WORD SENSE DISAMBIGUATION TO ENHANCED NATURAL LANGUAGE QUESTIONS FOR PILGRIMS

BY

AMMAR FUAD O ARBAAEEN

A thesis submitted in fulfilment of the requirement for the degree of Doctor of Philosophy in Computer Science

Kulliyyah of Information and Communication Technology
International Islamic University Malaysia

JUNE 2022

ABSTRACT

The tremendous growth in the field of data science and the widespread usage of information retrieval techniques has enabled users to retrieve accurate information. The diverse data availability in various Knowledge Base (KB) formats introduces several challenges to deliver concise and precise information corresponding to human queries. This requires a user to be familiar with the structures of KB and use a formal query language for the system to effectively understand the query. Question Answering (QA) systems have been introduced to enable users to post questions in Natural Language (NL) and infer specific answers instead of lists of documents. Such a system requires the capability of both critical analysis on questions and inference on answers selection. NL question analysis module is a fundamental step that impacts the QA system performance. It aims to transform users' NL questions into representations of a structured format suitable to query across KBs. Literature showed, the major challenge in NL question transformation is language ambiguity that may occur at a lexical-semantic level. Moreover, various challenging questions require handling ambiguities based on a certain condition such as seeking instructions or advice. Therefore, the motivation of this study is to propose a Knowledge-based Sense Disambiguation (KSD) method for resolving the problem of lexical ambiguity associated with NL questions. This algorithm is designed by incorporating question's metadata (date/GPS), context knowledge, and domain ontology, into a shallow NL processor. It aims at enhancing the accuracy of the word sense disambiguation process in questions analysis module to effectively returns potential answers corresponding to questions posed in QA systems. This work explores the use of the proposed KSD method to support pilgrims in expressive queries to obtain accurate information via a mobile QA application. Therefore, the validity of the proposed solution has been supported by two experiments to evaluate the accuracy performance. First, in vitro experiment was carried out as a standalone task to evaluate the KSD as a word sense disambiguation method in comparison with the baselines WordNet Most Frequent Sense (MFS) and the simplified version of the Lesk method on the same condition and test dataset. Second, in vivo experiment was performed to evaluate the effectiveness of the KSD method in a QA application in comparison to the MFS in the context of a pilgrimage domain. The results obtained from both experiments have revealed the feasibility of the proposed solution to effectively cope with lexical ambiguity in NL questions as well as to contribute to QA system performance improvement.

