata-rata akurasi dari jawaban.

$$\text{Akurasi keseluruhan} = \frac{22 \text{ jawaban benar}}{30 \text{ Testing pertanyaan}} \times 100 = 73\%$$

Tabel 1 Dokumen yang Relevan

Rank#	Ayat	Similarity (%)
1.	Dan (ingatlah), ketika Kami menjadikan rumah itu (Baitullah) tempat berkumpul bagi manusia dan tempat yang aman. Dan jadikanlah sebagian maqam Ibrahim tempat salat. Dan telah Kami perintahkan kepada Ibrahim dan Ismail: "Bersihkanlah rumah-Ku untuk orang-orang yang thawaf, yang itikaf, yang ruku dan yang sujud". (ID Doc: 125)	41.00%
2.	Mereka bertanya kepadamu tentang bulan sabit. Katakanlah: "Bulan sabit itu adalah tanda-tanda waktu bagi manusia dan (bagi ibadat) haji; Dan bukanlah kebajikan memasuki rumah-rumah dari belakangnya, akan tetapi kebajikan itu ialah kebajikan orang yang bertakwa. Dan masuklah ke rumah-rumah itu dari pintu-pintunya; dan bertakwalah kepada Allah agar kamu beruntung. (ID Doc: 189)	37.29%
3.	Dan orang-orang yang akan meninggal dunia di antara kamu dan meninggalkan istri, hendaklah berwasiat untuk istri-istrinya, (yaitu) diberi nafkah hingga setahun lamanya dan tidak disuruh pindah (dari rumahnya). Akan tetapi jika mereka pindah (sendiri), maka tidak ada dosa bagimu (wali atau waris dari yang meninggal) membiarkan mereka berbuat yang makruf terhadap diri mereka. Dan Allah Maha Perkasa lagi Maha Bijaksana. (ID Doc: 240)	21.48%
4.	Kepunyaan Allahlah segala apa yang ada di langit dan apa yang ada di bumi. Dan jika kamu melahirkan apa yang ada di dalam hatimu atau kamu menyembunyikan, niscaya Allah akan membuat perhitungan dengan kamu tentang perbuatanmu itu. Maka Allah mengampuni siapa yang dikehendaki-Nya dan menyiksa siapa yang dikehendaki-Nya; dan Allah Maha Kuasa atas segala sesuatu. (ID Doc: 284)	1.25%

Rank#	Ayat	Similarity (%)
5.	Dan berperanglah kamu sekalian di jalan Allah, dan ketahuilah sesungguhnya Allah Maha Mendengar lagi Maha Mengetahui. (ID Doc: 244)	1.23%

Tabel 2 Hasil Top Document

Rank#	Ayat	Skor
1.	Dan (ingatlah), ketika Kami menjadikan rumah itu (Baitullah) tempat berkumpul bagi manusia dan tempat yang aman. Dan jadikanlah sebagian maqam Ibrahim tempat salat. Dan telah Kami perintahkan kepada Ibrahim dan Ismail: "Bersihkanlah rumah-Ku untuk orang-orang yang thawaf, yang itikaf, yang ruku dan yang sujud". (ID Doc: 125)	10.41

Tabel 3. Daftar Query Pertanyaan

No	Q = Siapa
1	Siapa yang mengetahui rahasia langit dan bumi?
2	Siapa penghuni neraka?
3	Siapa penghuni surga?
4	Punya siapa langit dan bumi?
5	Siapa pencipta langit dan bumi?
6	Kepada siapa Allah menjelaskan tanda-tanda kekuasaanNya?
7	Siapa yang termasuk orang-orang merugi?
8	Siapa yang diperintahkan Allah membersihkan rumahNya?
9	Siapa yang termasuk orang zalim?
10	Siapa yang mendapat laknat Allah, Malaikat, dan manusia?
No	Q = Kapan
1	Sampai kapan orang beriman menjadi penghuni surga?
2	Kapan Sulaiman dikatakan mengerjakan sihir?
3	Kapan Allah akan mengadili orang Yahudi atau Nasrani?
4	Kapan Allah akan mengumpulkan kamu sekalian?
5	Kapan dihalalkan bagi kami bercampur dengan istri kami?
6	Kapan berlaku hukum qishaash?
7	Sampai kapan ibu menyusui anaknya?

8	Sampai kapan orang meninggal harus menafkai istrinya?
9	Kapan suatu pohon berubah?
10	Kapan waktu dilakukannya ibadah haji?
No	Q = Dimana
1	Di mana malaikat Harut dan Marut?
2	Di mana tempat syi'ar Allah?
3	Di mana tempat orang beriman?
4	Di mana tempat orang kafir?
5	Di mana sihir diajarkan?
6	Di mana tempat khusus di sisi Allah?
7	Di mana Allah mengambil janji dari Bani Israel?
8	Di mana tempat haji atau sa'i?
9	Di mana dihasilkan buah dua kali lipat?
10	Dimana rumah Allah?

Dari hasil evaluasi yang sudah dilakukan pada tabel 4 di peroleh akurasi pada setiap tipe kategori pertanyaan “Siapa”, “Kapan”, dan “Dimana” dengan 10 macam query pertanyaan, masing-masing 90%, 80%, dan 50%, hal ini membuktikan akurasi yang diperoleh meningkat dari akurasi pada penelitian sebelumnya dengan masing-masing akurasi 60%, 60%, dan 40% [7].

