MULTILINGUAL DYNAMIC LINKING OF WEB RESOURCES

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

UĞUR DÖNMEZ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

SEPTEMBER 2014

Approval of the thesis:

MULTILINGUAL DYNAMIC LINKING OF WEB RESOURCES

submitted by **UĞUR DÖNMEZ** in partial fulfillment of the requirements for the degree of **Master of Science in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen ______________
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. Adnan Yazıcı ______________
Head of Department, **Computer Engineering**

Prof. Dr. Ahmet Çoşar ______________
Supervisor, **Computer Engineering Department, METU**

Assist. Prof. Dr. Yeliz Yeşilada ______________
Co-supervisor, **Computer Engineering Dept., METU NCC**

Examining Committee Members:

Prof. Dr. Adnan Yazıcı
Computer Engineering Department, METU ______________

Prof. Dr. Ahmet Çoşar
Computer Engineering Department, METU ______________

Prof. Dr. İsmail Hakkı Toroslu
Computer Engineering Department, METU NCC ______________

Assist. Prof. Dr. Yeliz Yeşilada
Computer Engineering Department, METU ______________

Assist. Prof. Dr. Enver Ever
Computer Engineering Department, METU NCC ______________

Date: ______________

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: UĞUR DÖNMEZ

Signature :

ABSTRACT

MULTILINGUAL DYNAMIC LINKING OF WEB RESOURCES

Dönmez, Uğur
M.S., Department of Computer Engineering
Supervisor : Prof. Dr. Ahmet Çoşar
Co-Supervisor : Assist. Prof. Dr. Yeliz Yeşilada

September 2014, 121 pages

The World Wide Web is successful for locating, browsing and publishing information by its scalable architecture. However, the Web suffers from some limitations. For example, links on the Web are embedded in documents. Links are only unidirectional, ownership is required to place an anchor in documents, and authoring links is an expensive process. The embedded link structure of the Web can be improved by Semantic Web. By using Semantic Web components, existing Web resources can be enriched with additional external links. COHSE is a project that aims to solve this problem. Although COHSE is successful in dynamic linking, it is monolingual. In our work, we aim to solve this problem. MOHSE is a project that dynamically link Web resources in different languages. In order to achieve our goal, firstly, we researched multilingual controlled vocabularies, their representation, and their use in dynamic linking. Secondly, we investigated existing multilingual controlled vocabularies. Thirdly, we extended COHSE infrastructure to support multilingualism. Finally, we performed experiments with this extended infrastructure. Outcomes show that MOHSE is useful and increases user's Web performance. This thesis contributes to multilingualism on the Web. MOHSE is the prior work that supports multilingual dynamic linking.

Keywords: Dynamic Linking, Multilingualism, Information Retrieval, Semantic Web

ÖZ

WEB KAYNAKLARININ ÇOK DİLLİ DİNAMİK BAĞLANMALARI

Dönmez, Uğur
Yüksek Lisans, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi : Prof. Dr. Ahmet Çoşar
Ortak Tez Yöneticisi : Yrd. Doç. Dr. Yeliz Yeşilada

Eylül 2014 , 121 sayfa

Dünya Çapında Ağ ölçeklenebilir mimarısı ile bilginin yerleştirilmesi, taranması ve yayınlanmasında başarılıdır. Fakat Web giderilmesi gereken bazı limitler içerir. Örnek olarak Web'deki bağlantılar dokümanlanların içine gömülmüştür. Bağlantılar sadece tek taraflı çalışmaktadır. Ayrıca kaynaklara bağlantı eklenmek için o kaynağın sahibi olmanız gerekmektedir. Anlam bilimsel ağ ile Web'in gömülü bağlantı yapısı geliştirilebilir. Anlam bilimsel ağ bileşenleri ile Web kaynakları harici bağlantılar eklenerek zenginleştirilebilir. COHSE bu problemi çözmeyi amaçlayan bir projedir. COHSE dinamik bağlamada başarılı olsa da, tek dillidir. MOHSE Web kaynaklarını farklı dillerde bağlayan bir projedir. Amacımıza ulaşmak için, ilk olarak, çok dilli kontrollü kelimeleri, simgelenmelerini ve dinamik bağlamada nasıl kullanılacaklarını araştırdık. İkinci olarak, var olan kontrollü kelime gruplarını inceledik. Üçüncü olarak, COHSE altyapısını çok dilliliği destekleyecek şekilde geliştirdik. Son olarak, geliştirdiğimiz altyapı ile deneyler yaptık. Sonuçlar MOHSE'nin kullanışlı olduğunu ve kullanıcıların Web kullanım performansını arttırdığını gösterdi. Bu tez Web'deki çok dilliliğe katkıda bulunmaktadır. MOHSE çok dilli dinamik bağlamayı destekleyen öncü çalışmadır.

Anahtar Kelimeler: Dinamik Bağlama, Çok Dillilik, Bilgi Edinme, Anlam Bilimsel Ağ

To my family and friends

ACKNOWLEDGMENTS

I would like to express my sincere appreciations to my advisor Assist. Prof. Dr. Yeliz YEŞİLADA for her guidance, support, encouragement and positive attitudes throughout my study. I am grateful for having chance of working with her. I would like to thank a number of individuals who have helped me complete this master's thesis. Metu Ceng department head Prof. Dr. Adnan YAZICI and and Prof. Dr. Ahmet ÇOŞAR helped me a lot while I was in METU-NCC. I am grateful Assist. Prof. Dr. Enver EVER who helped me morally. I am grateful to Prof. Dr. Ali TAŞIRAN who helped me for evaluating my statistical data.

I am very grateful to every people I know during my research assistantship in METU-NCC. Şükrü ERASLAN was always with me when I needed some help. Abraham and Ryan prepared questions for my evaluation task. It was very important for me.

I want to thank my friends who supported me morally and helped me a lot. Also, I am very grateful to Sercan YÜKSEK for reviewing my writings. I want to thank Selin for reviewing my writing, doing my pilot evaluation, and entering the evaluation data to computer. I want to thank Elgin for supporting me about SPSS, and getting permission from METU ethical community. My dear friend İlkcan provided invaluable support for me. I would like to thank to him a lot. Additionally, I would like to thank the guys in Smyrna Bistro who provided their computers and canteen for my evaluation process.

I would like to thank Dream Theater who plays perfect songs which motivated me a lot while I was working.

I would like to thank all the participants who attended my evaluation study.

I am very grateful that my family encouraged me throughout the study.

TABLE OF CONTENTS

CHAPTERS

APPENDICES

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

CLIR	Cross Language Information Retrieval
CLLD	Cross-lingual Link Discovery
COHSE	Conceptual Open Hypermedia Service
DLS	Distributed Link Service
EU	European Union
KOS	Knowledge Organization System
GATE	General Architecture for Text Engineering
GEMET	The General Multilingual Environmental Thesaurus
KIM	Knowledge and Information Management
KS	Knowledge Service
LIR	Linguistic Information Repository
LSI	Latent Semantic Index
MOHSE	Multilingual Open Hypermedia Service
MONNET	Multilingual Ontologies for Networked Knowledge
NTCIR	National Institute of Informatics Testbeds and Community for Information Access Research
OWL	Web Ontology Language
RDF	Resource Description Framework
RS	Resource Service
SKOS	Simple Knowledge Organization System
SW	Semantic Web
SWB	Semantic Web Browser
WEB	World Wide Web
W3C	World Wide Web Consortium

CHAPTER 1

INTRODUCTION

The World Wide Web is one of the greatest innovations of the 20th century. It is very successful for locating (URLs), browsing (HTTP) and publishing (HTML) information by its scalable architecture. However, Web suffers from a number of limitations. The Web is the concept of linking resources. Typically, links on the Web are embedded within documents. Although embedding links is a simple and supporting Web scalable architecture, it is also one of the limitations of the Web. Links can only be unidirectional, ownership is required to place an anchor in documents, sometimes authoring links is an expensive process, etc. To solve these limitations Semantic Web [8] has been proposed. With the advances in the Semantic Web, the embedded link structure of the Web can be improved and extended. Additional Semantic Web components can be used to add links dynamically between resources by using the existing infrastructure of the Web. A key drive for Semantic Web advances is to improve machine processing on the Web, however including semantic for machine-processing can also be used to improve the linking and navigation of Web pages intended for human end-users.

Another problem of the Web is that it is very difficult to find the relevant information in the really wide web. This problem can also be solved by using of Semantic Web technologies. Semantic Web allow users to create smart and enriched queries for finding relevant information.

Web resources can be annotated with semantic mark-up using knowledge representaion languages, such as Resource Description Framework (RDF) [11] or Web Ontology Language (OWL) [4]. RDF-Schema provides a standard way of defining vocabularies. With semantic annotation, two important task of the semantic web can be achieved: (i) extracting and hyper linking named entities and (ii) finding relevant documents in accordance with entities [42].

COHSE [5, 30] is a project that aims to solve the problems given above. In order to achieve that COHSE uses Semantic Web technologies, in particular semantic annotation and knowledge resources such as controlled vocabularies and ontologies. COHSE provides different hypertext views of Web documents according to the choice of domain ontologies or vocabularies. To illustrate, Bechhofer et al (2005) [3] shows how COHSE is used to link biological Web documents by using Gene Ontology, Carr et al (2004) [15] presents how Sun's Java tutorial pages can be augmented by using an ontology that describes Java and object oriented programming. Yesilada et al (2008) [78] present how COHSE is used to dynamically link independent sub-sites of SUN Microsystems, and Bechhofer et al (2008) [6] explain how COHSE is used to dynamically link Web resources about infectious diseases as part of a life science project.