"We don't need more information. We need more meaning." Paul Salopek

خلاصة البحث

Abstract In Arabic

إزالة الغموض في معنى الكلمة لتحسين أسئلة اللغة الطبيعية للحجاج

أدى النمو الهائل في مجال علم البيانات والاستخدام الواسع لتقنيات استرجاع المعلومات إلى تمكين المستخدمين من استرداد المعلومات الدقيقة. يقدم توافر البيانات المتنوعة في تنسيقات قواعد المعرفة المختلفة العديد من التحديات لتقديم معلومات موجزة ودقيقة تتوافق مع الاستفسارات البشرية. يتطلب هذا أن يكون المستخدم على دراية بهياكل قواعد البيانات ويستخدم لغة استعلام رسمية للنظام لفهم الاستعلام بشكل فعال. تم تقديم أنظمة الإجابة على الأسئلة لتمكين المستخدمين من طرح الأسئلة باللغة الطبيعية واستنتاج إجابات محددة بدلاً من قوائم المستندات. يتطلب مثل هذا النظام القدرة على التحليل اللغوي على الأسئلة والاستدلال على اختيار الإجابات. يعد تحليل ومعالجة أسئلة اللغة الطبيعية خطوة أساسية تؤثر على أداء معظم الأنظمة ذات الصلة مثل أنظمة الإجابة على الأسئلة والاستفسارات. ويهدف إلى تحويل أسئلة اللغة الطبيعية للمستخدمين إلى تنسيق تمثيل منظم مناسب للاستعلام عبر قواعد المعرفة. أظهر المسح الأدبي أن التحدي الرئيسي في تحليل وتحويل أسئلة اللغة الإنجليزية هو الغموض اللغوي الذي قد يحدث على المستوى المعجمي الدلالي. علاوة على ذلك، تتطلب العديد من الأسئلة الصعبة معالجة الغموض بناءً على البيانات الوصفية للسؤال الخاصة بحالة معينة مثل طلب التعليمات أو المشورة. لذلك، فإن الدافع من هذه الدراسة هو اقتراح طريقة جديدة لإزالة الغموض القائم على المعرفة، تهدف إلى تعزيز دقة عملية توضيح المعنى المعجمي في أسئلة اللغة الطبيعية. تم تصميم هذه الخوارزمية من خلال دمج معلومات البيانات الوصفية للسؤال، ومعرفة السياق، وأنطولوجيا المجال، عن طريق تطوير معالج لغوي تقني. يساعد على تحديد المعنى المقصود وتعيين المعنى المناسب للكلمات الغامضة التي تحدث في أسئلة اللغة الإنجليزية فقط للإجابة بشكل فعال على الأسئلة المطروحة في مجال الاهتمام. يستكشف هذا العمل الجديد والفريد من نوعه استخدام النهج المقترح لدعم الحجاج في الاستفسارات التعبيرية للحصول على معلومات ومعرفة دقيقة عبر تطبيق الاستفسارات مع ضمان الجودة على الهواتف الذكية. نظرًا لأن النهج المقترح هو جزء من بنية نظام استفسارات، فقد تم دعم صحة النهج المقترح من خلال تجربتين لتقييم أداء الدقة في حل الغموض اللغوي المرافق مع أسئلة اللغة الطبيعية للمستخدمين. أولاً، تم إجراء التجربة في المختبر كمهمة قائمة بذاتها لتقييم النهج المقترح كطريقة لإزالة الغموض عن معنى الكلمة بالمقارنة مع المعنى الأول لـ WordNet والنسخة المبسطة من النهج الرسمي Lesk المستخدم لتوضيح معنى الكلمة على نفس الحالة ومجموعة بيانات الاختبار. تم إجراء التجربة الثانية لتقييم تأثير النهج المقترح مقارنة مع WordNet بالمعنى الأول في سياق تطبيق استفسارات الحج. أظهرت النتائج التي تم الحصول عليها من كلتا التجربتين جدوى النهج المقترح للتعامل بشكل فعال مع الغموض المعجمي في أسئلة اللغة الطبيعية لاستنتاج الإجابات المحتملة وكذلك المساهمة في تحسين أداء نظام الإجابة على الأسئلة. تشير النتائج أيضا إلى أن المنهجية العلمية المقترحة حققت أداءً قابلاً للمقارنة وأفضل دقة من المناهج العلمية الرسمية والخطوط الأساسية في سياق مجال الحج والعمرة.

"لسنا بحاجة إلى مزيد من المعلومات. نحن بحاجة إلى مزيد من المعنى".
Paul Salopek

APPROVAL PAGE

The thesis of Ammar Fuad O Arbaaeen has been approved by the following:

Asadullah Shah
Main Supervisor

Muhamad Sadry Abu Seman
Co-Supervisor

Adamu Abubakar Ibrahim
Co-Supervisor

Akram Mohammed Zeki Khedher
Internal Examiner

Abd. Samad Bin Hasan Basari
External Examiner

Meftah Hrairi
Chairman

DECLARATION

I hereby declare that this thesis is the result of my own investigations, except where otherwise stated. I also declare that it has not been previously or concurrently submitted as a whole for any other degrees at IIUM or other institutions.

Ammar Fuad O Arbaaeen

Signature ..... Date 23/06/2022

INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA

DECLARATION OF COPYRIGHT AND AFFIRMATION OF FAIR USE OF UNPUBLISHED RESEARCH

WORD SENSE DISAMBIGUATION TO ENHANCED NATURAL LANGUAGE QUESTIONS FOR PILGRIMS

I declare that the copyright holders of this thesis are jointly owned by the student and IIUM.

Affirmed by Ammar Fuad O Arbaaeen

……..…….. 23/06/2022

Signature ……..
Date

ACKNOWLEDGEMENTS

Praise be to Allah (SWT) the almighty for giving me strength, ability with light to learn, understand, patience with health and focus to accomplish my goals. Although the path has not been smooth, it's been a wonderful journey full of challenges, frustrations and memories that I'm unable to express in words. I've grown from all of these experiences and learned that there is always a new chapter ready to start. And for all of these, I am deeply grateful. May peace and blessing of Allah be upon our Beloved Prophet Muhammad (SAW), his family and his companions.

