Sains Malaysiana 49(9)(2020): 2113-2118
http://dx.doi.org/10.17576/jsm-2020-4909-09

Query Translation for Multilingual Content with Semantic Technique

(Terjemahan Pertanyaan untuk Kandungan Pelbagai Bahasa dengan Teknik Semantik)

NORITA MD NORWAWI*, SUNDRESAN A/L PERUMAL, EMRAN HUDA & WAKA JENG

ABSTRACT

Cross-lingual information retrieval (CLIR) allows user query in a different language from the language of target resources. Thus, translation is the key element in the query processing. There are three translation approaches: query, document, or hybrid query-document. However, query translation is very challenging due to the polysemy problem. Different linguistic nature of the languages will lead to ambiguity of meaning subsequently user's true intention could be misinterpreted. This paper presents a semantic technique on query translation for a multilingual knowledge repository to improve the query processing. Offline translated documents or parallel corpora in English, Arabic, and Malay language including Jawi *text was used as the data. Set of keywords were constructed preidentified by expert related to prophetic food. These keywords were annotated with the relevant Quranic verses, Hadith texts, Manuscript text images and scientific article determined by expert. The synonym and context-based translation was annotated together with the specific keyword. A query will do a three-way pattern match based on the keyword indexing list that link to the relevant documents. A one-stop knowledge repository on prophetic food was developed as a proof of concept using sources are from al-Quran, Hadith, classical manuscript, and scientific articles verified by experts to ensure the content authenticity and integrity.*

Keywords: Cross lingual information retrieval; one stop knowledge repository; prophetic food; query translation; semantic technique

ABSTRAK

Dapatan semula maklumat silang bahasa (CLIR) membolehkan pertanyaan pengguna diajukan dalam bahasa yang berbeza daripada bahasa bahan sumber sasaran. Oleh itu, terjemahan menjadi kunci utama dalam pemprosesan pertanyaan. Terdapat 3 jenis pendekatan terjemahan: terjemahan pertanyaan, dokumen atau pertanyaan-dokumen hibrid. Walau bagaimanapun, terjemahan pertanyaan adalah mencabar berpunca daripada masalah polisemi. Gaya linguistik pelbagai bahasa yang berbeza menimbulkan kesamaran makna yang menyebabkan hasrat sebenar pengguna boleh disalah tafsir. Kajian ini membentangkan teknik semantik terjemahan pertanyaan repositori pelbagai bahasa untuk menambahbaik pemprosesan pertanyaan. Dokumen sumber yang diterjemahkan secara manual atau corpora selari dalam Bahasa Inggeris, Arab dan Melayu termasuk teks Jawi digunakan sebagai data kajian. Set kata kunci telah dikenal pasti oleh pakar bidang berkaitan dengan makanan sunnah. Kata kunci ini dianotasikan dengan ayat-ayat Al-Quran teks Hadith, teks dan imej manuskrip dan artikel saintifik yang berkaitan oleh pakar bidang berkenaan. Perkataan sinonim dan terjemahan secara konteks dianotasikan juga kepada kata kunci berkaitan. Setiap pertanyaan akan menggunakan 3 kaedah pemadanan ke atas senarai indeks kata kunci yang akan menghubungkan kepada dokumen yang relevan. Repositori pengetahuan sehenti berkaitan makanan sunnah dibangunkan sebagai bukti konsep menggunakan sumber daripada Al-Quran, Hadith, manuskrip klasik dan artikel saintifik yang disahkan oleh pakar bidang untuk menjamin kesahihan dan integriti.

Kata kunci: Dapatan semula maklumat silang bahasa; makanan sunnah; repositori pengetahuan sehenti; teknik semantik; terjemahan pertanyaan

INTRODUCTION

With the advancement of computer, Internet and communication technology, information access has advanced into many tools and applications such as the information retrieval (IR), question answering tasks, summarization, multimedia information retrieval, text mining, and clustering and Web information retrieval. However, today, due to the diversity of information and language barriers, keeping up with communication and cultural interchange across the globe is very challenging. As for the IR, there are 4 categories of retrieval which are: mono-lingual, bi-lingual, cross-lingual (CL), and

multi-lingual (ML). Cross-language and multi-lingual IR is a retrieval process in which the user fires query in one language to retrieve documents from another language. The query is posed in the source language and the language of the relevant document is the target language.

