Semantically Annotated Corpus Model of Indonesian Translation of Quran: an Effort in Increasing Question Answering System Performance

Husni Teja Sukmana
husniteja@uinjkt.ac.id
Dept. Informatics Engineering

Ria Hari Gusminta
ria.gusmita@uinjkt.ac.id
Dept. Informatics Engineering

Yusuf Durachman
yusuf_durachman@uinjkt.ac.id
Dept. Information Systems

Asep Fajar Firmansyah
asep.airlangga@uinjkt.ac.id
Dept. Information Systems

Syarif Hidayatullah State Islamic University
Jakarta, Indonesia

Abstract. **This paper presents our work in defining a model to build a semantic-based corpus of Indonesian Translation of the Quran. This task was being an effort to increase Question Answering System (QAS) performance on Indonesian Translation of the Quran that has been developed in 2013[1]. As the QAS run on three kinds of question type i.e. Siapa (Who), Kapan (When), and Di mana (Where), this model designed specifically to deal with those question types. The model is ready to be implemented and evaluated. Since we want to measure how significance semantic approach is on QAS performance, evaluation will be conducted by using the same scheme used in [1]. Furthermore, our model also will be assessed by comparing it with other corpus developed by using graph database model.**

Keywords—Indonesian Translation Quran; semantic-based corpus; Question Answering

I. INTRODUCTION

In Indonesia, moslem is the biggest population. Data from Indonesian National Statistical Bureau showed that on 2010, there are 207,176,162 moslem population compared to 237,641,326 the total Indonesian population [4]. In other word, percentage of moslem population against total number of Indonesian people is about 87.1%.

Some research purposed in providing Islam-related information for moslem population in Indonesia have been conducted which one of them started in 2012 [2]. This research developed a Question Answering System on Khulafaur Rasyidin History (termed as QASKR) written in Indonesian Language. [1] was trying to cleared up lacks on QASKR by combining standard architecture on [2] and Rule-based method that successfully applied (performance percentage above 80%) on [5], and then employed the new architecture on Indonesian Translation of Quran (termed as QASIQ). Surprisingly, it's result showed the unexpected achievement where it only accomplished on 53.3 %. One that remains on these research was they used corpus without linguistics information and it contributed on low performances. This following figures out how was that corpus take a part on the research.

```
<PERSON>Umar</PERSON> juga merupakan sahabat
<PERSON>Rasulullah Saw</PERSON>. yang terkemuka dan
seorang yang paling zuhud terhadap dunia. Telah
diriwayatkan darinya sebanyak 539 hadis.  Beberapa
orang yang meriwayatkan hadis darinya ialah
<PERSON>Usman bin Affan</PERSON>, <PERSON>Ali bin
Abu Thalib</PERSON>, <PERSON>Thalhah bin
Ubaidillah</PERSON>, <PERSON>Sa'ad bin Abi
Waqqash</PERSON>, <PERSON>Abdurrahman bin
Auf</PERSON>, <PERSON>Ibnu Mas'ud</PERSON>,
<PERSON>Abu Dzar</PERSON>, <PERSON>Amr bin
Abasah</PERSON> dan anaknya
<PERSON>Abdullah</PERSON>, <PERSON>Ibnu
Abbas</PERSON>, <PERSON>Abdullah bin
Zubair</PERSON>, <PERSON>Anas bin Malik</PERSON>,
<PERSON>Abu Hurairah</PERSON>, <PERSON>Amr bin
Ash</PERSON>, <PERSON>Abu Musa Al-Asy'ari</PERSON>,
<PERSON>Barra' bin Azib</PERSON>, <PERSON>Abu Said
Al-Khudri</PERSON>, dan masih banyak lagi sahabat
lainnya.
```

Above document was delivered by document retrieval component on QASKR for question "*Siapakah yang meriwayatkan hadis dari Ali bin Abu Thalib?*" The document only equipped by named entity tag such as PERSON, TIME, and LOCATION. Algorithm used on answer extraction component was calculate distance of question keyword "*meriwayatkan*" and word with PERSON tag, and choose word with shortest distance as the answer. In this case, system delivered Usman bin Affan which was the incorrect answer. We are totally agree that this document talks about Umar instead of Ali bin Abu Thalib.

