# Generating Weighted Vector for Concepts in Indonesian Translation of Quran

Syopiansyah Jaya Putra[1], Khodijah Hulliyah[2], Nashrul Hakiem[2], Rayi Pradono Iswara[2]
Asep Fajar Firmansyah[1]
[1]Information Systems [2]Informatics Engineering
UIN SyarifHidayatullah Jakarta, Indonesia
{syopian, khodijah.hulliyah, hakiem, rayi.iswara, asep.airlangga}@uinjkt.ac.id

## ABSTRACT
This paper presents a work in generating Weighted Vector for each Concept in Indonesian Translation of Quran (ITQ). This task is done in aiming to provide a resource needed in implementing a semantic-based question answering system (QAS) for Indonesian ITQ, particularly in retrieving semantically related verses. Semantic approach on QAS employs Ontology concepts of the domain. Since there is no Ontology for ITQ remains, we built one by utilizing the existing Ontology from Quranic Arabic corpus (http://corpus.quran.com/). Furthermore, each leaf concept that enriched by related Quran verse (as its instance) had a representation vector of terms that occur in the corresponding Quran verse to express how strength the concept in relates with verse terms. This vector is assigned with a weight resulted from applying TFIDF method. From 222 leaf concepts in the Ontology, we applied the process only to those that categorized as a member group of Person, Location, and Time named entity. They are 107 in a total. The result shows that the most strength concept in association with verse terms is syaitan which is scored at 0.895 of 1. In overall, 16.82 % concepts had score that more than 0.4, following by 14.95%, 23.36% and 11.21% concepts scored at more than 0.3 ,0.2 and less than 0.1 respectively, and finally the rest ones were the biggest in volume where 33.64% concepts obtained score more than 0.1 and less than 0.2.

## CCS Concepts
**• Information systems → Information retrieval → Document representation→Content analysis and feature selection**.

## Keywords
Weighted Vector; Concepts; Indonesian Translation of Quran.

## 1. INTRODUCTION
Muslims have a guidance of life that is Quran and Hadith. Allah revealed the Quran by using the Arabic language. The challenges in translating the verses of the Quran into Indonesian are in one word can have several meanings, and it leads to different understandings. Therefore, need to be built ITQ Ontology.


*iiWAS '16, November 28-30, 2016, Singapore, Singapore*

DOI: http://dx.doi.org/10.1145/3011141.3011218

The existence of ITQ Ontology is very useful in conducting research, especially in Question Answering System (QAS). QAS is the process of identifying exact answer in responding to a question asked by the user [1]. The working of QAS is looking for the closest relationship between the question word and answered words.

Since there is no Ontology for ITQ remains, we built one by utilizing the existing Ontology from Quranic Arabic Corpus (http://corpus.quran.com/). Furthermore, each leaf concept that enriched by related Quran verse (as its instance) had a representation vector of terms that occur in the corresponding Quran verse to express how strength the concept in relates with verse terms.

However, the most crucial thing when building Ontology is to define the concepts and its derivatives. To see the relationship between words or sentences, appeared various techniques is used to perform weighting vector in finding the nearest value between one word and other words. These techniques often used by search engines as a tool in scoring and ranking a document's relevance given by user query.

Two types of term weighting schemes: Unsupervised Term Weighting Schemes (UTWS) and Supervised Term Weighting Schemes (STWS) [2]. TF-IDF, a short for Term Frequency–Inverse Document Frequency is one of some methods that it is often used as a weighted factor in UTWS [3]. Moreover, TFIDF has been used in order to provide the main concepts of the domain for Ontology construction.

This research proposes a work in generating weighted vector for each concept in ITQ. To improve the performance of previous question answer (QA) for ITQ in [4][5], we adopted a semantic approach to that QA system. This task is needed to provide a property to retrieve semantically related verses on ITQ.

## 2. RELATED WORKS
In that respect, [6] presented research a novel method using vector evaluation for Arabic text categorization. The proposed method uses a categorized Arabic documents corpus, and then the weights of the tested documents' words are calculated to determine the document keyword which will be compared with the keywords of the corpus categorized to determine the tested document's best category. The proposed method determines the keywords of the tested document by weighting each of its words and then comparing these keywords with the keywords of the testing Corpus categorizes.

