
Proceedings Paper:
Liu, Z, Yang, L and Atwell, E orcid.org/0000-0001-9395-3764 (2019) The Semantic Annotation of the Quran Corpus Based on Hierarchical Network of Concepts Theory. In: 2018 International Conference on Asian Language Processing (IALP). IALP 2018, 15-17 Nov 2018, Bandung, Indonesia. IEEE , pp. 318-321. ISBN 978-1-7281-1175-9

https://doi.org/10.1109/IALP.2018.8629241

The Semantic Annotation of the Quran Corpus Based on Hierarchical Network of Concepts Theory

Zhiying Liu, Lijiao Yang*
Institute of Chinese Information Processing,
Beijing Normal University
UltraPower-BNU Joint Laboratory for Artificial Intelligence
Beijing, China
liuzhy@bnu.edu.cn, yanglijiao@bnu.edu.cn

Eric Atwell
School of Computing, University of Leeds
Leeds, UK
E.S.Atwell@leeds.ac.uk

Abstract—**This Quran as the central religious text of Islam is widely regarded as the finest work in classical Arabic literature and plays an important role in Islam world. This paper studied and analyzed the Quran Chinese and English data, built the Quran Chinese and English words semantic knowledge base in which the grammar and semantic information of the Quranic words were described based on HNC theory, built the Quran semantic annotation corpus in which the part of speech and semantic description were annotated. The corpus with semantic annotation can help us identify the same meaning with different word forms. This paper proposed a new method of semantic analysis to solve the semantic similarity problem of natural language processing, which will benefit both the research on the semantic analysis in Natural Language Processing and the development of the Islamic Cultures.**

Keywords-Quran; semantic annotated corpus; semantic knowledge base

I. INTRODUCTION

Quran as the central religious text of Islam is widely regarded as the finest work in classical Arabic literature (Wikipedia) and plays an important role in Islam world.

The Quran is the primary scripture of the faith of Islam. It is the single most important reference for all matters of faith, social practice, the contemplation of law and the understanding of the Divine. It is widely regarded as the finest work in classical Arabic literature. [1]

The Quran is divided into 30 paras (juz in Arabic), 114 chapters (surah), 6236 verses (ayah). It has been analyzed, interpreted, annotated and studied for over a thousand years. The development of computer technology made it possible to do the research with more advanced and powerful ways. Quranic Arabic Corpus was built as an annotated linguistic resource which shows the Arabic grammar, syntax and morphology for each word in the Holy Quran. The corpus provides three levels of analysis: morphological annotation, a syntactic treebank and a semantic ontology. [2]

However, few studies were done on the Chinse Quran in natural language processing. The Quran research in Chinese mainly focused on the culture, philosophy, annotation studies etc. The computer technology and internet are mostly used as the ways of spreading the Quran. [3] The language research on the Quran is mainly concerned with the vocabulary. Studying on the semantics of Quran in Chinese with the computer technology hasn't be reported yet.

This research contributes to the area of semantic analysis and automatic identification and it addresses the challenging task of measuring semantic similarities between any two verses in Quran both in Chinese and English on the base of the Hierarchical Network of Concepts (HNC) Theory.

The analytical and identification research on the Quran can increase the understanding of Islam for people and help Islamists better understand doctrines, which can benefit the development of the Islamic Cultures.

II. RELATED WORK

The Quranic Arabic Corpus (http://corpus.quran.com) is a collaboratively constructed linguistic resource initiated at the University of Leeds, with multiple layers of annotation including part-of-speech tagging, morphological segmentation [4] and syntactic analysis using dependency grammar [5]. A new approach to linguistic annotation of an Arabic corpus was presented: online supervised collaboration using a multi-stage approach. [6] All these show that the research of Quranic Arabic corpus achieved great success.

The HNC (Hierarchical Network of Concepts) theory as a new approach for natural language understanding designed a concept space which can provide a multidimensional semantic knowledge representation method covering words, sentences and discourses. HNC theory constructs an entire theoretical framework for natural language, gives the complete formal description for the meaning of the language. [7] In this Framework, we had built word semantic knowledge base, developed the semantic analysis engine system which can analysing the Chinese language. [8] Various HNC application systems were developed and used like Chinese-English patent machine translation system. By using the semantic analysis engine, HNC Chinese-English Machine Translation system achieved better results in the proceeding of the sentences and discourses. [9]

III. THE BUILDING OF QURANIC CHINESE-ENGLISH BILINGUAL CORPUS

In order to analyze the Quran, we need to build a Quranic Chinese corpus. Zhao X.C. built a Quranic Chinese Corpus with word segmentation and POS (part of

speech) annotation based on the Chinese version translated by Ma Z.W for analyzing the characteristics of Hui Nationality-style Chinese [10]. However, the Quran Chinese version translated by Ma J. is most widely used and accepted by academia. In this case we choose the Ma J version as our Chinese Corpus materials.