On the other hand, handling heterogenous knowledge sources is challenging due to the diversity of perspectives and context of the knowledge. The level of complexity increases since these resources are in different languages from each other. Today, CLIR and MLIR becomes more important due to the rapid progress of Internet technology that opens the boundary of language and culture. Hence, translation is the key signature task in the query processing due to the multilingual nature. There are three main approaches in handling the translation which are: document, query or hybrid query document (Agbele et al. 2018) as shown in Table 1.

TABLE 1. Comparison of the three translation approach (Agbele et al. 2018)

SN	Parameter	Query Translation	Document Translation	Query-Document Translation
1	Extra storage space	Not needed	Needed	Not needed
2	Ambiguity	maximum	Minimum	More than both query and document
3	Information Retrieval	Bilingual	Bilingual	Bilingual and Multilingual
4	Transition time	Less	More than query	More than both query and document
5	Flexibility	Highly	Less	Less

TABLE 2. Comparison of the three translation approach (Agbele et al. 2018; Jena & Rautaray, 2019)

Translation Approach	Dictionary Based	Corpora Based	Machine Translation Based
Ambiguity	High	Low	Low
Offline Translation	Possible	Possible	Not Possible
Working Architecture	Visible as like white box testing	Visible as like white box testing	Works similar to black box testing
Development expenses	Less expensive	More expensive than DBT	More expensive
Translation Availabilty	Highly available in many languages	Available only in few languages	Available only in few languages

Based on Table 1, Agbele et al. (2018) highlighted that the query-document translation approach does not require extra storage space but have issues related to language ambiguity and speed due to synonymy and polysemy problem.

Synonymy refers to multiple words with common meaning whereas polysemy refers to words with multiple meanings. Synonymous and polysemous words will reduce the recall and precision rates (Azad & Deepak 2019). Due to different in linguistic nature, language

ambiguity will lead to misinterpretation of user's true intention especially for translation. Hence, processing a query and its translation for CLIR and MLIR with heterogeneous knowledge resources ensuring right context and diverse perspectives is a challenge (Abusalah et al. 2005; Aldhlan et al. 2010; Elayeb & Bournas 2016; Sharmal & Morwal 2015).

As for the query translation, there are three sources of knowledge: dictionary, corpora, or machine learning. Table 2 illustrates the comparison of these techniques (Agbele et al. 2018; Jena & Rautaray 2019).

In this study, corpora-based query translation was used which has low ambiguity than dictionary based, and moderate development cost compared to machine learning based translation as presented in Table 2.

This paper presents a semantic technique for query translation to handle the cross-lingual ambiguity of the diverse resources where query and targeted text are possibly in Arabic, Malay language, or English. Users may pose query in English, Arabic or Malay to search for relevant text in Al-Quran, Hadith, classical Malay manuscript in Roman or *Jawi* text and scientific articles related to prophetic food. Search results will embed the direct match with the semantic technique of using translation, synonym and context based on meaning as determined by expert.

Materials and Methods

A knowledge repository was prepared consists of Quranic verses, Hadith text, classical manuscript text, and scientific articles related to prophetic food. Initially, date and goat's milk were selected as the scope of the study. Begin with content analysis and systematic literature review conducted by domain experts in Quran and Sunnah studies, Manuscript on Malay Remedies, Food Science and Health Sciences. All relevant resources were identified and gathered from al-Quran, Hadith, manuscript, and scientific articles. A knowledge taxonomy on dates and prophetic food was build based on the literature, concept, and relations to the Quranic, Hadith, manuscripts, and scientific articles. Based on this taxonomy, and keywords pre-identified by experts, a structural connection was designed for the specialized search engine on prophetic food which later implemented as a web and mobile application called NAISSE.

Results and Discussion

Based on the knowledge collected, a taxonomy of knowledge was constructed on dates and goat's milk linking to the resources that map these knowledge resources to pre-identified keywords (Tawil et al. 2016). These keywords identified by domain experts were tagged with the translation in the other two languages, its synonym or contextual meaning. For example, *kurma* in Bahasa Malaysia could mean ر مت, بطر, لخن in Arabic. The Quranic, Hadith, and classical manuscript were translated offline and become the parallel corpora. Scientific articles were represented by their URL address and its details including abstracts. Figure 1 illustrates the representation of the knowledge repository.