In [1], presented that this study aims to propose an Ontology-based question answering approach for the domain of Islamic Fatwa. The Ontology has been constructed using a collection of

Fatwas, which has been collected from Ibn Uthaymeen-Prayer Fatwas. Several pre-processing tasks have been applied in order to eliminate the irrelevant data (e.g. numbers, non-Arabic letters, and punctuation). Furthermore, Term Frequency-Inverse Document Frequency (TFIDF) has been used in order to provide the main concepts of the domain for Ontology construction.

In conjunction with that, [7] did comparison three methods of Weighted Vector, TFIDF term weighting, the vector space model and the cosine similarity measure for relevance score calculation. He did research with several questions; to assign weights marking the importance of these tokens for the documents, the mechanisms determine which documents are retrieved and the relevance score calculated that finally determines the ranking.

# 3. METHODOLOGY

In order to generate Weighted Vector for each Concept in ITQ, there are several resources needed as depicted in the following figure:

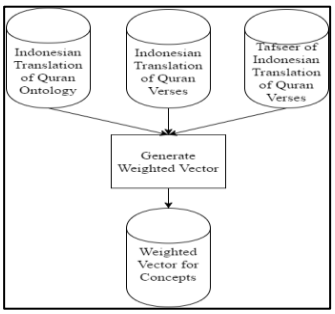


**Figure 1. Architecture of Generating Weighted Vector for Concepts in TIQ**

As seen in Figure 1, there are three resources needed to conduct a process to generate Weighted Vector for Concepts in ITQ. They are ITQ Ontology, ITQ Verses, and Tafseer of ITQ Verses.

From our best knowledge, there is no Ontology exist for ITQ. To deal with this, we built our own Ontology for ITQ. This Ontology construction utilized the one that has already developed that being a part of Quranic Arabic Corpus [8]. As the Quranic Concepts are written in English, we did an adjustment on them that appear inappropriately in ITQ.

Since there is no digital version of the last two resources, we gathered them from an International Quranic project output which has a digital translation of the Quran in many languages in the world[9]. We selected the one that translated by Ministry of Religious Affairs, Republic of Indonesia and Muhammad Quraish Shihab et al for ITQ Verses and Tafseer of ITQ Verses respectively.

Generating the Weighted Vector for Concepts in ITQ Ontology was conducted by implementing Term Frequency (TF) Inverse Document Frequency (IDF) method. TF refers to a number of occurrences of particular term in a document. Meanwhile, IDF is used to represent how important a certain term in regard to its existence in all documents. TF and IDF is calculated based on this following equation [8];

```
TF = (number of times a term in the document occurs)/(total number of terms found in that document)
```

```
IDF = log(total number of documents/number of documents where term appear in)
```

# 4. WEIGHTED VECTOR FOR CONCEPTS IN INDONESIAN TRANSLATION OF QURAN ONTOLOGY

There are three resources needed in generating Weighted Vector for Concepts in ITQ Ontology. Those resources were collected in their particular way as explained in detail below.

## 4.1 Providing Indonesian Translation of Quran Ontology

Construction of ITQ Ontology was accomplished by utilizing an existing Quranic Ontology on Quranic Arabic Corpus [9]. This following figure depicts sample of Quranic Ontology in [9].

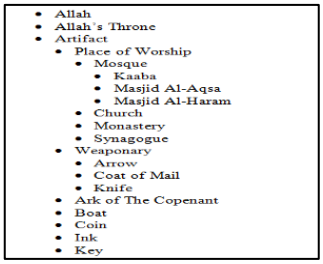


**Figure 2. Sample of Quranic Ontology from Quranic Arabic Corpus**

We applied all properties of Quranic Ontology in [9] including Concepts, Hierarchy, and Instances. We got 284 Concepts and 254 Individuals. Individual is assigned with verse location where the concept is explained in. We set a format to write Individual value as follows:

```
Surat_[chapter_number]_ayat_[verse_number]
```

In above format, Surat means Quran Chapter and ayat refers to Quran Verse.