Although it has been demonstrated that COHSE is successful in linking Web resources, it is unfortunately monolingual. Multilingualism is one of the challenges for the Semantic Web [7] thus, the aim of the MOHSE project is to dynamically link Web resources in different languages. Moreover, European Commission promotes multilingualism on the Web to preserve linguistic diversity in Europe [1]. Therefore, this project will make significant contribution to promote multilingualism on the Web and is closely related to the European Commission's multilingualism objective. To achieve our objective, MOHSE has the following aims:

1. Research the state of the art multilingual controlled vocabularies, their representation, and their use in dynamic linking of Web resources.

2. Investigate existing multilingual controlled vocabularies that include terms from

[1] European Commission Multilingualism, http://ec.europa.eu/education/languages/

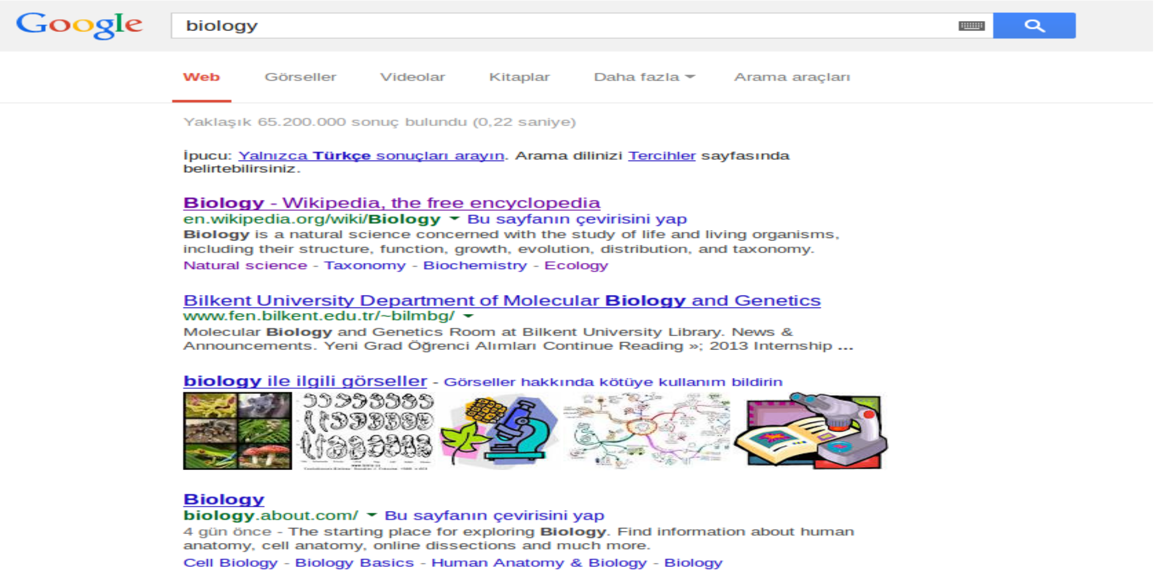
both Turkish and English.

3. Extend COHSE infrastructure to dynamically link Web resources in Turkish and English.

4. Experiment with this extended infrastructure to demonstrate how multilingual dynamic linking can be achieved.

The core role of MOHSE Semantic Web Application is to enable user interpret third-party materials on the Web which is related to selected ontology and expand their knowledge on the topic when they search something on the Web. In brief, MOHSE enriches English Web resources with Turkish resources.

Here we explain how MOHSE works with an example scenario. Assume that user wants to learn about biology and searches it on the Web. When the user searches it with a typical search engine such as Google or so, the first result shown is the Biology page on Wikipedia as shown in the Figure 1.1. The Wikipedia page is full of biological terms and if the user is not familiar with them, it is hard to understand the context (see the Figure 1.2). By using MOHSE biology related terms can be annotated as shown in the Figure 1.3. The user can easily click annotated terms and find more about these terms as shown in the Figure 1.4.



Yaklaşık 65.200.000 sonuç bulundu (0,22 saniye)

İpucu: Yalnızca **Türkçe** sonuçları arayın. Arama dilinizi Tercihler sayfasında belirtebilirsiniz.

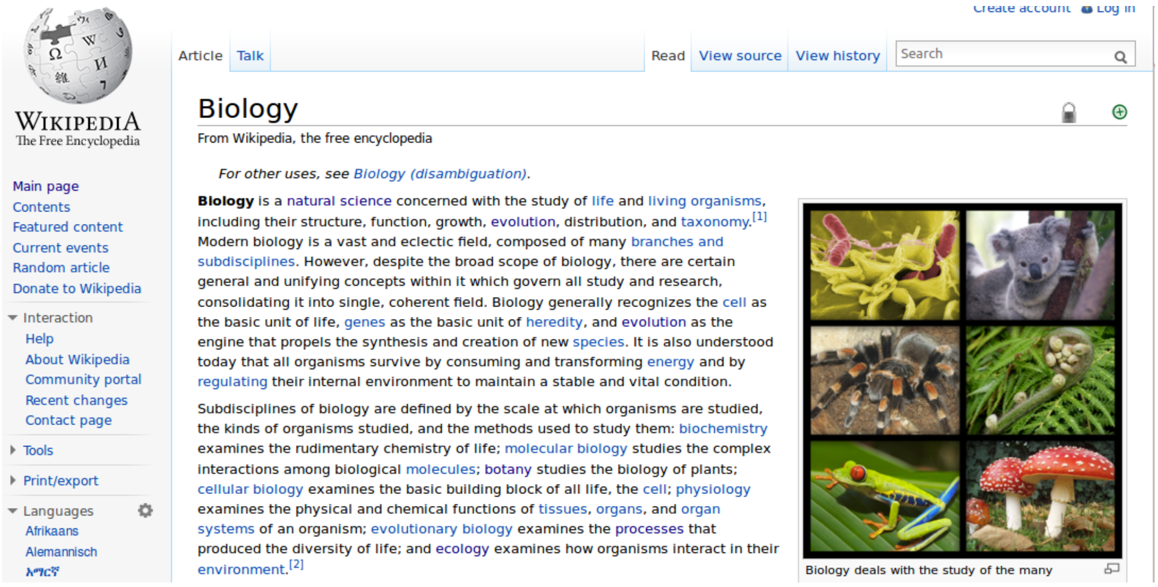
Biology - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Biology** Bu sayfanın çevirisini yap
Biology is a natural science concerned with the study of life and living organisms, including their structure, function, growth, evolution, distribution, and taxonomy.
Natural science - Taxonomy - Biochemistry - Ecology

Bilkent University Department of Molecular **Biology** and Genetics
www.fen.bilkent.edu.tr/~bilmbg/
Molecular **Biology** and Genetics Room at Bilkent University Library. News & Announcements. Yeni Grad Öğrenci Alımları Continue Reading »; 2013 Internship ...

biology ile ilgili görseller - Görseller hakkında kötüye kullanım bildirin

Biology
biology.about.com/ Bu sayfanın çevirisini yap
4 gün önce - The starting place for exploring **Biology**. Find information about human anatomy, cell anatomy, online dissections and much more.
Cell Biology - Biology Basics - Human Anatomy & Biology - Biology

Figure 1.1: Google Search Results When Searching Biology



Biology

From Wikipedia, the free encyclopedia

For other uses, see Biology (disambiguation).

Biology is a natural science concerned with the study of life and living organisms, including their structure, function, growth, evolution, distribution, and taxonomy.[1] Modern biology is a vast and eclectic field, composed of many branches and subdisciplines. However, despite the broad scope of biology, there are certain general and unifying concepts within it which govern all study and research, consolidating it into single, coherent field. Biology generally recognizes the cell as the basic unit of life, genes as the basic unit of heredity, and evolution as the engine that propels the synthesis and creation of new species. It is also understood today that all organisms survive by consuming and transforming energy and by regulating their internal environment to maintain a stable and vital condition.