Another example come from QASIQ where document with highest number of term with answer type tag existence was pointed as selected documents to be used in answer extraction. These documents shows the case:

```
Allah berfirman: "Hai Adam, beritahukanlah kepada mereka
nama-nama benda ini". Maka setelah diberitahukannya
kepada mereka nama-nama benda itu, Allah berfirman:
"Bukankah sudah Ku-katakan kepadamu, bahwa sesungguhnya
Aku mengetahui rahasia langit dan bumi dan mengetahui apa
yang kamu lahirkan dan apa yang kamu sembunyikan?"
```

```
Ayat 284 - Kepunyaan Allah-lah 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
```

Both above documents returned for question "*Siapa yang memiliki langit dan bumi?*" Since the question type is "*Siapa*", then the answer type is PERSON. First document has higher number of term with PERSON tag than second number, so that system provided answer from wrong document.

In order to settle such above problems, the corpus is necessary to be annotated with semantics. By enriching the corpus with semantics, system will able to know meaning of the sentence and so that utilize it to extract correct answer from correct document.

This paper presents a work in construct a model to build semantically annotated corpus of Indonesian Translation of Quran. As the corpus was purposed to increase existing Question Answering System performance on Indonesian Translation of Quran, the model was deal with three question types i.e. Siapa (Who), Kapan (When), and Di mana (Where). Additionally, the model was designed to enrich the corpus with other linguistics information such as morphology analysis and part of speech.

This paper is organized as follows: the next section presents related work; Section 3 provides the methodology and proposed model; Section 4 presents the ongoing implementation of this research; Section 5 conclude this paper with a summary and future work.

II. RELATED WORK

There are several works on building a semantically annotated corpus aimed at various system on information retrieval or natural language processing fields.

The need of development or adaptation of text mining (TM) tools on Narrative information in Electronic Health Records (EHRs) and literature articles and , led [5] to create a new annotated corpus (PhenoCHF), focusing on the identification of phenotype information for a specific clinical sub-domain, i.e., congestive heart failure (CHF). The corpus is unique in this domain, in its integration of information from both EHRs (300 discharge summaries) and literature articles (5 full-text papers). The annotation scheme, whose design was guided by a domain expert, includes both entities and relations pertinent to CHF. Two further domain experts performed the annotation, resulting in high quality annotation, with agreement rates up to 0.92 F-Score [5].

In 2014, [6] has designed and developed TEMIS, a syntactically and semantically annotated corpus of Italian Legislative Texts. The whole corpus has been dependency annotated and a subset has been enriched with frame-based information by customizing the formalism of FrameNet project. In both cases, a number of domain-specific extensions of the annotation criteria developed for the general language has been foreseen [6].

Ontology usage was also take a part in semantic annotation research field. [7] Focused on creating a framework of semantic annotation on Urdu language web documents base on ontology. Urdu language exist in an unstructured format, and this was being a background of this research. The framework uses domain specific ontology and context keywords instead of NLP (Natural Language processing) techniques. The experiment has been conducted to evaluate the presented annotation framework. The set of corpora used in the experiment belong to the online classified ads posted on the online Urdu newspapers [7].

Islamist extremist has been a topic chosen by [8] so that it constructed a semantically annotated collection of Islamist extremist stories. This corpus is called as N2 (Narrative Networks) corpus. N2 is unique represented by its three profiles. First, every text in the corpus is a story, which is in contrast to other language resources that may contain stories or story-like texts, but are not specifically curated to contain only stories. Second, the unifying theme of the corpus is material relevant to Islamist Extremists, having been produced by or often referenced by them. Third, every text in the corpus has been annotated for 14 layers of syntax and semantics, including: referring expressions and co-reference; events, time expressions, and temporal relationships; semantic roles; and word senses [8].