Due to repetition applied in the Quran, there are 37 Individuals share their existence for several different Concepts. Furthermore, we had 200 unique Individuals in total. In case we found a repetition on Individual, we appended a number of times the Individual being used to Individual's name. This number is written after an underscore character. For instance, Individual "surat_71_ayat_23" has been determined as Individual for 5 different Concepts, and so they were written in this following way:

```
surat_71_ayat_23_1
surat_71_ayat_23_2
surat_71_ayat_23_3
surat_71_ayat_23_4
surat_71_ayat_23_5
```

Considering the Ontology was written in English, we did translation on all Concepts by using Bing Translator Application Programme Interface (API) so that they can be fit with text in ITQ. Unfortunately, the translation results did not meet our expectation where there were several Concepts appeared in incorrect form. This means that those terms are not the same with their referred-term in ITQ. There are two things contributed in this case i.e. lexical difference and missing a single quote character. We show some of the incorrect translation results due to they differ in lexical in Table 1.

**Table 1. Incorrect Translation Result Terms Due To Lexical Difference**

| Translation Result Term | Term in ITQ |
|---|---|
| Gunung Arafat | 'Arafat |
| Ahmad | Muhammad |
| Al judi | Bukit Judi |
| Babilonia | Kerajaan Sulaiman |
| Ara | Buah tin |

In Table 1, we can see that Ahmad was slightly different with Muhammad in that it was missing several letters. Gunung Arafat and Al judi came with another case where referred-term in ITQ was their term part. The worst case was found on Babilonia and Ara as they were totally different with Kerajaan Sulaiman and Buah tin in ITQ.

Another discover thing on translation result was there were several terms that incorrect due to missing a single quote character as listed in the following table.

**Table 2. Incorrect Translation Results Due to Single Quote Missing**

| Translation Result Term | Term in ITQ |
|---|---|
| Aad | 'Ad |
| Syuaib | Syu'aib |
| YajujMajuj | Ya'jujMa'juj |
| Yakub | Ya'qub |
| Yauq | Ya'uq |

For all incorrect translation results, we did manual correction by looking for their referred term in ITQ and replace them with the referred term. For those translation terms that do not have referred term in ITQ, we tackled them by applying manual interpretation on the verses they were explained in. Table 3 shows number of correct translation result terms and the incorrect ones.

**Table 3. Number of Correct and Incorrect Translation Results**

| No. Correct Translation Results | No. Incorrect Translation Results | |
|---|---|---|
| | Lexical Difference | Single Quote Missing |
| 194 | 80 | 10 |

We constructed ITQ Ontology by utilizing Protege, an open-source Ontology editor.

## 4.2 Providing Indonesian Quran Verses

Since we did not find any digital Indonesian Translation of Quran verses published by the authorized party, we gathered it from [9]. In [9], there are many translations of Quran in different languages including Indonesian. We chose the one that translated by Indonesian Ministry of Religious Affairs. This digital ITQ verses was written in this following format:

```
Chapter number|verse number|verse content
Display of some ITQ verses can be seen in Figure 5 below:
1|1|Dengan menyebut nama Allah Yang Maha Pemurah lagi Maha Penyayang.
1|2|Segala puji bagi Allah, Tuhan semesta alam.
1|3|Maha Pemurah lagi Maha Penyayang
```

**Figure 3. Display of Digital Indonesian Quranic Text**

## 4.3 Providing Tafseer of Indonesian Quranic Verses

Due to the same case with ITQ verses, we also got the Tafseer of ITQ verses from [9] as it provides the Tafseer as well. We selected Tafseer written by Muhammad Quraish Shihab et al. This Tafseer is structured as the same as structure of ITQ verses outlined below:

```
Chapter number|verse number|tafseer content
```

In this Tafseer of ITQ verses, every first verse completed by an introduction to the chapter. This introduction is enclosed by a couple of double square brackets. Tafseer of the first verse was given after introduction part. Figure 5 demonstrates how the Tafseer was presented. The underline sentences are the introduction of chapter 3 in the ITQ. Meanwhile, Tafseer of the first verse in that chapter was written in Italic mode.

## 4.4 Construct a Document for Each Leaf Concept in the Ontology

Before construct all document for each leaf Concept in the Ontology, we selected all leaf Concept in the Ontology of ITQ. Leaf Concept is the one that has no child class in Ontology. For each leaf concept, there is one instance of it that in this case refers to verse location it is explained in. Since the Weighted Vector is generated as one resource for semantic-based question answering system that accommodates three kinds of factoid question type i.e. Who, Where, and When, it is needed to classify the leaf Concepts based on answer type for each those question type that are known as Person, Location, and Time. The classification process was done by analyzing each Concept manually and determined whether it belongs to one of answer type or not. From the classification results, we gained 62, 19, and 9 Concepts for Person, Location, and Time type respectively. A sample of the result of the classification is given in the following tables.