Subdisciplines of biology are defined by the scale at which organisms are studied, the kinds of organisms studied, and the methods used to study them: biochemistry examines the rudimentary chemistry of life; molecular biology studies the complex interactions among biological molecules; botany studies the biology of plants; cellular biology examines the basic building block of all life, the cell; physiology examines the physical and chemical functions of tissues, organs, and organ systems of an organism; evolutionary biology examines the processes that produced the diversity of life; and ecology examines how organisms interact in their environment.[2]

Biology deals with the study of the many

Figure 1.2: Biology Wikipedia Page

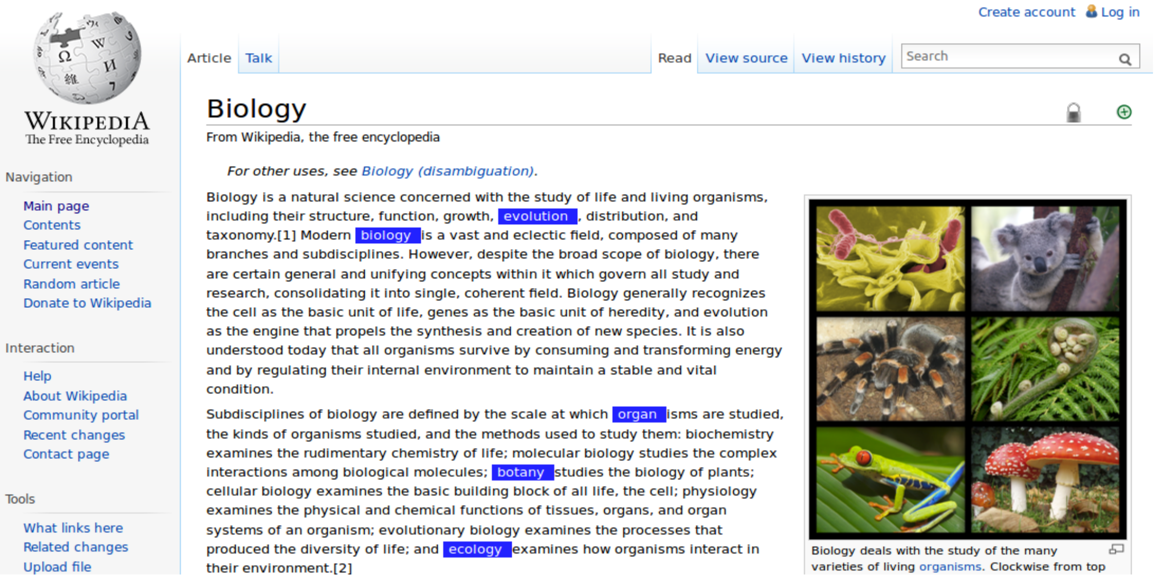
1.1 Contributions

The most important contribution of MOHSE is multilingualism on the web. MOHSE's first aim is to dynamically link Web resources in different languages. In order to achieve this purpose, MOHSE uses multilingual thesaurus, multilingual dynamic linking and text translation techniques. Although there are a lot of Semantic Web browsers (we cover them in Section 3.2), they do not contribute anything for multilingualism.

Semantic Browsers are well studied but there is no enough effort for evaluating them. In this thesis, we conduct detailed user survey for evaluating MOHSE. Our user survey will give significant contribution for evaluation of Semantic Web browsers.

1.2 Thesis Outline

The outline of this thesis is organized as follows. Chapter 2 presents an overview of the related technologies to this thesis. In particularly, it gives an overview of Semantic Web. In Chapter 3 related work about the thesis is explained. It covers multilingual



Biology

From Wikipedia, the free encyclopedia

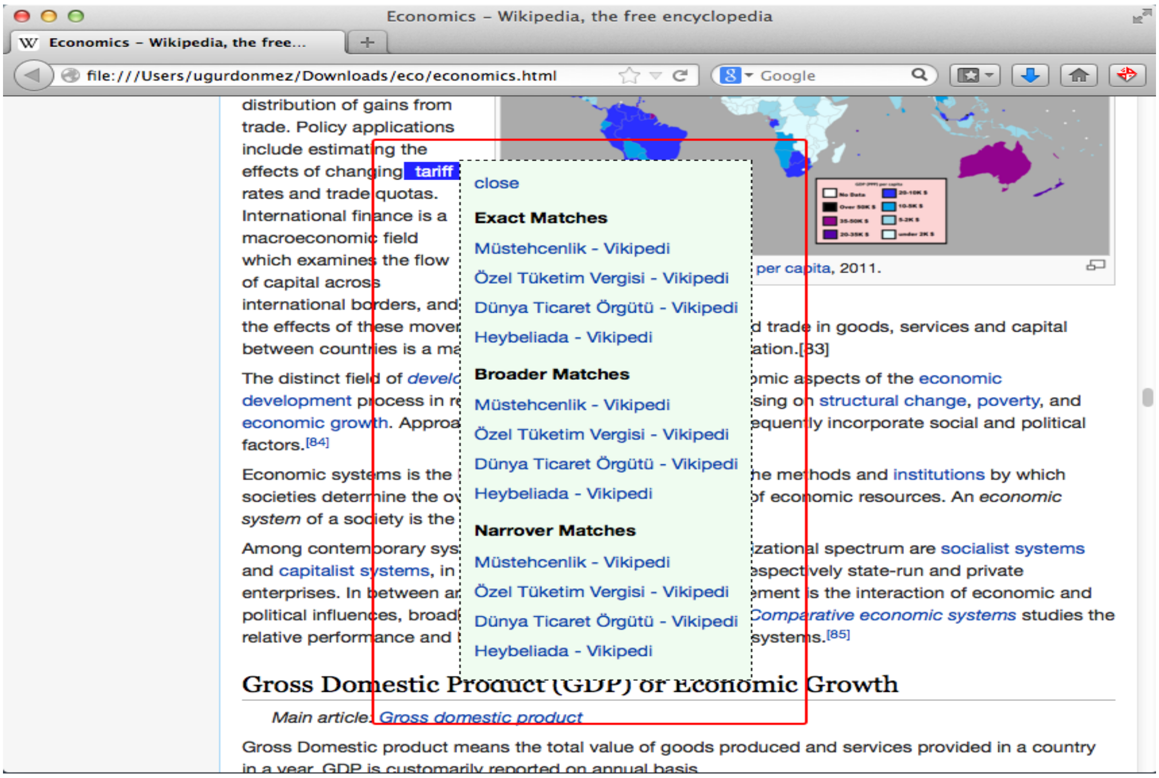
For other uses, see Biology (disambiguation).

Biology is a natural science concerned with the study of life and living organisms, including their structure, function, growth, evolution , distribution, and taxonomy.[1] Modern biology is a vast and eclectic field, composed of many branches and subdisciplines. However, despite the broad scope of biology, there are certain general and unifying concepts within it which govern all study and research, consolidating it into single, coherent field. Biology generally recognizes the cell as the basic unit of life, genes as the basic unit of heredity, and evolution as the engine that propels the synthesis and creation of new species. It is also understood today that all organisms survive by consuming and transforming energy and by regulating their internal environment to maintain a stable and vital condition.

Subdisciplines of biology are defined by the scale at which organ isms are studied, the kinds of organisms studied, and the methods used to study them: biochemistry examines the rudimentary chemistry of life; molecular biology studies the complex interactions among biological molecules; botany studies the biology of plants; cellular biology examines the basic building block of all life, the cell; physiology examines the physical and chemical functions of tissues, organs, and organ systems of an organism; evolutionary biology examines the processes that produced the diversity of life; and ecology examines how organisms interact in their environment.[2]

Biology deals with the study of the many varieties of living organisms. Clockwise from top

Figure 1.3: Enriched Biology Page by MOHSE


Economics – Wikipedia, the free encyclopedia
Economics – Wikipedia, the free...
file:///Users/ugurdonmez/Downloads/eco/economics.html
Google
No Data
Over 50K $
35-50K $
20-35K $
20-10K $
10-5K $
5-2K $
under 2K $

distribution of gains from trade. Policy applications include estimating the effects of changing tariff rates and trade quotas. International finance is a macroeconomic field which examines the flow of capital across international borders, and the effects of these mover

between countries is a ma

per capita, 2011.

d trade in goods, services and capital

ation.[83]

The distinct field of develo

development process in re

economic growth. Approa

factors.[84]

omic aspects of the economic

sing on structural change, poverty, and

equently incorporate social and political

Economic systems is the

societies determine the ov

system of a society is the

he methods and institutions by which

of economic resources. An *economic*

Among contemporary sys

and capitalist systems, in

enterprises. In between ar

political influences, broad

relative performance and

zational spectrum are socialist systems

espectively state-run and private

ement is the interaction of economic and

Comparative economic systems studies the

systems.[85]

close

Exact Matches

Müstehcenlik - Vikipedi

Özel Tüketim Vergisi - Vikipedi

Dünya Ticaret Örgütü - Vikipedi

Heybeliada - Vikipedi

Broader Matches

Müstehcenlik - Vikipedi

Özel Tüketim Vergisi - Vikipedi

Dünya Ticaret Örgütü - Vikipedi

Heybeliada - Vikipedi

Narrover Matches

Müstehcenlik - Vikipedi

Özel Tüketim Vergisi - Vikipedi

Dünya Ticaret Örgütü - Vikipedi

Heybeliada - Vikipedi

Gross Domestic Product (GDP) or Economic Growth

Main article: Gross domestic product

Gross Domestic product means the total value of goods produced and services provided in a country in a year. GDP is customarily reported on annual basis

Figure 1.4: MOHSE Link Box

dynamic linking, Semantic Web browsers, text retrieval, and multilingual vocabulary mapping. In Chapter 4, architecture of COHSE is examined and detailed architecture of MOHSE is given. In Chapter 5, evaluation process of MOHSE, and evaluation results are presented. Chapter 6 concludes the thesis and covers future work.

CHAPTER 2

BACKGROUND

In this chapter, we give an overview of the background work for this thesis. Section 2.1, we include information about Semantic Web and Semantic Web technologies. In section 2.2, we cover multilingual controlled vocabularies, and SKOS and OWL. In section 2.3, we examine text retrieval methods and algorithms. Finally, in Section 2.4 we provide an overview of thesaurus. In Section 2.5, we summarize the chapter.