Special thanks to my Professor. Dr Asadullah Shah for his continuous support, encouragement and leadership, and for that, I will be forever grateful.

I wish to express my appreciation and thanks to those who provided their time, effort and support with their prayers to achieve my aspirations particularly my supportive brothers & sister, and my dear friends. To the respectful members of my dissertation committee Dr Muhamad Sadry and Dr Adamu Abubakar, thank you for sticking with me. I must also express my thanks to the members of IIUM who provided direct or indirect support towards my achievement. I must also express my deep thanks to the ministry of higher education – the Saudi Arabian' government who's funding my education to achieve my dreams.

Special gratitude to my supportive wife and daughters Rital, Ruba, and Rahaf for their love and patience through my journey. You are all now at the starting, it is my wish that your educational journeys would be successful beyond your imagination.

Finally, I would like to dedicate this work to the soul of my wonderful mum who just passed away five months before the completion of my journey and to the soul of my incomparable father who could not see this achievement. Who always prayed for me. May Allah (SWT) bless them and make their graves a garden and grant them the highest levels of paradise.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATION

AI	Artificial Intelligence
CS	Computer Science
DBR	Design-Based Research
FAQ	Frequently Asked Question
IE	Information Extraction
IR	Information Retrieval
IT	Information Technology
KB	Knowledge Base
KSD	Knoweldge-based Sense Disambiguation
MFS	Most Frequent Sense by WordNet
MRDs	Machine Readable Dictionaries
NE	Named Entity
NL	Natural Language
NLP	Natural Language Process
OWL	Ontology Web Language
POS	Part Of Speech
QA	Question Answering
RAD	Rapid Application Development
SQL	Structured Query Language
UI	User Interface
WSD	Word Sense Disambiguation
XML	Xtensible Markup Language

CHAPTER ONE

INTRODUCTION

1.1 MOTIVATION

Information retrieval (IR) technology assists the process of retrieving and searching for information. The overall approaches employed by IR uses keywords to obtain relevant information from various heterogeneous resources (Y. Jiang, 2020), thereby offering a constraint which holds a precise conceptualization of the user' expression as well as the meaning of their content. Upon searching, the user obtains a list of relevant documents which might consists of the desired information. Nevertheless, users may look for a certain answer, rather than a number of documents. Although IR techniques could be highly successful and obtain relevant information, users encounter complicated linguistic difficulties to obtain concise and precise information. Question Answering (QA) is a Computer Science (CS) discipline that provides techniques to deal with this issue by making it potential for users to express queries and infer concise and precise answers in Natural Language (NL).

Literature (Dimitrakis, Sgontzos, & Tzitzikas, 2020; Ishwari et al., 2019) reveals that the major challenge for a question answering (QA) system is the ambiguity of language during the conversion process of a natural language question into structured representation format. Mainly, this ambiguity results from two or more senses of the same word which negatively impact the process of extracting accurate answer from a knowledge base. Starting from this point, the key objective of this research is to investigate the importance of word sense disambiguation to support the capabilities of

lexical ambiguity resolution in questions processing module. Therefore, this chapter introduces a knowledge-based sense disambiguation method that can effectively resolve the problem of lexical ambiguity associated with NL question posed in a QA system for a domain of interest.

1.2 BACKGROUND OF THE STUDY

Tremendous growth in the field of information and communication technologies has made available a huge amount of disparate information that is stored in the form of textual documents and data. The technological evolution is transforming and extending the principles of website architecture from documents (traditional database) to data (semantic data representation). This revolutionary approach supports the creation of a common framework that allows data to be used and shared across various application areas in a meaningful linking data format. It also supports the process of automatic data integration and knowledge management such as knowledge sharing and knowledge extraction. Semantic technology uses ontology as a knowledge representation where data are described and stored in the form of rich vocabulary and expressive properties and object relations of domain in machine readable format. Data may contain appropriate information that offers valuable knowledge to users in a specific domain. Therefore, the content description and query processing techniques are crucial to obtain information resources and provide accurate results to users.

The emerging field of Information Retrieval (IR) technology supports the process of finding and retrieving information. The IR process involves the information retrieval based on keywords from different heterogeneous resources (Al-Harbi, Jusoh, & Norwawi, 2017). Therefore, it has limitations in gaining precise results in content

meaning of user expression in the query. The user receives a list of relevant documents upon request of the query and s/he may find the desired information within the list. However, users may request accurate and comprehensive answers rather than a list of documents. Although literature shows that IR techniques have demonstrated tremendous success in retrieving relevant information, the users still face many complex challenges for the extraction of desired information in natural languages. Therefore, QA techniques are useful in dealing with this complexity, which allows users to express and extract precise information and knowledge in natural languages (Ojokoh & Adebisi, 2019).