III. ANNOTATION LAYER

In order to let the system able to interpret information in the verse of Indonesian Translation of Quran, annotation is done in several layer [8] as listed below:

1. Syntax
   This layer was purposed for two kinds of linguistics information annotation such as morphology and part of speech analysis. Object at this layer is sentence.
2. Referential structure
   In Indonesian Translation of Quran, there are many terms act as pronoun. To deal with this fact, annotation was also applied on referential structure. From this structure, system will able to gather value that referred by particular pronoun.
3. Semantics
   Semantics layer is designed to accommodate three question types implemented on existing Question Answering System i.e. Siapa (WHO), Kapan (When), and Di mana (Where).

3.1. Syntax Annotation

We applied syntax annotation on each verse of Indonesian Translation of Quran. This data was supposed to be taken from official version of Indonesian Translation of Quran from the Ministry of Religious Affairs of the Republic of Indonesia. Unfortunately, we could not find it in digital form. Finally, we got a digital version of Indonesian Translation of Quran from a site that is *http://tanzil.net/trans/*. It contains the translation of Quran in various languages, including Indonesian. Especially for Indonesian Translation, there are several versions of it as follows: Ministry of Religion Affair of the Republic of Indonesia, Muhammad Quraish Shibab, and Jalal ad-Din al-Mahalli and Jalal ad-Din as-Suyuti. We took the one from the first version. This is Indonesian Translation of Quran that we have taken:

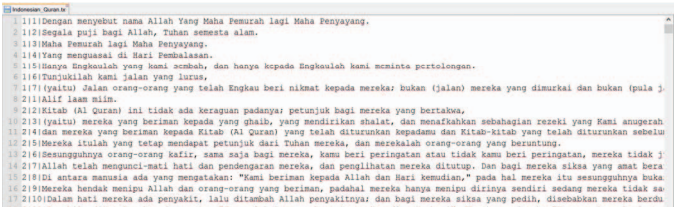

Figure 1. Display of Indonesian Translation of Quran Text

We applied text preprocessing on all verse of Indonesian Translation of Quran, started from morphology analysis and part of speech analysis. Both process are done by utilize existing library of each of process. Morphology analysis was conducted by utilize MorphInd [9]. Analysis of part of speech was run by using an Indonesian Part of Speech tagger service created by Faculty of Computer Science, University of Indonesia [10]. These following figures shows result of morphology and part of speech analysis on syntax layer.

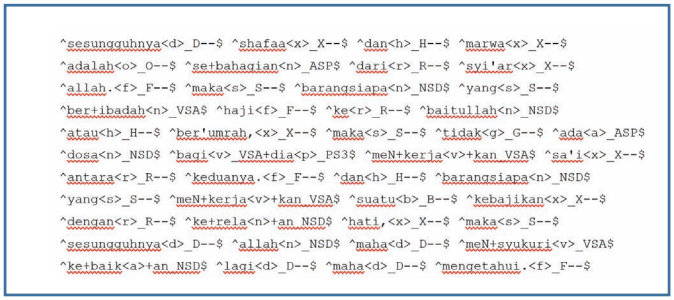

Figure 2. Result of Morphology Analysis

Figure 3. Result of Part of Speech Analysis

3.2. Referential Structure Annotation

As we need to have list of pronouns in the Indonesian Translation of Quran to construct referential structure annotation layer, there was a preprocessing done aiming to get term that has named entity PERSON. This could be accomplished by using ontology of Indonesian Translation of Quran. Since we could not find existing ontology of Indonesian Translation of Quran, we developed it by reusing other existing ontology i.e. Arabic Quran ontology [11]. Table 1 gives list of sample of Indonesian Translation of Quran concept that extracted from Arabic Quran ontology:

Table I
List of Sample of Indonesian Translation of Quran Concept

Aad	Allat	Al Uzza
Aazar	Anak-anak Adam	Allah
Abu Lahab	Anak-anak Israel	Ayyub
Adam	Anak Lembu Emas	Baal
Agama	Anggur	Babi
Ahmad	Anjing	Babilonia
Al Ahqaf	Ansor	Badar
Al Judi	Ara	Baduy
Al Marwah	Artefak	Bagal
Al Quran	Ashabul Kahfi	Bagian Tubuh
	Awan	Bahasa

After got the list of Quran concept, we select those that represent Person as the sample of them is listed as follows:

Table II
List of Sample Quran Concept

Aad	Harut	Luqman
Aazar	Hud	Luth
Abu Lahab	Iblis	Malaikat
Adam	Ibrahim	Manusia
Ahmad	Idris	Marut
Allah	Ilyas	Maryam
Anak-anak Adam	Ilyasa	Mesias
Anak-anak	Imran	Mikail
Israel	Isa	Muhammad
Ayyub	Ishak	Musa
Firaun	Ismail	Nuh
Habil	Izrail	Orang Romawi
Haman	Jalut	Para Penggali Parit
Harun	Jibril	Penduduk Aikah
	Jin	

In order to conducted annotation of referential structure, we find location of existence of each concept that represent person (called as person concept) in the Indonesian Translation of Quran. Furthermore, we find pronouns that refer to particular person concept and created a dictionary to record triple of pronoun, location, and person concept.

Table III
Sample of Referential Structure Dictionary

mereka***chapter_29_verse_38***Aad
kamu***chapter_6_verse_74***Aazar
dia***chapter_111_verse_1***Abu Lahab
kamu***chapter_2_verse_35***Adam
seorang Rasul***chapter_61_verse_6***Ahmad
Aku***chapter_58_verse_21***Allah
mereka***chapter_17_verse_70***Anak-anak Adam

3.3. Semantics Annotation

Semantics annotation in this research is conducted with special purposed i.e. to increase performance of Question Answering System (QAS) that has been developed before [1]. By reason of it, semantics model was designed in line with terminology of three kinds of question type on that QAS. Those question type are Siapa (Who), Kapan (When), and Di mana (Where).

We defined two main component of semantics model and annotation on the Indonesian Translation of Quran i.e. entity classes and entity relationship. The model is described as follows:

- [Person] does what, [Person] is what
- Something happens on [Time]
- Something happens at [Location]

As Indonesian text syntax can be vary in term of active and passive sentence, semantics model also accommodate appearance of entity class in a reverse order.

To fulfill resource of semantics model, we provided terms from entity name Time and Location (entity name Person terms came from referential structure annotation task). Furthermore, to accomplish the annotation we find location of appearance of each term in the Indonesian Translation of Quran and implement the semantics model. Table IV shows sample of Quran terms

with entity name Time. It is followed by table V to display sample of Quran terms with entity name Location.

Table IV
Sample of Quran Terms with Entity Name Time

Hari Sabtu
Hari Kemudian
Bulan Haram
Seribu tahun kurang lima puluh tahun
Bulan Ramadhan
Beberapa hari yang ditentukan
Hari raya
Jumát
Malam Kemuliaan

Table V
Sample of Quran Terms with Entity Name Location

'Arafat
Masjidil Haram
Al Masjidil Aqsha
Al Masjidil Haram
Shafaa
Marwa
Babil
Badar
Bakkah
Hunain
Iram
Makkah
Madinah
Yatsrib
Al Ahqaaf
Mesir
Saba
Bukit Judi
Gunung Thursina

Sample of semantics annotation on the Indonesian Translation of Quran is depicted on Figure 4.