2.1 Semantic Web

Berners who is the creator of Semantic Web, et al. (2001) [8] stated that most of the Web's content today is designed for humans to read, not for the computer program to manipulate meaningfully. Computers can adeptly parse Web pages for layout and routine processing. In general computers have no reliable way to process semantics. However, the Semantic Web brings structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users. Semantic Web is an extension of current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation.

Jupp et al. (2008) [39] stated that two key Semantic Web technologies endorsed by the World Wide Web Consortium (W3C) [1] are the Resource Description Framework

[1] http://w3.org

(RDF) [2] and the Web Ontology Language (OWL). RDF provides a base vocabulary for describing resource and ad-hoc relationship between them. OWL is an extension of RDF and provides a vocabulary for building ontologies. Ontologies are used to encode knowledge about particular domain in the form of the entities and the relationship between them. An ontology language like OWL has well defined semantics, which facilities computational interpretation of statements. Ontologies provide the content to annotate documents on the web. This Semantic Markup provides a mechanism for computers to interpret a document's contents. Whilst ontologies offer great promise, the library and information sciences have a history of using knowledge artefacts, known as Knowledge Organization Systems (KOS), to classify and index documents [65]. The KOS provide support for document retrieval and navigation applications. The KOS can vary from simple dictionaries to more complex structures like thesauri and controlled vocabularies, which introduce structure to the knowledge in the form of associative relationships between concepts. The W3C have published the Simple Knowledge Organisation System (SKOS) [47], a standard vocabulary for representing KOS like structures. SKOS has a serialisation into RDF that facilitates the use of SKOS in Semantic Web applications. OWL and SKOS may appear similar at first, but have both been developed to fulfill different purposes. Choosing which to use depends largely on application requirements. We choose to use SKOS for thesis because SKOS support multilingualism and support to navigate through relationship according to related, broader and narrower. Moreover, GEMET which is multilingual thesaurus we are using, can easily be converted to SKOS format instead of OWL. Therefore, we decided to use SKOS format in MOHSE.

Web resources can be annotated with semantic mark-up using knowledge representaion languages, such as RDF [11] or OWL [4]. RDF-Schema provides a standard way of defining vocabularies. With semantic annotation, two important task of the semantic web can be achieved: (i) extracting and hyperlinking named entities and (ii) finding relevant documents in accordance with entities [42]. Therefore, vocabularies which are represented by OWL or SKOS, can use MOHSE to annotate Web content.

[2] http://w3.org/RDF

WWW is, as the name suggests, a hypothetical Web linking resources from every single server in the world. Since links are dynamic, they are stored within resources. Beside being a simple and scaleble approach, embedding links has some limitations too: links can only be unidirectional, ownership is required to place an anchor in a document, sometimes authoring links is an expensive process, etc [38]. With stated advances in the Semantic Web, the embedded link structure of the Web can be improved and extended. Components of Semantic Web can be utilized to add dynamic links between resources by using the existing infrastructure of the Web. A key feature of Semantic Web advances is to improve machine processing on the Web, yet it is still not used to link and navigate the Web pages intended for human end-users from different linguistic background.

2.2 Multilingual Controlled Vocabularies

There are variety of different styles of knowledge representation, ranging from formal ontologies to taxonomies and controlled vocabularies [6]. Ontologies are seen as a cornerstone of the Semantic Web [34]. In brief, an ontology provides a machine-interpretable description of a domain interest. The languages used to represent what called as ontologies (and thus the models themselves) vary considerably, and can range from simple taxonomies or hierarchies through to rich, formal, logic based languages such as OWL [4]. Increased formality (in terms of precise semantics) can remove ambiguity in the representation and facilitate the use of machine processing. The Semantic Web community has focused a great deal of effort on the standardization of language for representation. RFD and OWL are W3C recommendations as stated before. There are now a huge number of published ontologies using these representations: one could argue that OWL is the most successful knowledge representation language we have ever seen (in term of content size). At the other end of the problem, there are a large number of vocabularies around which are not intended to be formal ontologies, but rather are controlled vocabularies to be used for annotation, information retrieval or organization of information resources. These sources often contain the kinds of information that we believe are useful when presenting information to human users (as opposed machines) - for example, synonyms, lexical variants and ”scope notes” or definitions providing valuable context for a human reader. COHSE’s knowledge sources on a thesaurus model is likely to provide a better ”fit” with the navigation models that we wish to support [6]. SKOS provides a standardised collection of relationship (broader/narrower/related) which can be used by our application [47]. Although these relationships may not have precise semantics that come with OWLs super and sub-class relationships, in this context, the looser interpretation is, in fact, more appropriate to the task in hand.

2.3 Text Retrieval

The goal of text retrieval system is to present the user with a set of items that will satisfy his or her information need. Information need can be represented as query and items, which are found according to the query, called documents [55]. In our system MOHSE, query is anchors found in the Web resources, and documents are links added to these anchors.

In literature, two types of text retrieval systems are used: (i) exact match and (ii) ranked retrieval [55]. Exact match text retrieval systems provide an unranked set of documents which satisfy user’s query. Most existing text retrieval systems fall into this category. In MOHSE we also used exact match method because using exact match is easier than the ranked retrieval. In ranked retrieval, system attempts to impose a total order on the documents in such a way that the most useful documents are near the top of the list. There are three types of ranked retrieval system: (i) ranked boolean, (ii) probabilistic, and (iii) similarity based. Ranked retrieval is out of MOHSE scope; therefore, we will not go further to examine this topic.

2.3.1 Multilingual Text Retrieval

In this section, we provide information about multilingual text retrieval because we are trying to add Turkish links to English Web resources; therefore, we need to examine multilingual text retrieval.

According to Oard et al (1998) [55] there are three main approaches emerged in the literature: (i) text translation, (ii) thesaurus based approaches and (iii) corpus based approaches. The former two approaches are discussed below and the latter one is discussed in Section 3.3.

Text Translation

Text Translation is most straightforward approach to multilingual text retrieval. Although Turkish English text translation methods were examined [73, 26, 56, 76], we did not use text translation in MOHSE because they are able to produce high quality translation only in limited domains. For increasing success rate of translation, we should create controlled vocabulary and grammar for both English and Turkish, however these control make translation process difficult.

Multilingual Thesauri

Thesaurus based techniques have some advantages. Thesauri can represent relationship between terms and concepts in a way that humans find understandable, thesaurus based text retrieval allows users to exploit better queries. Moreover, domain knowledge can be encoded in the thesaurus. Thesaurus can be used for formulating better queries in multilingual retrieval by using ontology knowledge included in thesaurus. Query can be enriched by using ontology knowledge. Because of these benefits, we decided to use thesaurus based techniques in MOHSE.

2.4 Thesaurus

AGROVOC [3] is multilingual thesauri developed and maintained by the Food and Agriculture Organization (FAO) of the United Nations. It contains more than 40,000

[3] AGROVOC, http://aims.fao.org/standards/agrovoc/about

concepts in up to 22 languages covering topics related to food, nutrition, agriculture, fisheries, forestry, environment and other related domains. The GEneral Multilingual Environmental Thesaurus (GEMET)[4] is another large thesaurus which has been developed by European Topic Centre on Catalogue of Data Sources (ETS / CDS). The overall goal of GEMET is to create general terminology for the environment. GEMET consists of a number of themed multilingual vocabularies which are available in 27 languages and with more than 6,000 descriptions. GEMET is available online for browsing and can be extracted in different formats such as SKOS for public use. GEMET provides support for various themes such as agriculture, biology, chemistry, energy, food, human health, pollution, tourism, etc. The current version of GEMET assures a complete numerical equivalence in some languages such as Basque, Bulgarian, Dutch, Finnish, French, German, Hungarian, Italian, Norwegian, Portuguese, Russian, Slovenian and Spanish. However, some descriptions are still not complete for the languages such as Danish, Slovak, Swedish and Greek, and they are still under development. Although we could not find much information about Turkish support in GEMET, we could see on their Website that GEMET is also available in Turkish. Since GEMET can easily be downloaded from in both English and Turkish in SKOS format which is supported by MOHSE, we have decided to use GEMET in our experiments.

2.5 Summary

In this chapter, firstly we describe Semantic Web technologies which are used in our thesis. Semantic Web allows us to increase machine process on web resources. Web resources can be enriched by ontology knowledge included in SKOS or OWL vocabularies by using Semantic Web technologies. Secondly we describe multilingual controlled vocabularies and how they can be used in this work. Thirdly, text retrieval methods, which are used in MOHSE are described. Lastly, we make short introduction about thesaurus and we give brief information about GEMET used in our thesis.

[4] Eionet GEMET Thesaurus ,http://www.eionet.europa.eu/gemet/

CHAPTER 3

RELATED WORK

In this chapter, we provide detailed discussion on related work about our thesis. In section 3.1, we give information about multilingual dynamic linking. In section 3.2, we examine current Semantic Web browsers in detail and compare them. In section 3.3, we research about multilingual text retrieval methods. In section 3.4, we include information about multilingual vocabulary and ontology mapping. In section 3.5, we describe multilingual translation. Section 3.6 covers thesaurus.