IR based on QA system involves the research and methods from NLP, Information Extraction (IE), Automatic Summarisation, Knowledge and Database Management. In QA applications, the user obtains a more concise and relevant answer to the NL questions from the stored documents, while in IR applications, the user searches by keywords as input and receives a relevant list of documents based on her/his query. NLP supports a QA system with expression techniques to understand the question and provide the appropriate response in a natural language format. Moreover, NLP components are also used to analyse the questions returned by the IR system and provide valuable meaning of the terms given in the question. On the other hand, IE components utilise resources, such as entity tagging, template elements, template relation, correlated elements and general elements to obtain the concise and relevant information from the stored documents. Therefore, a QA system attracts many researchers who are achieving significant enhancements in different domains such as biomedicine, weather forecasting and tourism (Al-Harbi et al., 2017).

Table 1.1 Differences between QA and IR (Guda, Sanampudi, & Manikyamba, 2011)

Function	Information	Question Answering
Input	Keywords	NL questions
Output	List of documents	The phrases as well as terms including the

The basic architecture of a QA system consists of three major modules: question processing module, a documents and passages processing module, and an answer processing module, illustrated in Figure 1.1 (Ojokoh & Adebisi, 2019). The question processing module is responsible for analysing the received questions, understanding and identifying what is being asked. Analysing and classifying the questions is performed on the basis of different methods which range from keyword extraction to deep linguistic-based methods (Pundge, 2016). The syntactic and semantic aspects of the questions are examined to understand the purpose of that particular questions. As a result of the analysis, the extracted keywords are used by IR as the query term for retrieving the candidate and relevant information in a document processing module. This relevant information is further analysed in the passage processing task using methods such as sentence splitting, part-of-speech (POS) tagging and parsing. Finally, the answer processing module uses the result of the analysis to infer the precise answer based on the question's representation and ranking of candidate answers.

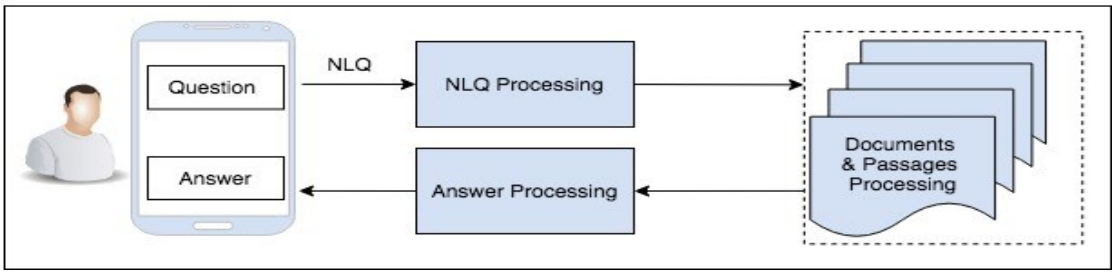

Figure 1.1 Basic Pipeline Architecture of QA System (Ojokoh & Adebisi, 2019)

A close look at the literature on QA systems, It shows clearly that QA systems are classified on the basis of several factors such as domain dimension, question types and the type of analysis done on question (Mishra & Jain, 2016a). In terms of domain, most QA systems are classified into two types: a web-based question answering system (open domain) and a restricted question answering systems (closed domain). This study focuses on the restricted domain to meet its objectives. In this respect, literature shows that most QA system classify NL questions into two types, namely, complex questions and factoid questions. The simple facts are needed to answer a factoid question, while summarising and grouping paragraphs are needed to answer a complex question. In the present research, the focus is placed on questions which contain multiple words and belong to WH-types of question while ignoring the degree of complexity. A thorough literature review shows that the answers are extracted for the NL question from either a structured, semi-structured or unstructured database of the QA system (Guo et al., 2020). Question answering system dimensions is discussed in detail in Chapter 2.