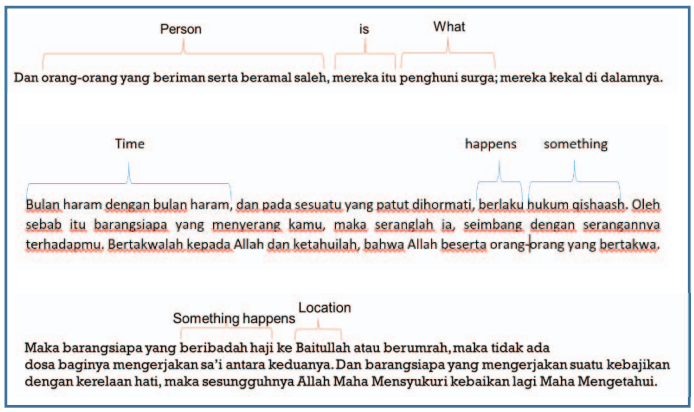

Figure 4. Sample of Semantics Annotation on the Indonesian Translation of Quran

IV. CONCLUSION

We have presented a model to do semantically annotation on the Indonesian Translation of Quran corpus. This model is ready for the implementation in order to evaluate its performance in term of increasing Question Answering System performance. We hope there will be a valuable information we contribute after evaluate the corpus regarding the effectiveness of semantic approach on the corpus.

REFERENCES

[1] Ria Hari Gusmita, Yusuf Durachman, Asep Fajar Firmansyah I.A., Husni Teja Sukmana, Adam Suhaemi, "A Rule-based Question Answering System on Relevant Documents of Indonesian Quran Translation", in Proceedings of The 3nd International Conference on Cyber & IT Service Management, IEEE Xplore, 3 – 6 November, 2014, Jakarta, Indonesia

[2] Zidny Naf'an and Ria Hari Gusmita, "Development of an Indonesian Question Answering System about Khulafaur Rasyidin's History", in Proceedings of The 1st International Conference on Cyber & IT Service Management in Conjunction with the ITIL v.3 Workshop, Training and Certification, 9 – 11 November, 2012, Bandung, Indonesia

[3] Dwi Anggraeni Meynar, Implementasi Question Answering Sistem dengan Metode Rule-Based padaTerjemahan Al-Qur'an Surat Al-Baqarah, Skripsi Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Pertanian Bogor, 2007

[4] http://www.bps.go.id/

[5] Noha Alnazzawi, Paul Thompson and Sophia Ananiadou, "Building a semantically annotated corpus for congestive heart and renal failure from clinical records and the literature", in Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi), April 26-30 2014, Gothenburg, Sweden

[6] Giulia Venturi, "Design and Development of TEMIS: a Syntactically and Semantically Annotated Corpus of Italian Legislative Texts", in Proceedings of Semantic Processing of Legal Texts Workshop, 2012

[7] Quratulain Rajput, "Ontology based semantic annotation of Urdu language web documents", in Proceedings of International Conference on Knowledge-Based and Intelligent Information & Engineering Systems, September 15 – 17, 2014, Gdynia, Poland

[8] Mark A. Finlayson, Jeffry R. Halverson, Steven R. Corman, "The N2 corpus: A semantically annotated collection of Islamist extremist stories", in Proceedings of Language Resources and Evaluation Conference, May 26 – 31, 2014, Iceland.

[9] Septina Dian Larasati, Vladislav Kuboň, and Daniel Zeman, "Indonesian Morphology Tool (MorphInd): Towards an Indonesian Corpus", in proceedings of proceedings of the Workshop on Systems and Frameworks for Computational Morphology, August 2011, Zurich, Switzerland

[10] http://fws.cs.ui.ac.id/RESTFulWSStanfordPOSTagger/

[11] http://corpus.quran.com/ontology.jsp

[12] Alwi, H, Soenjono Dardjowidjojo, Hans Lapoliwa, Anton M. Moeliono, Tata Bahasa Baku Bahasa Indonesia, PT Balai Pustaka, Departemen Pendidikan dan Kebudayaan Republik Indonesia, Jakarta, 1998