**Table 4. Sample of Classification Result**

| Person | Location | Time |
|---|---|---|
| Aazar | 'Arafat | BulanRamadhan |
| Abu Lahab | Al Ahqaaf | HariKiamat |
| Adam | Al Masjidil Aqsha | HariSabtu |

Generate of document that relates to each concept in the Ontology was done by applying this following steps:

1.Open file that contains list of leaf Concept along with its related verse

2.For each line in the file:

- Extract each line to get concept name and verse location
- Open file namely with verse location
- Open file that contains Tafseer based on verse location
- Open file output to write verse and Tafseer content. This file is named with a combination of leaf concept name and Tafseer term as outlined below:
  - Concept name_tafsir
  - Get verse and Tafseer content from respective related file and write them into the output file

Each produced document contains verse and tafseer related with a particular concept. Both of them is written on a separate line in order to simplify detection of verse and tafseer in answer extraction phase. This following figure describes how both verse and tafseer are written in each document.

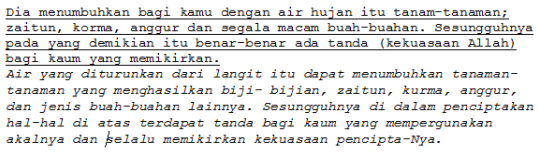


**Figure 4. Display of Verse and Tafseer in the Document**

In Figure 4, a sample of document content is given. The one written in underline mode is verse and the other with Italic format is tafseer.

## 4.5 Text Preprocessing on Each Document

To let the system able to manipulate text inside each document, a text preprocessing is applied in form of Tokenization, Stemming, and Part of Speech (POS) tagging. Tokenization is conducted by splitting a sentence on space character before applying stopwords removal. This following is sample of Tokenization result:

```
The verse is:
Atau kamu mengira bahwa orang-orang yang mendiami gua dan (yang mempunyai) raqim itu, mereka termasuk tanda-tanda kekuasaan Kami yang mengherankan?
List of Token:
 orang-orang        raqim              menghe
 mendiami           tanda-tanda        rankan
 gua                kekuasaan
```

**Figure 5. Result of Tokenization on Text Document**

In Figure 5, words with underline mode are removed before doing Tokenization. This is because they are stopwords in Indonesian. Afterward, tokenization is done on the rest of terms and resulted 7 tokens.

For both Stemming and POS tagging, we utilized an existing library created by Information Retrieval Lab in University of Indonesia i.e. Indonesian POS tagger [10] and Indonesian Stemmer [11] respectively.

```
The verse is:
Dan mereka selalu diikuti dengan kutukan di dunia ini dan (begitu pula) di hari kiamat. Ingatlah, sesungguhnya kaum 'Ad itu kafir kepada Tuhan mereka. Ingatlah kebinasaanlah bagi kaum 'Ad (yaitu) kaum Huud itu.
The stemming result is:
dan mereka selalu ikut dengan kutuk di dunia ini dan (begitu pula) di hari kiamat. ingat, sungguh kaum 'ad itu kafir kepada tuhan mereka. ingat binasa bagi kaum 'ad (yaitu) kaum huud itu.
```

**Figure 6. Sample of Stemming Result on Document**

Figure 6 presents Stemming process on a verse where words written in Italic mode are those that have affixes and so then Stemming is applied on them. As the result, those words exist in their stem called as a lemma. In another side, POS tagging is implemented to get word class on the sentence. We give the result of POS tagging in Figure 7 below.

```
The verse is:
Atau kamu mengira bahwa orang-orang yang mendiami gua dan (yang mempunyai) raqim itu, mereka termasuk tanda-tanda kekuasaan Kami yang mengherankan?
Some of POS Tagging result is:
<element id = '0'>
        <word>Atau</word>
        <postag>con</postag>
</element>
<element id = '1'>
        <word>kamu</word>
        <postag>pr</postag>
```

**Figure 7. Sample of POS Tagging Result**

Figure 7 depicts each word in the sentence has been analyzed and given related class that represented by postag tag. Word classes in the given example are con (conjucntion), pr (pronoun), vb (verb), and nn (noun).