3.1 Multilingual Dynamic Linking

The Web can be considered as a closed hypermedia system since the links are embedded into pages [46]. Whereas an open system in hypermedia simply is the one where the reader is given the same access as the author [22]. Open Hypermedia Systems are well researched by the hypermedia community where a number of systems have been developed including MicroCosm, Chimera and Devise Hypermedia [58]. Several systems have also been introduced to provide an open hypermedia system on the Web including DLS, DHM/WWW, Webvise and Arakne Environment [10]. Using these systems, readers can create links and other hypermedia structures on top of arbitrary Web pages; can share these links with others through the use of external linkbases. Conceptual Hypermedia Systems specify hypertext structure and behavior in terms of a well-defined conceptual scheme. Since constructing hypertext links manually is inconsistent and error-prone, these systems sought to bring conceptual models to bear in order to support linking. Although conceptual hypermedia systems address the problem of hard-coding and hand-crafting links, systems such as MacWeb [53] and SHA [19] provided closed systems rather than supporting in the wide (or wider) of the Web. We can also find some example of the conceptual hypermedia systems on the Web, for example Semantic Wikis (e.g., Platypus Wiki [1]). They introduce explicit models of knowledge in support of management of content. This combines hypertext and meta-data, but is again primarily concerned with organizing and managing resource under a single control point. Compared to the systems above we can say that COHSE marries open and conceptual hypermedia systems by using the Web infrastructure. According to Bechhofer et al (2008) [6] Conceptual Open Hypermedia Systems treats links separately from document and using conceptual model such as ontologies guide the navigation. This approach allows the construction of hypertext link structures by using knowledge representation such as ontologies. Therefore using ontologies for dynamic linking is possible as done in COHSE [5, 30]. Although MOHSE and COHSE have features in common, the most important aim of MOHSE is to dynamically link Web resources in different languages.

Internationalization is an important topic on the Web[2]. The overall objective of internationalization is to ensure the Web resources are available in different languages for different cultures. There are a number of projects such as MultilingualWeb [3] that aims to promote standards and best practices for the multilingual Web. There has been also significant work in understanding the requirement of developing and maintaining Web sites in different languages [35]. Creating bilingual Web sites may not be so difficult, however when one considers more than two different languages, design and development can become challenging. These challenges can be both technical such as how to maintain the same content in more than one language [52, 69, 70] and cultural such as how to ensure that the content is suitable to different cultures in different languages [33, 66]. Although these would be interesting to investigate, they are beyond the scope of this thesis. The focus of our thesis is about dynamically linking Web resources that already exist in different languages. These resources can belong to the same organization or they can belong to third party. When we look at

[1] http://platypuswiki.sourceforge.net/

[2] W3C Internationalization Activity, http://www.w3.org/International/

[3] MultilingualWeb: Standards and best practices for the Multilingual Web. http://www.multilingualweb.eu/

the literature, there are quite a lot of work for designing and maintaining multilingual Websites, but we cannot find any work on dynamic linking of multilingual Web resources. Although there are similar work on dynamically linking Web resources in English, none of them explores how Web resources can be linked in different languages.

3.2 Semantic Web Browsers

The term Semantic Web Browser (SWB) refers to any browser which: (i) uses at least one KOS to support browsing, (ii) is able to identity and highlight useful terms in the Web page being visited, (iii) enables the semantic interpretation of these Web pages and adds semantic hyperlinks to their highlighted terms, (iv) gathers additional information from the highlighted terms, which may involve access to external services [24]. Toward this end, several semantic web browser have been developed. In this section we investigate these browsers and we compare them with MOHSE. Table 3.1 shows the comparison of these browsers. We compare them based on the following features: (i) Ontology selection methods of SWB's differ from each other. KIM and PERSON use generic ontology which user can not change. GATE and SemWeb do not use any ontology. Other browsers allow users to select ontology according to context. (ii) Although GATE, Piggy Bank, Magpie, and Power Magpie create framework for SWB, others do not. (iii) Sealife Browser, KIM, COHSE, and MOHSE do not make information extraction. (iv) SWB's differ according to their deployment methods which are portlet, browser extraction and standalone application. (v) GATE, Piggy Bank and Magpie do not make dynamic linking. (vi) Only MOHSE is multilingual between all the SWBs.

Table 3.1: Comparison of Different Features of Semantic Web Browsers

SWB	Ontology Selection	Framework for SWB	Information Extraction	Deployment	Dynamic Linking	Multilingual
Sealife Browser	user selected	no	no	portlet	yes	no
KIM	generic	no	no	browser extension	yes	no
GATE	no	yes	yes	no	no	no
Piggy Bank	user selected	yes	yes	browser extension	no	no
Magpie	user selected	yes	yes	browser extension	no	no
Power Mappie	user selected	yes	yes	browser extension	yes	no
SemWeb	no	no	yes	browser extension	yes	no
VieW	user selected	no	yes	application	yes	no
PERSON	generic	no	yes	browser extension	yes	no
COHSE	user selected	no	no	portlet	yes	no
MOHSE	user selected	no	no	browser extension	yes	yes

The most important semantic browser for MOHSE is **SeaLife Browser** because it dynamically link Web resources to relevant services according to browser's background knowledge. This is accomplished using eScience's growing number of Web/Grid Services and its XML-based standards and ontologies. The browser identifies terms in the pages being browsed through the background knowledge held in ontologies. Semantic hyperlinks were identified by ontology term's servers and services [63]. MOHSE and SeaLife Browser have common features in using ontologies and linking resources dynamically, however Sealife Browser is not multilingual.

GATE [20] is a framework and graphical development environment which enables users to develop and deploy language engineering components. GATE provides tools such as tokenizers, part-of-speech taggers, gazetteer lookup components, pattern-matching grammars, coreference resolution tools and others that aid the construction of various NLP and especially IE applications. GATE is also a framework for content and annotation management. By using the GATE architecture, a number of successful applications are developed for various language processing task. The MUSE system is a multi-purpose Named Entity recognition system which is capable of processing texts from widely different domains and genres. The MUMIS (MUltiMedia Indexing and Searching environment) system uses Information Extraction components developed with GATE.

Knowledge and Information Management (KIM) is a platform for automatic semantic annotation, web page indexing and retrieval [59]. It uses named entities as a foundation for formulating semantic relationships in a document, and assigns ontological definitions to the entities in the text. The platform uses a massive populated ontology of common upper-level concepts (e.g. locations, organizations, dates or money) and their instances. KIM is a based on the GATE platform [20].

Like MOHSE, KIM is deployed as a browser plugin. Both KIM and MOHSE uses click&go hypermedia paradigm. However MOHSE differs from KIM in a number of ways. While KIM is coupled with a specific, large knowledge base, MOHSE is open with respect to multilingual ontologies, allowing user to select a particular semantic viewpoint, and use this to enrich the browsing experience by the help of MOHSE Firefox extension and MOHSE Knowledge Service.

Another Semantic Web Browser deployed as a browser extension is **Piggy Bank** [36]. Piggy Bank allows users to automatically produce Semantic Web information while they are consuming Web information. They can collect individual information items and store them locally in Semantic Web format. Also Piggy Bank uses screen scrapers to structure information located in Web pages if Web page does not include semantic data. Moreover, it offers a system called Semantic Bank for sharing items collected by users, between users. In contrast with MOHSE, Piggy Bank is not about dynamically linking documents. Instead, it allows users to create, share, import and reuse semantic information on Web Source.

Magpie [25] is another Semantic Web browser. Magpie provides a semantic layer to web pages based on user selected ontologies like MOHSE. Not only it offers a browser extension to users, but also gives developers a generic framework for creating Semantic Web applications. It uses named entities to reveal semantic relationships and associates entities with ontological definitions. It is stated that Magpie Semantic Web browser can increase user search performance [74]. However Magpie does not offer dynamic linking of web resources. Therefore **PowerMagpie** [31] has been developed to solve this problem. To perform this, it uses a Semantic Web gateway called Watson [21].

Another dynamic linking approach is **SemWeb** [61]. SemWeb is an extension of Mozilla Firefox Web browser. SemWeb adds a semantic layer to Web documents: it annotates Web pages using a Linked Data[4] domain and creates context-based hyperlinks on Web pages to guide users to find relevant pages. In addition, the information presented to the user is personalized based on a novel behavior. Although dynamic linking approach of SemWeb is same as the MOHSE, MOHSE is superior with se-

[4] Linked Data, http://linkeddata.org/

mantic annotation since it uses user selected ontologies.

VIeW [12] is a system that combines ontologies, web-based information extraction and automatic hyperlink to enrich Eeb documents with additional relevant background information. Semantic Browser supports user selected ontologies for semantic annotation like MOHSE. VIeW predefined ontologies are created by Protege Ontology Editor [29]. Web resources are founded by Google API for dynamic linking. MOHSE Resource Service gives different solution for crawling and indexing web resources by a search agent.

Another Semantic Web Browser deployed as a Firefox Browser Extension is **PERSON** [1]. The most important feature of PERSON is that PERSON not only offer a semantic layer with web pages, but also on RSS data of user choice. MOHSE differs from PERSON in several ways. Semantic annotation of PERSON works only generic driven ontology provided by PERSON. Knowledge base of PERSON created by Google Image API, Google Map API, Wikipedia API and RSS reader however MOHSE adds external web resources by crawling and indexing web documents.