The advance in QA application is challenging as it involves the complex development of various methods in a question processing module. The fundamental part of a QA system is known as the question processing module, whose quality of processing greatly affect the overall performance of QA application. The literature shows that there are several methods classified into two major types: statistical methods (Quarteroni & Manandhar, 2009) and linguistic methods (Bhoir & Potey, 2014; Mishra & Jain, 2016a). Some knowledge-based QA systems further rely on rule-based mechanisms, besides the NLP techniques, in order to classify the question features. Further details about the methods are provided in the literature review (Chapter 2). The literature review clearly shows that the current methods used by a QA system still suffer from numerous obstacles, such as ambiguity (Höffner et al., 2017; Ojokoh & Adebisi,

2019). The deployment of statistical methods in a question processing module analyses the frequency of word occurrence in documents but ignores many aspects including words order and lexical semantics of terms. In the present research, the focus is placed on the question processing module that utilises the NLP method to correctly determine the intended meaning of the posed question. NLP methods are classified into two major types: deep NLP and shallow NLP processing. The deep NLP utilises the complex parsing algorithms, such as top-down parser, which can deeply analyse the whole structure of the sentence resulting in low coverage and more time consumption (M. C. Yang, Lee, Park, & Rim, 2015). On the other hand, the shallow processing generates a partial internal structure which is sufficient enough to answer a NL question with faster processing and high coverage (Al-Harbi et al., 2017). The detailed review of both methods is presented in Chapter 2.

The question processing module is used as part of a natural language interface of a QA system prototype in the current research. Literature clearly showed that the major challenge faced by NLP based question processor is the usage of a natural language which is full of ambiguity (Corrêa, Lopes, & Amancio, 2018; Höffner et al., 2017). This ambiguity in NL question leads to an irrelevant answer or no answer. Ambiguity is defined as a barrier to human language understanding (Y. Wang, Wang, & Fujita, 2020). The lexical ambiguity is used in the present research to handle more than one interpretation for an individual word. For instance, if the term "chair" is used, then it presents different lexical definitions, including "a seat for a person" and "the officer who presides at the meeting of organizations." Moreover, handling complex ambiguities associated with NL questions may require metadata of the question (date/GPS) which provides some semantic structure to particular question such as seeking advice in a critical situation (Pillai, Veena, & Gupta, 2018; Y. Wang et al.,

2020). For instance, “what are the hajj pillars?” the sense of “pillars” in the context of the above question refers to “a fundamental principle or practice” defined by WordNet 3.1. Whereas “pillars” in the following context “how should I reach the pillars?” refers to “anything that approximates the shape of a column or tower” such as the stoning of the devil pillars at Mina area, mainly because the question posed by a user at Mina surrounding area during the Hajj season. Similarly, “date”, “bank”, “stone” as a target words associated with seeking instruction NL questions, where the question’s metadata (date/GPS) knowledge identifier may support the semantic capabilities processing to determine the appropriate sense. This ambiguity challenges is widely encountered in the literature but still require many efforts from research community to resolve numerous problems of lexical ambiguity (Aouicha, Ali, & Taieb, 2016; Mennes & van Gulik, 2020).

The term used for the resolution of lexical ambiguity is known as word sense disambiguation (WSD) which improves the performance of the tasks that are sensitive to lexical such as QA. This WSD is an intermediate task which does not end in itself but rather requires the processing at each level to complete the NLP task. The details of WSD methods are thoroughly discussed in Chapter 2.

This study mainly focuses on the resolution of the lexical ambiguity associated with NL questions by introducing a knowledge-based sense disambiguation method that comprises a dictionary-based and an AI-based method. Moreover, this KSD method also integrates the context knowledge and specific domain ontology into a shallow NLP. WordNet 3.1 provides the context knowledge based on surrounding entities in the question, while lexicon knowledge is formalized in ontology. The shallow NLP-based processor includes syntactic and semantic processing. This proposed method is used in

the current study for the evaluation of the QA system prototype developed for the pilgrimage domain.

1.3 STATEMENT OF THE PROBLEM

The emerging field of data extraction using intelligent methods for solving the problem of lexical ambiguity is showing quite promising results. However, Natural Language Processing (NLP) is still facing several challenges to determine the precise meaning of various ambiguities. Besides, the extraction of accurate information from a knowledge base (KB) remains challenging in question answering (QA) systems. The main reason behind their failure is the incapability to deal effectively with semantic issues in the natural language (NL) question processing module (Dimitrakis et al., 2020; Jabalameli, Nematbakhsh, & Zaeri, 2020; Y. Jiang, 2020; Rodrigo & Peñas, 2017).