## 4.6 Generate a Quranic Concept Weighted Vector by Implement TFIDF method

In order to generate Weighted Vector for Concepts in ITQ Ontology, we implemented TFIDF method as explained in [12][4][8]. Input of this process is a list of concept and related verses. Since this Weighted Vector will be used as a property in previous QAS for ITQ, then only Concepts from Person, Location, and Time type were taken. We have classified Concepts into those three type and resulted 90 Concepts. We also selected related verse for each Concept.

# 5. EXPERIMENT RESULT

As explained in previous part, we categorized the Concepts into three groups based on applied named entity (NE) i.e. Person, Location, and Time. The result of generation the Weighted Vector for Concepts in ITQ showed that the highest Weighted Vector score was 0.895 that belonged to concept syaitan. Inversely, concept harut, marut and kerajaan sulaiman got the lowest score for their Weighted Vector that is 0,0248. This following table lists Weighted Vector score in range for Concepts with named entity Person, Location, and Time.

**Table 5. Distribution of Weighted Vector for Concepts of Named Entity Person, Location and Time**

| Named entity Group | Range of Weighted Vector Score | Number of Concepts |
|---|---|---|
| Person | > 0.4 | 14 |
| | > 0.3 and < 0.4 | 11 |
| | > 0.2 and < 0.3 | 21 |
| | > 0.1 and < 0.2 | 26 |
| | < 0.1 | 5 |
| Location | > 0.4 | 2 |
| | > 0.3 and < 0.4 | 5 |
| | > 0.2 and < 0.3 | 3 |
| | > 0.1 and < 0.2 | 8 |
| | < 0.1 | 6 |
| Time | > 0.4 | 2 |
| | > 0.3 and < 0.4 | 0 |
| | > 0.2 and < 0.3 | 1 |
| | > 0.1 and < 0.2 | 2 |
| | < 0.1 | 1 |

Table 5 shows that distribution of Weighted Vector score of Concepts in ITQ for each named entity group has five range of score in common. They are (> 0.4), (>0.3 and <0.4), (> 0,2 and < 0,3), (> 0,1 and < 0,2) and (< 0.1). Person, location and time named entity Concepts were mostly expressed by their related documents in scale between 0.1 and 0.2. Person NE is a group with the highest number of Concepts that graded at the highest, 0.4. In another side, Location and Time NE stay the same for that kind of number. In regard with the lowest range, Concepts in Time NE shows the best achievement in that there is only one concept placed at the range. In contrast, even Person NE has a quite big number of Concepts in the highest range, it also remains with many Concepts in the lowest range compared with Location and Time NE. Interestingly, Concepts in Location NE are spread conversely on highest and lowest range and number of the Concepts. Each concept is not only completed by Weighted

Vector score but also with verse location where the concept exist in (called as Document). This is depicted as follows.

'aad***document_3055***0.399494349640794
anak-anak adam***document_989***0.107025317799908
'arsy***document_5924***0.419358750927329
berhada-berhala***document_1145***0.223229770971577
biji-bijian***document_5785***0.456120873598406

**Figure 8. Sample of Weighted Vector for Concepts list**

Document selected for each Concept denotes that the document represents the related Concept strongly compare with others documents that also has the Concept lexically inside. In another side, score gathered that acts as Weighted Vector describes how strong the document express the concept. This score is range from 0 and 1. We show how a selected document portrays a concept clearly than other documents that also contain the Concept.For Concept Penduduk Aikah, it exists in 4 verses in ITQ as follows:

**Table 6. All Verses that Contain Concept Penduduk Aikah**

| No. | Verse Location | Verse Content | |
|---|---|---|---|
| | | Indonesian | English |
| 1. | Chapter 15, verse 78 | Dan sesungguhnya adalah penduduk Aikah itu benar-benar kaum yang zalim, | And the companions of the thicket were [also] wrongdoers |
| 2 | Chapter 26, verse 176 | Penduduk Aikah telah mendustakan rasul-rasul; | The companions of the thicket denied the messengers |
| 3 | Chapter 38, verse 13 | dan Tsamud, kaum Luth dan penduduk Aikah. Mereka itulah golongan-golongan yang bersekutu (menentang rasul-rasul). | And [the tribe of] Thamud and the people of Lot and the companions of the thicket. Those are the companies. |
| 4. | Chapter 50, verse 14 | dan penduduk Aikah serta kaum Tubba' semuanya telah mendustakan rasul-rasul maka sudah semestinyalah mereka mendapat hukuman yang sudah diancamkan. | And the companions of the thicket and the people of Tubba'. All denied the messengers, so My threat was justly fulfilled. |