Moreover, we choose the English Quran version as a parallel corpus since the meaning of the Quran can always be better understood when two languages are compared with. In the Quranic Arabic Corpus, seven English versions are collected and annotated, in which we choose the version translated by Yusuf Ali as our English Corpus when we found that Yusuf Ali version matched Chinese version better. For example,

TABLE I. VERSE 2:6 OF DIFFERENT TRANSLATIONS

English Version
Sahih International: Indeed, those who disbelieve - it is all the same for them whether you warn them or do not warn them - they will not believe.
Pickthall: As for the Disbelievers, whether thou warn them or thou warn them not it is all one for them; they believe not.
Yusuf Ali: As to those who reject Faith, it is the same to them whether thou warn them or do not warn them; they will not believe.
Shakir: Surely those who disbelieve, it being alike to them whether you warn them, or do not warn them, will not believe.
Muhammad Sarwar: Those who deny your message will not believe whether you warn them or not
Mohsin Khan: Verily, those who disbelieve, it is the same to them whether you (O Muhammad Peace be upon him) warn them or do not warn them, they will not believe.
Arberry: As for the unbelievers, alike it is to them whether thou hast warned them or hast not warned them, they do not believe.
Chinese Version
Ma J: 不信道者，你对他们加以警告与否，这在他们是一样的，他们毕竟不信道。

The sentence has an excellent match between Chinese and English. 不信道者(those who reject Faith) in Chinese is equivalent to the phrase those who reject Faith in English. 道(Faith) in Chinese matched the word Faith in English perfectly in Yusuf Ali version, while other English versions don't give the clear matching word.

Word segmentation in Chinse text is an important step of preprocessing. We use the ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System) to segment the Quran Chinese text, then corrected the word segmentation manually to ensure the accuracy of the segmentation. However, we found that the accuracy of ICTCLAS is not so higher as it is supposed to be. The reason is that Quran as a classical work contains many ancient words and items. So, we need to import the Quran domain words which are obtained by segmenting. After continuously importing the new words and updating the results of segmenting, we got the excellent results of word segmentation like below.

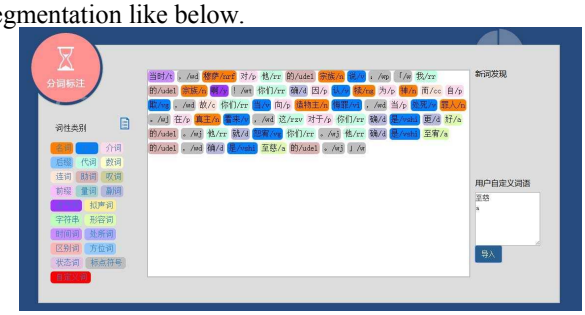

Figure 1. Example of ICTCLAS Segmentation Result

In the original segmenting result, the word 恕宥 (turn towards) was separated into two single Chinese characters 恕(forgive) and 宥(excuse). When you imported the new word 至慈(Most Merciful) followed with its part of speech a (means adjective), the segmenting result is corrected as above. Two more new words 至慈(Most Merciful), 至宥(Oft-Returning) are also added in this paragraph.

Now I have finished the first three chapters' segmentation with segmentation tags and part-of-speech tags.

IV. QURANIC CHINESE-ENGLISH KNOWLEDGE BASE

HNC word knowledge base has already been built containing more than 110,000 words. It is supposed that this knowledge base can be imported and used as Quranic knowledge base. But it is not. Several aspects should be taken into consideration.

A. Word and sense selection

The HNC word knowledge base as the general knowledge resource includes quantity of words with more than one meaning items. And word sense disambiguation is one of the most difficult problems in natural language processing.

In fact, the word in the Quran text is limited. Statistically, there are 29940 Chinese characters, 16446 words, 2265 lexical items in the first three chapters. Therefore, it is unnecessary to import all the word items from the HNC word knowledge base, especially when some of words may cause segmentation confusion.

Besides, the word sense in the Quran text is relatively single without many ambiguations. For example, the word 部(bu) in dictionary has two different meanings when it is used as one-character word. One means the unit of an organization which is not shown in the Quran text; the other is used as a measure word which appears in Quran text.

这部经，其中毫无可疑，是敬畏者的向导。(This is the Book; in it is guidance sure, without doubt, to those who fear Allah.)

On the other hand, the domain terms in Quran should be added into the knowledge base, such as 报应日(Day of Judgement), 阿赖法特(Arafat), 拜星教徒(Sabian), 至慈(most merciful), 至赦(Oft-forgiving).