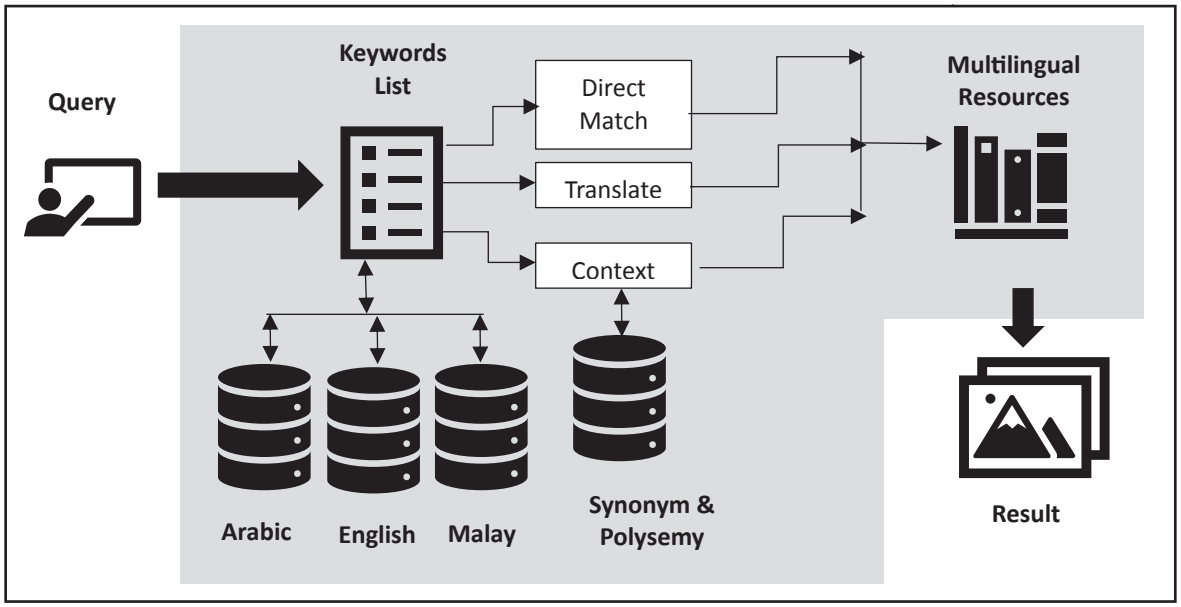

FIGURE 1. Schematic diagram of the multilingual knowledge repository

In order to resolve the linguistic ambiguity issue due to query translation, a semantic technique was proposed as a mechanism that will accept the query in either of the three languages then matched either exactly, based on translation or context match to the relevant documents. Keywords list contains possible keyword identified by expert relevant to the scope and resources. Each keyword will record the list of relevant and specific ID of the Quranic verses, Hadith text, manuscript snippet images and the scientific articles. It will be linked to an index of prophetic food names tagged with its translation and scientific name as shown in Figure 2.

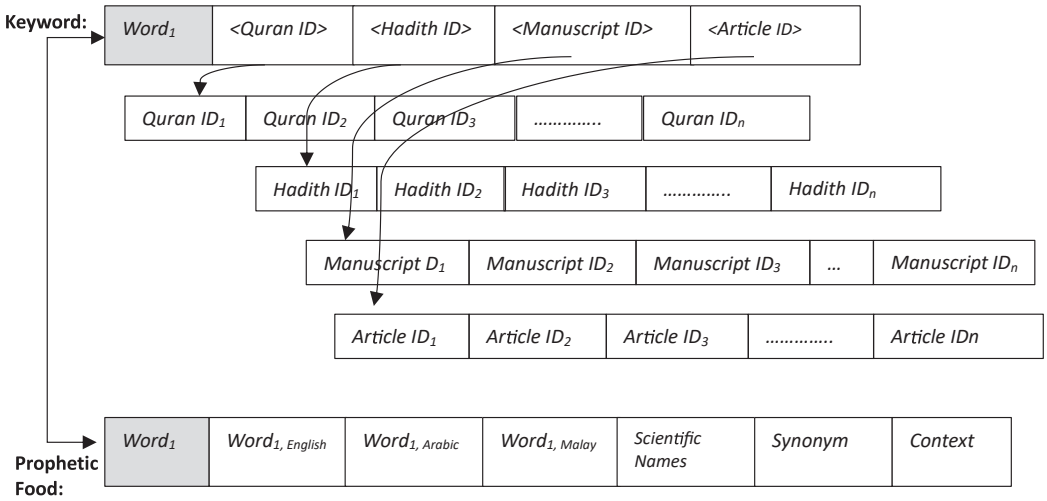

FIGURE 2. Illustration of keywords list structure linked to prophetic food list

Referring also to Figure 1, the query will be matched through 3 different steps which are a direct match (***D***), dictionary-based match for the translation (***T***) and context-based match (***C***) through its synonyms (***S***) and semantic (***M***) defined by the knowledge taxonomy, ***O***. The semantic technique is represented formally as follows:

Assume ***L*** is the set of all matching words, ***w***, from all the resources Quran (***Q***), Hadith (***H***), Malay manuscript in roman text (S_r), Malay manuscript in *Jawi* text (S_j) and scientific articles (***J***). Thus,

$L = \{w|\ w \in \mathrm{D}\} \cup \{w|\ w \in T\} \cup \{w|\ w \in C\}$ where ***L*** are the list of matched keywords such that

Direct Match ***D***

$D_a = \{w|\ w \in \mathrm{Q}\} \cup \{w|\ w \in H\}$ for Arabic text

$D_m = \{w|\ w \in \mathrm{S_r}\} \cup \{w|\ w \in S_j\}$ for Roman or *Jawi* Malay text

$D_e = \{w|\ w \in J\}$ for English text

thus, $\boldsymbol{D} = D_a \oplus D_m \oplus D_e$ implies only an exact match with the same language as query or the source language.

Dictionary Based Translation Match ***T***

As for the dictionary-based translation ***T***,

$T_m = \{w|\ w \in \mathrm{Q}\} \cup \{w|\ w \in H\} \cup \{w|\ w \in J\}$ for query in Malay language, ***m***

$T_a = \{w|\ w \in \mathrm{S_r}\} \cup \{w|\ w \in S_j\} \cup \{w|\ w \in J\}$ for query in Arabic language, ***a***

$T_e = \{w|\ w \in \mathrm{Q}\} \cup \{w|\ w \in H\} \cup \{w|\ w \in \mathrm{S_r}\} \cup \{w|\ w \in S_j\}$ for query in English language, ***e*** and

$$\boldsymbol{T} = \{w_l|\ w_l \in T_a\} \cup \{w_l|\ w_l \in T_m\}\ \&\ l \neq a, l \neq m \oplus$$

$$\{w_l|\ w_l \in T_a\} \cup \{w_l|\ w_l \in T_e\}\ l \neq a, l \neq e \oplus$$

$$\{w_l|\ w_l \in T_m\} \cup \{w_l|\ w_l \in T_e\}\ l \neq m, l \neq e$$

where ***T*** is the set of translation of the keyword $\boldsymbol{w_l}$ in other language

Context based match ***C***

$$C = \{w_s | \, w_s \in N\} \cup \{w_m | \, w_m \in O\}$$

where w_s are words matched in the synonym list, ***N*** and w_m are words matched from the knowledge taxanomy, ***O***.

A CLIR application was developed called NAISSE (naisse.org). It is an online one-stop knowledge repository on prophetic food later developed as mobile apps. The knowledge sources are from al-Quran and Hadith which are in Arabic, classical manuscript in Malay language either written in Roman or *Jawi* text, and scientific articles which are mainly in English manage using a content management system (Norwawi et al. 2019). Results from NAISSE search engine on Quran were also compared to *https://quran.com*, a search engine for Al-Quran as shown in Table 3.

TABLE 3. Comparison search results between *quran.com* and *naisse.org*

Query (Single word)	No. of verses matched from *quran.com*	No. of verses matched from *naisse.org*	Similar verses	Synonym *naisse.org*	Context *naisse.org*	Spelling Variety (Arabic) *naisse.org*	Duplicate verse
Kurma (Malay)	26	29	19	1	4	4	1
Dates (English)	30	29	17	4	7	15	0
النخيل (Arabic with Diacritic)	1	29	1	5	4	18	1
بطر (Arabic without diacritic)	2	3	1	0	0	2	0