Tabel 4 Hasil Evaluasi

NO	SIAPA		KAPAN		DIMANA	
	B	S	B	S	B	S
1		1	1		1	
2	1		1			1
3	1		1			1
4	1		1			1
5	1		1		1	
6	1		1		1	
7	1		1			1
8	1		1			1
9	1			1	1	
10	1			1	1	
Persentase	90%	10%	80%	20%	50%	50%
Total	22 Benar / 30 pertanyaan * 100 = 73%					

5. KESIMPULAN

Dalam penelitian ini dilakukan percobaan dengan beberapa pertanyaan dari semua kategori pertanyaan yaitu “Siapa”, “Kapan”, dan “Dimana”. Dari hasil percobaan yang telah dibahas pada bab sebelumnya dapat diambil kesimpulan bahwa *Rule Based* cukup baik dan dapat digunakan untuk *QAS*.

Dari hasil evaluasi, dengan menambahkan NER dan tidak menggunakan lucene sebagai

pencari kandidat jawaban terbukti dapat membantu meningkatkan akurasi *Rule Based* menjadi lebih baik. Akurasi yang diperoleh lebih baik dari penelitian sebelumnya dari semua kategori pertanyaan. Pada penelitian sebelumnya pada pertanyaan “Siapa”, “Kapan”, “Dimana” masing-masing diperoleh akurasi 60%, 60%, dan 40%, sedangkan dalam penelitian ini masing-masing diperoleh akurasi sebesar 90%, 80%, 50%.

Saran pengembangan untuk lebih meningkatkan kinerja Sistem QAS ayat Al Quran dengan menambahkan Analisa Semantik antar ayat maupun kata agar jawaban yang di kembalikan terhadap pengguna lebih baik dan akurat, penggunaan *Part Of Speech Tagging* (POS Tagging) untuk membantu ekstraksi jawaban lebih baik, dan melakukan analisa lebih mendalam tentang *Neutral Language Processing* (NLP) khususnya untuk NER, ekstraksi jawaban, analisa *pragmatic*, dan *discours*.

6. REFERENSI

- [1] Anggraeny, Meynar Dwi. 2007. “Implementasi Question Answering System Dengan Metode Rule-Based Pada Terjemahan Al Qur’an Surat Al Baqarah”. Skripsi Program Studi Ilmu Komputer IPB Bogor.
- [2] Hammo, Hammo, dkk. 2002. “QARAB: A question answering system to support the Arabic language”. ACL 2002 Computational approaches to semitic languages, pp. 11.
- [3] Derici, Caner, dkk. 2013. “Türkçe Soru Cevaplama Sistemlerinde Kural Tabanlı Odak Çıkarımı Rule-Based Focus Extraction in Turkish Question Answering Systems”. Department of Computer Engineering. Boğaziçi University, 34342 Bebek, Istanbul, Turkey
- [4] Buscaldi, Davide. 2007. “N-gram vs. Keyword-Based Passage Retrieval for Question Answering”. in Evaluation of Multilingual and Multi-modal Information Retrieval. Springer-Verlag Berlin Heidelberg. pp. 377-384.
- [5] Homaidi, Ahmad. 2015. “Question and Answering System pada Buku Transformasi Pesantren Sukorejo dari Hutan Menjadi Pusat Pendidikan Menggunakan N-Gram Passage Retrieval”. Tesis Program Studi

- Teknologi Informasi STTS Surabaya.
- [6] Jovita, dkk. 2015. "Using Vector Space Model in Question Answering System," *Procedia Computer Science*, Vol. 59, No. Iccsci, pp. 305-311.
- [7] Gusmita, Ria Hari, dkk. 2014. "A rule-based question answering system on relevant documents of Indonesian Quran Translation". *International Conference on Cyber and IT Service Management, CITSM 2014*.
- [8] Mollá, Diego, dkk. 2006. "Named Entity Recognition for Question Answering". *Proceedings of ALTW*, p. 51-58.
- [9] Chen, Hsi-Hsi, 1998. "Named Entity Extraction for Information Retrieval". *Computer Processing Of Oriental Languages*, Vol. 12, No. 4, pp. 75-85.
- [10] Liu, Ye and Ren, Fuji. 2011. "Japanese Named Entity Recognition for Question Answering System". *IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)*.
- [11] Akour, Mohammed. 2011. "QArabPro: A Rule Based Question Answering System For Reading Comprehension Tests in Arabic". *American Journal of Applied Sciences*, Vol. 8, No. 6, pp. 652-661.
- [12] Departemen Agama, 1989. *Al Qur'an dan Terjemahannya*, Semarang: Toha Putra.
- [13] Grishman, Ralph. 1997. "Information Extraction : Techniques and Challenges". *Information Extraction A Multidisciplinary Approach to an Emerging Information Technology*, pp. 10-27, 1997.
- [14] Chinchor, Nancy and P. Robinson, 1997. "MUC-7 Named Entity Task Definition". *Proceedings of the Sixth Message Understanding Conference MUC6*.
- [15] Asian, Jelita. 2007. "Effective Techniques for Indonesian Text Retrieval". *School of Computer Science and Information Technology, Science, Engineering, and Technology Portfolio, RMIT University: Australia*.
- [16] Arifin, Agus Zainal. 2009. "Enhanced Confix Stripping Stemmer And Ants Algorithm For Classifying News Document In Indonesian Language". *International Conference on Information & Communication Technology and Systems*, Vol. 5, pp. 149-158.
- [17] Amin, Fatkhul. 2012. "Sistem Temu Kembali Informasi dengan Metode Vector Space Model". *Jurnal Sistem Informasi Bisnis* Vol. 18, No. 2. pp. 122-129.
- [18] Cios, J.K. dkk. 2007. *Data Mining A Knowledge Discovery Approach*, New York: Springer US.
- [19] Rillof, Ellen and Michael Thelen. 2000. "A Rule-Based Question Answering System For Reading Comprehension Tests". *ANLP/NAACL 2000 Workshop on Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems*. Vol. 6, pp. 13-19.