B. HNC semantic representation

This research focuses on the semantic analysis and comparison in Quran. Accordingly, the HNC semantic representation is extremely important. However, the HNC semantic representation is seldom used in the application system because the representation symbols are so complicated that it is impossible for the system to parse correctly.

In the Quran knowledge base, the HNC semantic representation of a word have to be simplified from both the form and the meaning aspects.

Formally, grammar attribute symbols can be removed from the semantic representation. The HNC semantic symbols were designed to describe both the grammar and semantic attributes originally. The grammar attributes serve the sentence structure analysis and the semantic attributes reveal the meaning of a word. The grammar

attributes have been separately represented as the Concept Categories in the HNC knowledge base because of their importance in syntactic analysis. Therefore, it is unnecessary to keep them in the semantic representation.

The grammar attributes include word category symbols v (verb), g (abstract noun), u (adjective and adverb), z (value noun), r (result noun), w (denoting-thing noun), p (denoting-person noun). After removing these grammar attributes, we can get clearer semantic representation of a word. For example,

TABLE II. SEMANTIC REPRESENTATION CONTRAST 1

Words	Before	After
罪过 (fault)	jg84e72	j84e72
资本 (capital sum)	gza24	a24
主宰 (god)	v44e61	44e61
儿子 (son)	p4099	4099
倾盆大雨 (rain)	rvw5089*2	5089*2

Semantically, the representation also needs to be simplified. Concepts can be divided into simple concepts and complex concepts. Some words' meanings can be expressed in simple concepts, while others can be expressed in complex concepts. Simple concepts can directly correspond to a node in the HNC semantic network, and complex concepts will connect with multiple nodes of semantic network. To simplify the complexity of the process, we assume that all the words' meanings is expressed in simple concepts. In this case, complex concepts representation should be modified to simple concepts.

For example,

TABLE III. SEMANTIC REPRESENTATION CONTRAST 2

Words	Before	After
预备(prepare)	v6500#v11e21;ug11e21	11e21
饮酒(wine)	v65221a|pw65221a	221a
消除(remove)	(v312,l10,(r53322;j84e72/g810;))	312

The word 消除 (remove) is originally represented as the symbol string (v312,l10,(r53322;j84e72/g810;)) which means eliminate (v312) the harm (r53322) or the bad (j84e72) mind (g810). After modifying, it is represented as the simple string 312.

The meaning of a word pointed to the unique node in the semantic network unless the word itself has more than one meaning. The complex association of meaning can be expressed as the relationship between nodes and nodes through subsequent concept-related knowledge base.

C. Ontology description

Different words in form may have the same meaning. The synonyms can be converged together as a node of semantic network. There is internal connection among these nodes. When words are all described as the nodes of the semantic network, the analysis and comparison of the words will be made possible. For example,

```
j8    ::    Ethical
attribute
   j80 :: righteousness and evil      |words:
      j800 ::                         |words:
      j801 :: righteousness           |words: 正义(righteousness);
      j802 :: evil                    |words: 不义(wrong);黑暗(dark)
   j81 ::true and false; solid and void |words:
      j810 ::                         |words: 真伪(right and wrong);
      j811 ::true                     |words: 诚实(true);真相(truth)
      j812 ::false                    |words: 伪信(Hypocrite);
      j815 ::solid                    |words: 诚实(true);
      j816 ::void                     |words: 妄言(desire);
   j82 :: virtue and vice             |words:
      j82e71 :: virtue                |words: 教化;
      j82e72 :: vice                  |words: 恶劣(vile);残杀(slay)
```

Figure 2. Example of HNC Ontology

The Node in HNC semantic network j810 represents the meaning true and false which can be divided into two aspects of j811(true) and j812(false). The words 不义 (wrong), 黑暗 (dark) have different forms but they all correspond to the same semantic representation.

D. Quran knowledge base

We will build the Quranic semantic knowledge base guided by the HNC theory. The knowledge base will provide grammar and syntax knowledge for the semantic analysis engine. A Chinese word in the knowledge base mainly include the attributes such as part-of-speech (POS), concept category (CC), HNC semantic representation, sentence category (SC), English translation (English), etc. For example,

TABLE IV. QURAN KNOWLEDGE BASE

Chinese	Pos	CC	HNC semantic representation	SC	English	Freq
阿丹	pn	pf	f30		Adam	8
阿赖法特	pn	wj2	wj2*6		Arafat	1
啊	interj	f5	f51		O	69
哀哉	interj	f5	f51		woe	3
艾列弗	interj	f5	f51		Alif	2
爱	v	v	7135	X20	love	4
爱慕	v	v	71359	X20	allure	2
安定	v	v	50a9ae8	S	assure	2
安宁	adj	u	50a9ae8		peace	3
安宁地	n	wj2	wj2		place of safety	1

These attributes are presented in different formal symbols and can be processed by the semantic analysis engine system.