Based on the result of Weighted Vector generation, Document_1880 is selected for concept Penduduk Aikah. This document come from the first verse in the table that is chapter 15, verse 78. We can see that only this chapter that explains about Penduduk Aikah representatively in which it describes who Penduduk Aikah (the companions of the thicket) is (the profile), and they are the wrongdoers. The second verse focuses on what was the Penduduk Aikah did to the messengers, where this is not to explain profile of Penduduk Aikah. The third and the fourth verse are even not put Penduduk Aikah as the only one thing to be explained since they also mentioned others companions such as Tsamud and the people of Lot(third verse) and people of Tubba' (fourth verse).

# 6. CONCLUSION AND FUTURE WORK

We have presented a work in generating Weighted Vector for Concepts in ITQ by implement TFIDF method. These Concepts were gathered from ITQ Ontology that previously defined. Since a preliminary task is for developing a Semantic-based QAS for ITQ, we only applied Concepts which is belong to named entity group Person, Location, and Time as these named entity group are also implemented in QAS. Based on our analysis, the resulted Weighted Vector for Concepts in ITQ can be a benefit in providing an alternative method for doing indexing process on ITQ text as in our previous researches on ITQ text Lucene still failed to accomplish indexing. Selected document for each concept has been proven indeed explain the concept representatively, and this is how indexing process goes for. In another word, the concept can be an index for the selected document.

Our future work will be implemented this Concepts Weighted Vector to build a Semantic-based QAS for ITQ and have them to be used in indexing process as well so that we can get better performance of the system.

# 7. REFERENCES


[1] M. Sheker, S. Saad, R. Abood, and M. Shakir, "Domain-Specific Ontology-Based Approach for Arabic Question Answering," J. Theor. Appl. Inf. Technol., vol. 83, no. 1, 2016.

[2] F. A. Zaghoul and S. Al-Dhaheri, "Arabic Text Classification Based on Features Reduction Using Artificial Neural Networks," in Computer Modelling and Simulation (UKSim), 2013 UKSim 15th International Conference on, 2013, pp. 485–490.

[3] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," J. Doc., vol. 28, pp. 11–21, 1972.

[4] R. Mohamed, M. Ragab, H. Abdelnasser, N. M. El-makky, and M. Torki, "Al-Bayan : A Knowledge-based System for Arabic Answer Selection," in Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), 2015, no. SemEval, pp. 226–230.

[5] S. Park, H. Shim, and G. G. Lee, "ISOFT at QALD-4 : Semantic similarity-based question answering system over linked data," in CLEF 2014 Conference, 2014, pp. 1236–1248.

[6] A. Odeh, A. M. Abu-Errub, Q. Shambour, and N. Turab, "Arabic Text Categorization Algorithm using Vector Evaluation Method," Int. J. Comput. Sci. Inf. Technol., vol. 6, no. 6, 2014.

[7] M. Dittenbach, "Scoring and Ranking Techniques - tf-idf term weighting and cosine similarity." Information Retrieval Facility, 2010.

[8] E. Gabrilovich and S. Markovitch, "Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis," in Proceedings of the 20th International Joint Conference on Artifical Intelligence, 2007, pp. 1606–1611.

[9] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. New York, NY, USA: Cambridge University Press, 2008.

[10] L. R. Group, "The Quranic Arabic Corpus," University of Leeds, 2011. [Online]. Available: http://corpus.quran.com/. [Accessed: 07-Oct-2016].

[11] H. Zarrabi-Zadeh, "Quran Translations," Tanzil, 2016. [Online]. Available: http://tanzil.net/trans/. [Accessed: 07-Oct-2016].

[12] J. Bao, N. Duan, M. Zhou, and T. Zhao, "Knowledge-Based Question Answering as Machine Translation," in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014, pp. 967–97.