Based on Table 3, the results show that *naisse.org* is comparable to *quran.com* for Malay and English word except in Arabic. Upon examining and comparing the results, Arabic language source of ambiguity is due to the diacritics sign, different way of spellings due to grammatical syntax rules. However, this could be overcome using translation, synonym, and context. For example, if the query is *Kurma* which is in Malay language, there are 6 exact matches with the Malay translation of the relevant Quranic verses, 1 synonym which is *Tamar* and the rest are either the translation into English and Arabic or context based on the synonym or semantic as determined by expert. Query translation using the parallel corpora is simple due to the translation already available for all the text in its respective documents as also recommended by Prasath et al. (2015) who use Tamil and English in their study. This study will continue with further improvement in handling the Arabic language as evident in Table 3. There are varieties of terms of Arabic related to *Kurma* and *Dates* for example نخيل, بطر, الّيتِفَ, نٌاوَنْص, ارّيقِنَ and in context of meaning ةِبَيّطَ ةِرَجَشَ, ريمِطُق, دّيضِنَ عُلْطَ and ةِنَيلِ نْمِ. The different spelling and diacritics sign will also influence the search result.

Conclusion

This study demonstrated the interconnection of knowledge on prophetic food from different sources such as Al-Quran, Hadith, Classical Manuscript, and Scientific Articles merged, collated, and presented in a one-stop repository for easy access. A multilingual knowledge repository, NAISSE, developed with three major keyword matching mechanism: direct match, dictionary-based translation and context based through synonym or its semantic represented by the knowledge taxonomy, structure of the keyword and resources identification. NAISSE will match the keyword search with multi documents using a semantic technique for its query translation based on the translation, synonym

and semantic. It can be expandable to other languages by developing the linguistic matching algorithm that suits the language.

There are lots of potential for the future works since this is an initial study for MLIR using Arabic, English, Malay including the *Jawi* text using data from Quran, Hadith, classical manuscript and articles related to prophetic food specifically dates and goat's milk. Future study may take advantage of the simple approach of offline document translation with hybrid with online query translation with semi-automated generation of keywords that need to be validated first by experts. In terms of search results, it can be improved by focusing on reducing the language ambiguity through parallel corpora based and handing the grammatical aspect of the language especially Arabic.

ACKNOWLEDGEMENTS

The authors express their appreciation for the Niche Research Grant Scheme, awarded by the Ministry of Education to Universiti Sains Islam Malaysia (USIM) with project code NRGS-P/FST/8404/52113.

REFERENCES

Abusalah, M., Tait, J. & Oakes, M. 2005. Literature review of cross-language information retrieval. *World Academy of Science, Engineering and Technology* 4: 175-177.

Agbele, K.K., Ayetiran, E.F. & Aruleba, K.D. 2018. Survey on cross-lingual information retrieval. *International Journal of Scientific & Engineering Research* 9(8): 484-491.

Aldhlan, K.A., Zeki, A.M. & Zeki, A.M. 2010. *Datamining and Islamic knowledge extraction: Alhadith as a knowledge resource.* In *Proceedings of the International Conference on Information and Communication Technology for Muslim World (ICT4M)*. IEEEE. H-21.

Azad, H.K. & Deepak, A. 2019. Query expansion techniques for information retrieval: A survey. *Information Processing & Management* 56(5): 1698-1735.

Elayeb, B. & Bournas, I. 2016. Arabic cross-language information retrieval: a review. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 15(3): 1-44.

Jena, G.C. & Rautaray, S.S. 2019. A comprehensive survey on cross-language information retrieval system. *Indonesian Journal of Electrical Engineering and Computer Science* 14(1): 127-134.

Norwawi, N.M., Perumal, S., Sempo, M.W., Huda, E. & Jeng, W. 2019. Multi-lingual content management system for prophetic food. In *Proceedings of the International Conference on Islamic Applications in Computer Science and Technologies (IMAN 2019)*. 27(28).

Prasath, R., Sarkar, S. & O'Reilly, P. 2015. Improving cross language information retrieval using corpus based query suggestion approach. In *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, Cham. pp. 448-457.

Sharma, M. & Morwal, S. 2015. A survey on cross-language information retrieval. *International Journal of Advanced Research in Computer and Communication Engineering* 4(2): 384-387.

Tawil, S.F.M., Ismail, R., Wahid, F.A., Norwawi, N.M. & Mazlan, A.A. 2016. Application of OASys approaches for dates ontology. In *Third International Conference on Information Retrieval and Knowledge Management (CAMP)*. IEEE. pp. 131-135.

Faculty of Science and Technology
Universiti Sains Islam Malaysia
71800 Nilai, Negeri Sembilan Darul Khusus
Malaysia

*Corresponding author; email: norita@usim.edu.my

Received: 23 January 2020
Accepted: 1 April 2020