V. SEMANTIC ANNOTATION CORPUS

Based on the Quranic knowledge base, we can annotate Quranic Corpus with semantic annotation by attaching the semantic representation behind each word automatically. When the word has more than one meaning, word-sense disambiguation needs to be done. We can use rules to identify which meaning of a word is used in context. Take the Quran verse 3:29 as an example,

你说：「你们的心事，无论加以隐讳，或加以表白，真主都是知道的。他知道天地万物。真主对于万事是全能的。」 Say: "Whether ye hide what is in your hearts or reveal it, Allah knows it all: He knows what is in the heavens, and what is on earth. And Allah has power over all things.

The following is the form annotated automatically. All the word-sense representations are loaded behind the word when it has one more meaning.

*Verse 3:29 你/400e32 说/23*1：「你们/400e32 的/l41/f14 心事/810，无论/lb1 加以/l00 隐讳/332，或/l44 加以/l00 表白/3319，真主/q821*3 都是/jl111 知道/8109 的/f14 。他/400e33 知道/8109 天地万物/jw 。真主/q821*3 对于/l02 万事/jw 是/jl111 全能/44e61 的/l41/f14 。」*

In this paragraph, there are two same words 的 (de) in different place with different word sense. The first word 的 (de) is a structural auxiliary (l41) which connects a pronoun 你们(you) and a noun 心事(mind), the second word 的 is a sentence final auxiliary (f14) which marks the end of a sentence.

On the basis of the above knowledge base, we will build the rule base. The rules will eventually be called by the semantic analysis engine when processing the Quranic texts. The rules of disambiguation about the word 的(de) are as follows:

[1] (0)CHN[的]&END%=> PUT(0, WS_KEY,f14)$

This rule means that word sense f14 is assigned to 的 if we found the word 的 and it is at the end of the sentence.

[2] (0)CHN[的]+(1) LC_CC[g;z;r;w;p] => PUT(0, WS_KEY,l41)$

This rule means that word sense l41 is assigned to 的 if we found the word 的 and behind it is a word with concept category g, z, r,w or p.

The corpus with semantic annotation can help us identify the same meaning with different word forms, and then we can compare the similarity of different verses in Quran.

VI. CONCLUSION

This paper studied and analyzed the Quran Chinese and English data, built the Quran Chinese and English words semantic knowledge base in which the grammar and semantic information of the Quranic words were described based on HNC theory, built the Quran semantic annotation corpus in which the part of speech and semantic description were annotated. We tried to identify the similarities of different chapters in Quran through the analysis of the annotated corpus. This paper used a new method of semantic analysis to solve the semantic similarity problem of natural language processing, which will benefit both the research on the semantic similarity analysis in Natural Language Processing and the development of the Islamic Cultures.

ACKNOWLEDGMENT

This work is supported by China Scholarship Council, National Language Committee Research Program of China (No. ZDI135-42), Research and Development of Question Answering for Intelligent Robots (230200001).

REFERENCES

[1] Alan Jones, The Koran, London 1994, ISBN 1842126091, opening page.

[2] http://corpus.quran.com

[3] Zhao G.J. The Spread, Translation and Research of the Qur'an in China, Gansu Social Science, 2009

[4] Dukes, K., & Habash, N. (2010). Morphological annotation of quranic Arabic. In Language resources and evaluation conference (LREC). Valletta, Malta.

[5] Dukes, K., & Buckwalter T. (2010). A dependency treebank of the quran using traditional arabic grammar. In Proceedings of the 7th international conference on informatics and systems (INFOS). Cairo, Egypt.

[6] Dukes, K., Atwell, E. and Habash, N. 'Supervised Collaboration for Syntactic Annotation of Quranic Arabic'. Language Resources and Evaluation Journal. 2011.

[7] Huang Z. HNC (hierarchical network of concepts) theory. CN: Tsinghua University Press, 1998 (in Chinese)

[8] Jin Y. Natural language understanding based on the theory of HNC (hierarchical network of concepts). CN: Science Press, 2005 (in Chinese)

[9] Zhu Y, Jin Y. A Chinese-English patent machine translation system based on the theory of hierarchical network of concepts, Journal of China Universities of Posts and Telecommunications, v 19, 2012, P140-146.

[10] Zhao X.C. The Quantitative Study on Hui Nationality-style Chinese based on the Corpus, 2013

[11] Cui G, Sheng Y.M. Annotation of Corpus, Journal of Tsinghua University (Philosophy and Social Sciences), 2000(1):89-94