Most QA systems transfer the user' NL questions into formal representations or structured queries. One of the essential challenges in this process of conversion is the difference between the phrases posed by users and the vocabulary deployed in KBs which could have more than one meaning (Mennes & van Gulik, 2020). Determining the accurate semantic representation of the question keywords is a significant task which impacts further processing modules such as document processing and answer extraction.

A close look at the relevant literature reveals that many scholars (Navigli, 2018; Raganato, Camacho-collados, & Navigli, 2017) report positive outcomes to a lexical solution on the topic of word sense disambiguation (WSD). In addition, several studies (Jabalameli et al., 2020; Mohammed, Shi, & Lin, 2018) have appraised this issue in the context of QA, but only a few have examined the role of WSD in returning potential

answers. Moreover, challenging questions may require handling ambiguities based on certain conditions, such as seeking instructions or advice in a critical situation (Pillai et al., 2018; Y. Wang et al., 2020).

Literature has shown that a restricted domain such as a pilgrimage domain where pilgrims are particularly reliant on information that is accessible on the internet. The diverse data availability in various KB formats introduces several challenges to infer concise accurate information corresponding to pilgrim questions. This requires a user to be familiar with the structures of KB and use a formal query language for the system to effectively understand the query (del Carmen Rodrıguez-Hernández, Ilarri, Trillo-Lado, & Guerra, 2016; Jabalameli et al., 2020; White, Richardson, & Yih, 2015). Hence, there is a need to develop an effective QA approach to bridge the gap between pilgrim's expressiveness and the formal representation of the KB to ensure the provision of accurate answers.

The following are summery of the problem statement in points:

- The major challenge faced by question processing module is the usage of a natural language that is full of ambiguity.
- The lexical semantic issue is mainly arising due to the difference between the KB representation of the terms and the information expressed by the input question which leads to irrelevant answers or no answer at all.
- Extracting the lexical semantic of a NL question presents challenges at syntactic and semantic levels for most QA systems.
- The diverse data availability in various KB formats introduces several challenges to infer concise accurate information corresponding to pilgrim questions.

1.4 RESEARCH OBJECTIVES

The research aims to achieve the following objectives:

1. To model a knowledge-based sense disambiguation method that is capable of determining the correct sense of the ambiguous words associated with NL questions.
2. To evaluate the proposed method in terms of accuracy word sense disambiguation performance.
3. To develop a model of a QA application prototype that determines the intended meaning of NL questions expressed by pilgrims and gives appropriate responses.
4. To evaluate the effectiveness of the proposed solution in a QA application based on accuracy performance.

1.5 RESEARCH QUESTIONS

To accomplish the objectives of this study, there are four research questions:

1. What is the most suitable method that can be used to resolve the problem of lexical ambiguity associated with NL questions?
2. What is the performance of the proposed method for word sense disambiguation in terms of accuracy?
3. How can the proposed method be developed into a QA application prototype for a pilgrimage domain?
4. What is the accuracy performance of the proposed approach in a QA application?

1.6 RESEARCH SCOPE

This research aims to resolve the problem of ambiguity limited to lexical type that arises in NL questions for restricted domain. The proposed approach is applied to a pilgrimage QA application in the English language only. Users are enabled to pose query in the form of WH-questions only through their android smartphones' natural language user interface (UI). In this research the answers are structured in a repository based on the semantic of the questions. The research is limited to the use of WordNet and domain ontology as resources of context and lexicon knowledge bases which may lead to low word sense disambiguation performance due to the limited definition and relations coverage.

1.7 SIGNIFICANCE OF THE RESEARCH

The challenges in handling lexical ambiguity have become increasingly important in the NLP community and related fields (Bevilacqua, Pasini, Raganato, & Navigli, 2021; Mennes & van Gulik, 2020). They may impact the performance of various related technologies such as machine translation, information extraction and question answering (Ben Abacha & Demner-Fushman, 2019; Y. Wang et al., 2020). This study will examine the significance of handling lexical ambiguity arises in the user's natural language (NL) question to infer accurate response in QA systems. The major contribution of this work lies in the formulation of a new knowledge-based sense disambiguation (KSD) method that aims at resolving the problem of lexical ambiguity occurs in NL question posed to a QA system. This algorithm is designed by incorporating question's metadata, context and lexicon knowledge resources into a