In this section, we investigated several SWBs. Our survey shows that Sealife Browser, Piggy Bank, Magpie, VieW and COHSE allow user to select desired ontology like MOHSE. Most of the SWBs except Piggy Bank and Magpie can dynamically link web resources. MOHSE also deployed as a browser extension like other SWBs. However none of the SWBs considered are multilingual except from MOHSE.

3.3 Multilingual Text Retrieval

We covered some multilingual text retrieval techniques in the Section 2.3. In this section, we discuss some additional techniques.

A number of techniques used for text retrieval makes use of statistical information, called usage. Statistical information about term usage can be gleaned form parallel corpora. Most of the methods used multilingual text retrieval techniques are based on inverse document frequency and term frequency calculations. The calculation is given below:

$$idf_i = log_2(\frac{Number\ of\ documents}{Number\ of\ documents\ with\ term\ i})$$

$$tf_i = \frac{Number\ of\ term\ t\ occurs\ in\ document}{Total\ number\ of\ terms}$$

Several techniques can be used to compare *tfidf*, however the simplest one is vector space technique. In the vector space model text is represented by a vector of terms. Vectors of terms are normalized by inner product of two vectors and normalized inner product used to compute cosine similarity measure [55]. SMART is an example technique which uses this approach and is developed by Salton at Cornell University [62].

Probabilistic retrieval technique is another technique which are more complex than vector space technique. This technique seek to estimate that a given document is relevant based on *tfidf* evidence [55]. INQUERY, developed by Croft and others, is an example of a probabilistic text retrieval system. They described their experience with a range of projects involving text retrieval in Spanish, Japanese and Chinese [18].

Term vector translation and latent semantic indexing (LSI) are statistical techniques that have been applied to multilingual text retrieval [55]. In term vector translation technique, set of *tfidf* term weights are mapped from one language to another. Latent Semantic Index examines the similarity of the context in which word appear, and creates a reduced dimension feature space in which words that occur in similar contexts are near each other. Cross language Latent Semantic Indexing is a fully automatic method for cross language document retrieval in which no query translation is required. Cross language information retrieval is another method for machine translation used in many applications [44, 43].

Although probabilistic and statistical retrieval techniques were used for information retrieval in Turkish [13], there is no work done for multilingual retrieval in Turkish. However, Turkish English cross language information retrieval is studied [16].

Although it is beyond scope of this thesis, the methods discussed in this section for multilingual text retrieval can be adopted in the future work.

3.4 Automated Multilingual Vocabulary Mapping

Ontology mapping or ontology alignment is a field of research that looks at how to determine correspondences between concepts [71, 17, 40, 45]. There are many ontologies available and in typical complex applications, it is not enough to use only one ontology for processing. Therefore, typically more than one ontology is used and some kind of alignment or mapping is necessary between these ontologies. Mapping could provide a common layer from which several ontologies could be accessed and hence could exchange information in semantically sound manners. Developing such mappings has been the focus of ontology mapping research. In this field there has been some work on ontologies in different languages as well. Therefore, this could be somehow considered as automatically creating multilingual vocabularies. However, the work exist in this area is very limited as automatically mapping multilingual ontologies is very challenging and complex.

Ontology mapping in different languages can be useful for application that are open and distributed on the Web. For this purpose, different techniques are used, for mapping in terms of lexical, semantic, and structured ones, as terms may be mapped by a measure of lexical similarity, by semantic evaluation, usually on the basis of semantic linguistic resources, or by considering the term positions in the ontological hierarchy. All these solutions are supplementary with each other. Therefore, integrating the distinct ones will end up with a better solution. There are a number of work that particularly focuses on multilingual ontology mapping. Liang and Sini

(2006) [45] describes a concept based mapping procedures. AGROVOC and Chinese Agricultural Thesaures were mapped by representing vocabularies in OWL format. Francesconi, Faro and Marinai (2008) [28] described schema based mapping framework [60]. They implemented the framework for mapping thesauruses EUROVOC, ECLAS, GEMET, UNESCO Thesaurus and ETT. Lemon [51] is a RDF ontology-lexicon model represent translation on semantic web. Lemon architecture describes specific modules for different types of linguistic description. For translation, a new module which represents translation relation between lexicons in different natural languages associated to the same ontology, is added to architecture. This architecture is used in MONNET project [5]. MONNET is working on solutions aiming that easily access ontology-based information in multiple languages. Jung, Hakanson an Harturg (2009) [37] provide an indirect ontology mapping algorithm between Swedish and Korean ontologies, however more effort is needed to implement and evaluate their algorithm. Ontologies can be translated or enriched by using Linguistic Information Repository method [50, 27]. The main feature of LIR is that it provides a complete and complementary amount of linguistic data that allow localization of ontology elements to a specific linguistic and cultural universe. Moreover, the method provides a unified access to aggregated multilingual data. Trojan, Quaresma and Vieira (2008) [71] describes a framework for mapping of multilingual description logic ontologies. Firstly, source ontology is translated to the target ontology by using lexical database or dictionary. Then target and translated ontologies are used for mapping process. The framework includes agent based transformation. Each agent is responsible for lexical, semantic and structural relations. The framework is used for translating ontologies from Portuguese to English and provided neat results. In addition, the system is used in some question answering systems. In the future, it is planned to make the system more practical for developing further testing in other languages and evaluating for manual testing.

[5] MONNET project, http://www.monnet-project.eu/Monnet/Monnet/English

3.5 Automated Multilingual Translation

A number of studies exists in the literature on multilingual translation. For instance, there are frameworks that aim to design a controlled vocabulary for multilingual translation. Mitamura (1999) [49] presents a framework that uses a controlled vocabulary for natural language translation. The aim of the controlled vocabulary is to accomplish coherent authoring of source text, stimulate smoother, direct writing, and to bring out better quality output. Also it is indicated that, for a better machine translation, it is better to reduce the lexical ambiguity. In order to do that, they mainly follow three steps. In the first step, they aim to limit each term to exactly one meaning whenever possible. Therefore, it reduces ambiguity. In the second step, if a word has more than one meaning, they determine only one meaning for that word. Then, for the other meanings, they try to find another word which is also in domain. These terms are then marked in the lexicon, so that it can be known whether this term has any other meaning which is determined by a different term in the domain. Lastly, if a term has more than one meaning, then these meanings should be stored separately for the same term in the domain. If more than one entry is enabled for the same term, then lexical ambiguity will occur. Moreover, there are other lexical constraints. To illustrate, they suggest that more determinants and less pronouns and conjunctions should be used so that potential ambiguity can be prevented. Also, they give some restrictions about the use of -ing and -ed. For example, instead of saying ”When starting the engine”, the sentence should be ”When he starts the engine”. Furthermore, they limit the use of acronyms and abbreviations because they want to guarantee the designed acronyms and abbreviations not to cause any ambiguity. For example the acronym "OF" which stands for "Oil Field" can be mistaken with the preposition ”of”. Finally, they put emphasize on spelling, capitalization, hyphenation, and the use of slash character. In summary, the framework presented is interesting and could be useful for our work, however these limitations, which are mentioned above will make it difficult for us to use it further.

It is better to make some limitations on word for machine translation. KANT Controlled English for multilingual machine translation creates some limitations for English to increase success rate on machine translation [49]. It creates controlled vocabulary and controlled grammar, however, it is concluded that more limitations make it harder to use controlled grammar and vocabulary.

3.6 Thesaurus

WordNet [48] is an online English Thesauri. It groups nouns, verbs, adjectives and adverbs in sets of cognitive synonyms, each expressing different concepts. Words which indicate same concept are interchangeable in many context. WordNet can be represented as a format of RDF/OWL. WordNet can be used in applications in word sense disambiguation and information retrieval, question answering systems, information extraction, authoring and summarization, anaphora resolution, and event tracking. Although WordNet is an extremely important lexical database and used in many research in computational linguistic where lexical knowledge of English required, it is monolingual. Therefore, a similar project had been conducted to create multilingual wordnets. EuroWordNet [75] is a project that develops a multilingual database with wordnets in several European languages, structured along the same lines as the Princeton WordNet. EuroWordNet is an important development for multilingualism and can be used for multilingual text retrieval and lexical transfer in machine translation. The available languages in EuroWordNet are Dutch, Italian, Spanish, French, German, Czech and Estonian, however; unfortunately it does not support Turkish. Another WordNet project is BalkaNet [72] aiming to develop controlled vocabularies in the following Balkan languages: Bulgarian, Greek, Romanian, Serbian and Turkish. As a part of BalkaNet project, Turkish WordNet has been developed in Human Languages and Speech Technologies Laboratory at Sabanci University, Istanbul [9]. However it is not available in SKOS format that can be used for our thesis. Therefore, we decided not to further explore this.

We examined some of the automatically created multilingual thesaurus for further researches. BabelNet [54] is a project aiming to build a very large multilingual semantic network by mapping Princeton WordNet and Wikipedia with an aid of machine translation. Universal WordNet [23] is an automatically created large scale multilingual lexical database that organizes over 800,000 words from over 200 languages in a hierarchically structured semantic network providing over 1,5 million links from word to word meaning. Firstly, initial wordnet was created by using existing monolingual wordnets. Then it was extended by using translation dictionaries, thesauruses, and paralyzed corpora. Statistical methods are used to link terms in different languages. ConceptNet [32] is another machine interpreted semantic network which collects its data from Open Mind Common Sense project [64] which is a Web based collaboration over 15,000 authors enter sentences to contribute to project. Users answer question via Web interface and this fills the gaps in the project. ConceptNet is a semantic network which used by computers for understanding text written by humans. ConceptNet was extended for Turkish by using dictionary definitions of semantic word relations [2].

As mentioned in Section 2.4 in this study GEMET is employed because it includes Turkish terms and it can be converted to SKOS format.

3.7 Summary

This chapter covers related work section of the thesis. Firstly, we talk about multilingual dynamic linking. We discuss open and closed hypermedia systems, and internalization on the web. Secondly, we mention Semantic Web browsers and conclude that there is no SWB which is multilingual. Thirdly, we make further examination about multilingual text retrieval methods. Then, we discuss automated multilingual vocabulary mapping. This section includes ontology mapping and alignment researches. After that we cover automated multilingual translation. We decide that it is better to make some limitations on words for machine translation. Lastly, we give detailed information about thesaurus.

CHAPTER 4

ARCHITECTURE

In our work, we aim to create an efficient, robust, and scalable architecture for multilingual dynamic linking. MOHSE platform provides services and infrastructure for semantic annotation, indexing and retrieval. To do this, it performs information extraction based on an ontology and a knowledge base. Therefore; MOHSE architecture consists of four components. First one is Knowledge Source which is responsible for determining possible words which can be linked to other Web resources in the target language. Second part is Resource Manager that provides the target page according to the given term. Third part is a browser extension which sends HTML data to server and add links dynamically to the user interface (HTML Source). The last part is server which is responsible for integrating all the parts.

The remainder of the chapter is organized as follows: In Section 4.1 COHSE architecture is described. MOHSE design requirements is explained in Section 4.2. Section 4.3 is about MOHSE architecture. Section 4.4 Knowledge Service, 4.5 Resource Service, 4.6 Firefox Plugin, and 4.7 MOHSE Server present MOHSE components. In section 4.8, we summarize the chapter.

4.1 Background - COHSE Architecture

In this thesis, we have extended the existing COHSE architecture. This architecture was presented in several publications [30, 14, 3, 5, 77, 6, 78]. Basically COHSE is composed of three modules.

- Knowledge Service that maintains the ontology. It provides semantic and word knowledge using a structured vocabulary.

- Resource Service provides resource discovery and mapping words or concepts to web pages.

- COHSE DLS provides presentation and delivery of results. It controls the user's interaction with the Knowledge Service and Resource Service (see Figure 4.1).

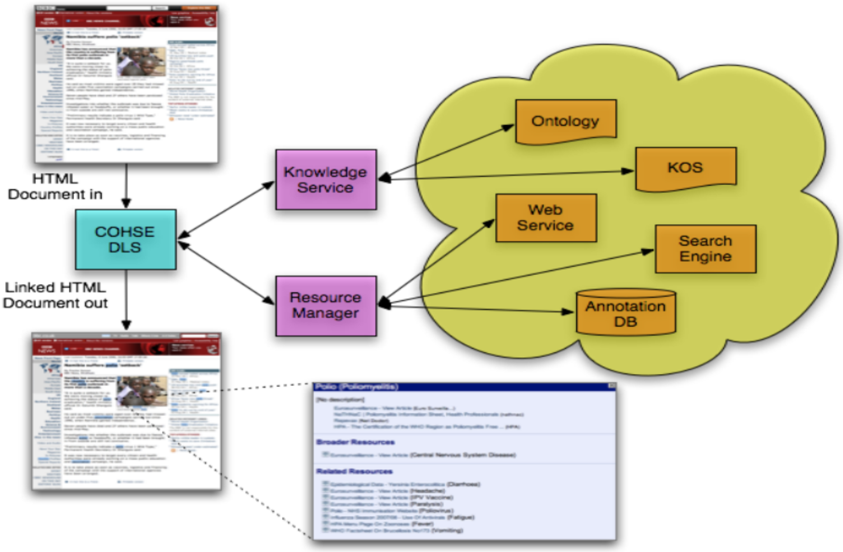


Figure 4.1: COHSE architecture: Architecture of the COHSE system showing how a plain web document is processed, the DLS uses the Knowledge Service and Resource Manager to add hyperlinks to documents and provide new link targets.

4.2 Design Requirements

MOHSE should support the interpretation of web pages. It should help users to make sense of the information presented in a web page. For example, if users are browsing web pages about a specific context like biology, MOHSE should enrich web pages according to this context. Moreover, web-based user interface should allow user to navigate the resources using both hypertext and semantic links.

Graphical user interface should enable user to switch to different ontology at any time. Because user may want to investigate the same document in a different context.

This is one of the important design features.

Web is a very dynamic structure and size of the web resources increases day by day. Moreover, some resources are deleted or deprecated in time. Therefore, sources dynamically added by MOHSE, should be updated regularly.

Enrichment process of MOHSE should not corrupt the layout of a page. Although web developers uses very different techniques for developing Web pages, MOHSE should work with the majority of Web resources. Also, process should be completed within acceptable time limitations. We will discuss process time in Section 5.3.

The key feature of MOHSE is that it should support multilingualism.

4.3 MOHSE Architecture

COHSE architecture meets all the requirements of MOHSE except multilingualism. COHSE interpret web pages and extend that according to given context. COHSE user interface supports switching between ontologies. Knowledge Service provides to load different RDF or OWL controlled vocabularies. Resource Service can be updated dynamically. Therefore; COHSE architecture is extended for providing multilingual constraints of MOHSE.

MOHSE is composed of four modules (see Figure 4.2) in order to meet the requirements mentioned in above section.

- MOHSE Knowledge Service extend COHSE Knowledge for supporting multilingualism.
- Resource Service is composed of search agent continuously crawl web resource and return them according to given query. The agent crawls Turkish web re-

sources.

- MOHSE was deployed as a browser plugin although COHSE was deployed as a portlet [78].

- MOHSE Server is the last module that control users' interaction with the Knowledge Service and Resource Service like COHSE DLS service [78].

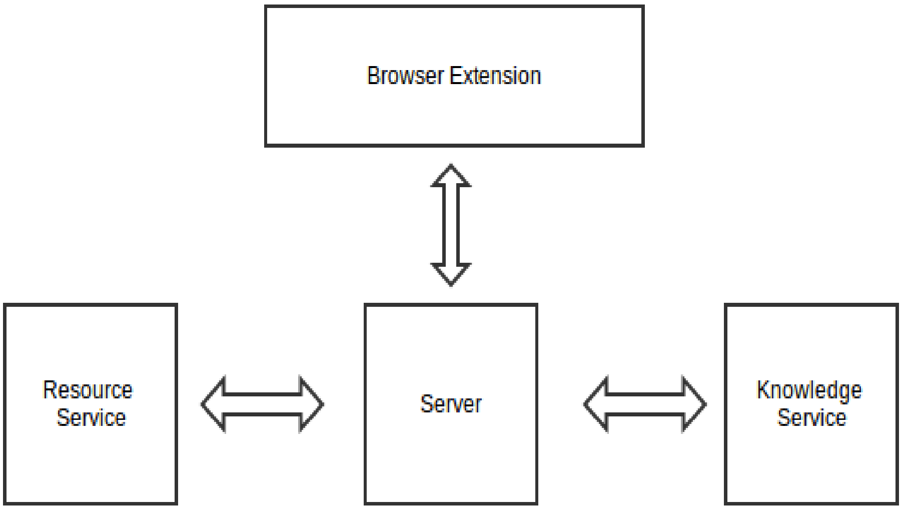


Figure 4.2: MOHSE modules

4.4 Knowledge Service

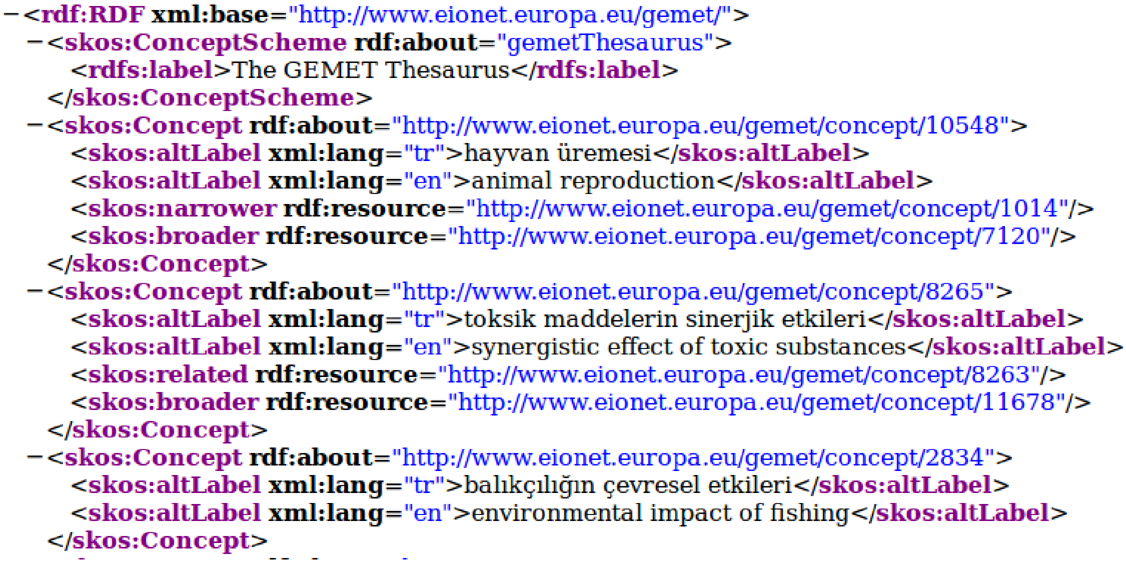
COHSE Knowledge Service [5, 30] supports interaction with ontologies by providing services such as mapping between concepts and lexical labels (synonyms), providing information about specialization (sub-classes) and generalization (super-classes) of concepts, description of concepts, etc. The service has a simple HTTP interface and can host third party ontologies represented by OWL and SKOS.

Even though COHSE Knowledge Service supports interaction between ontologies, there is no service that supports multilingualism. Therefore; we extended COHSE Knowledge Service to support multilingual operations. Extended Knowledge Service is capable of loading SKOS or OWL ontologies with language label.

Knowledge Service a standalone application deployed under Tomcat. MOHSE talks Knowledge Service by a rest API. Knowledge Service operations and example usage are given in Appendix G.

4.4.1 Creating Ontologies

Although GEMET is downloadable in SKOS format, this file does not include language label. Moreover, it is not separated according to context. Therefore, we decided to implement a simple tool which creates our SKOS ontologies according to selected context. It is a simple JAVA application which uses GEMET Web service API. The application takes GEMET context as an argument and creates a SKOS file including all the concepts in given context, Turkish and English label of concepts, and their broader, narrower and related relationship (see the Figure 4.3). Output of the tool can be loaded directly to MOHSE Knowledge Service.



```
- <rdf:RDF xml:base="http://www.eionet.europa.eu/gemet/">
  - <skos:ConceptScheme rdf:about="gemetThesaurus">
      <rdfs:label>The GEMET Thesaurus</rdfs:label>
    </skos:ConceptScheme>
  - <skos:Concept rdf:about="http://www.eionet.europa.eu/gemet/concept/10548">
      <skos:altLabel xml:lang="tr">hayvan üremesi</skos:altLabel>
      <skos:altLabel xml:lang="en">animal reproduction</skos:altLabel>
      <skos:narrower rdf:resource="http://www.eionet.europa.eu/gemet/concept/1014"/>
      <skos:broader rdf:resource="http://www.eionet.europa.eu/gemet/concept/7120"/>
    </skos:Concept>
  - <skos:Concept rdf:about="http://www.eionet.europa.eu/gemet/concept/8265">
      <skos:altLabel xml:lang="tr">toksik maddelerin sinerjik etkileri</skos:altLabel>
      <skos:altLabel xml:lang="en">synergistic effect of toxic substances</skos:altLabel>
      <skos:related rdf:resource="http://www.eionet.europa.eu/gemet/concept/8263"/>
      <skos:broader rdf:resource="http://www.eionet.europa.eu/gemet/concept/11678"/>
    </skos:Concept>
  - <skos:Concept rdf:about="http://www.eionet.europa.eu/gemet/concept/2834">
      <skos:altLabel xml:lang="tr">balıkçılığın çevresel etkileri</skos:altLabel>
      <skos:altLabel xml:lang="en">environmental impact of fishing</skos:altLabel>
    </skos:Concept>
```

Figure 4.3: GEMET SKOS file

4.5 Resource Service

The resource service is responsible for mapping concepts to resources. MOHSE resources are external links. Therefore, we need a web search agent for crawling and indexing of web resources and find a proper resource according to given query. Implementing a web search agent is not straightforward and out of the scope of this project, thus we decided to use Apache Solr[1] and Apache Nutch[2]. Apache Nutch is an effort to build an open source web search engine based on Java for the search and index component. It automatically crawls Web resources and store them. Apache Solr is an open source enterprise search platform from the Apache Lucene[3] project. Its major features include full-text search, highlighting, faceted search, dynamic clustering, database integration, and rich document (e.g., Word, PDF) handling. Providing distributed search and index replication, Solr is highly scalable and the most popular enterprise search engine.

Solr is written in Java and runs as a standalone full-text search server within a servlet container such as Apache Tomcat[4] or Jetty[5]. Apache Solr runs on port 8983. Web resources can be searched by HTML request. Crawling and indexing processes are described step by step in Appendix H.

4.6 MOHSE Firefox Extension

MOHSE Firefox Extension (see the Figure 4.4) controls the user's interaction with MOHSE. It is implemented by using JavaScript[6].

MOHSE button is added to Firefox toolbar. You can see the button in the Figure 4.5 in rectangle. When the user right click the MOHSE button, the contexts are listed. You can see the listed contexts in the Figure 4.6. When user select one of them, the URL of the page and selected context are sent to MOHSE server. MOHSE server enriches the HTML details are given in section 4.7. Then enriched HTML is sent to the extension. Enriched HTML page is directly replaced the original HTML. Enriched

[1] Apaphe Solr, http://lucene.apache.org/solr/

[2] Apache Nutch, http://nutch.apache.org/

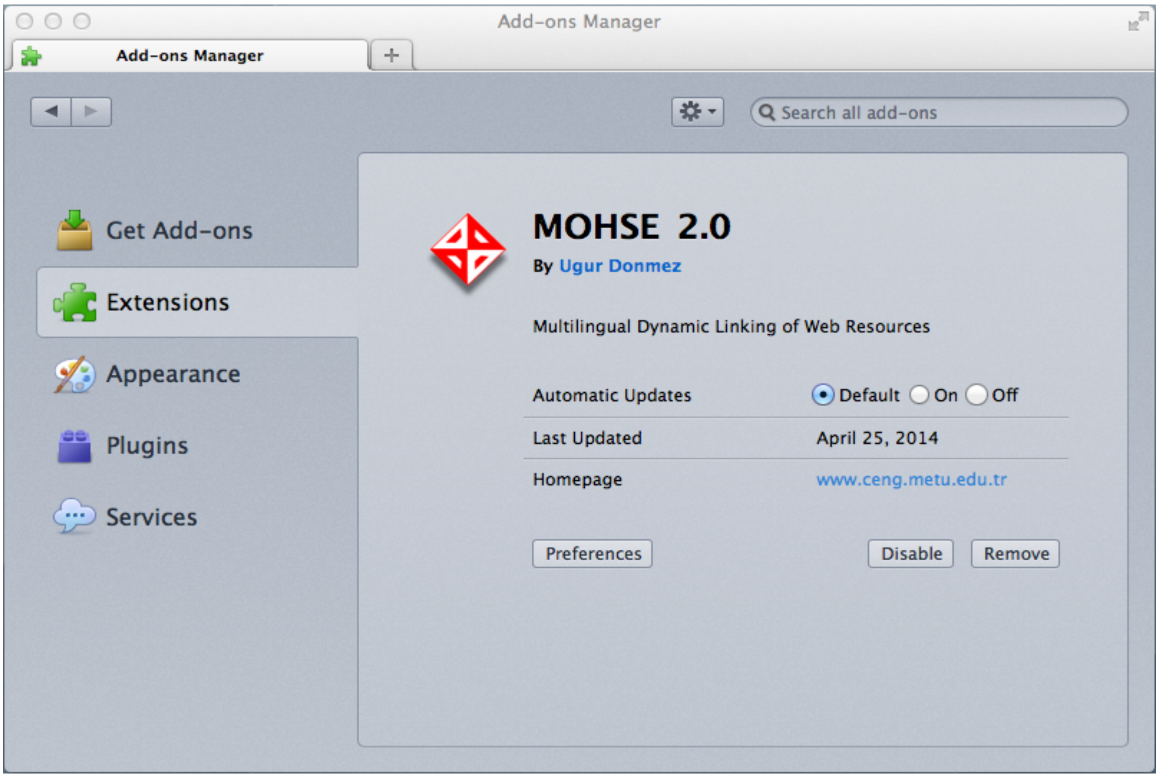
[3] Apache Lucene, http://lucene.apache.org/core/

[4] Apache Tomcat, http://tomcat.apache.org/

[5] Jetty, http://www.eclipse.org/jetty/

[6] JavaScript, https://developer.mozilla.org/en-US/docs/Web/JavaScript/Reference

HTML pages include MOHSE links. When user click one of them, a MOHSE link box will appear. The link box is composed of four components: (i) a list of exact resources links, (ii) a list of related resources links, (iii) a list of broader resources links, and (iv) a list of narrower resources links. Figure 4.7 shows the MOHSE link box.



MOHSE 2.0

By Ugur Donmez

Multilingual Dynamic Linking of Web Resources

Automatic Updates ● Default ○ On ○ Off

Last Updated April 25, 2014

Homepage www.ceng.metu.edu.tr

Figure 4.4: MOHSE Firefox Extension

4.7 MOHSE Server

Responsible for integrating all the modules. It is a Java application which controls all the data flow and process. Figure 4.8 and Figure 4.9 show flow chart and sequence diagram of MOHSE Server module.

Basically, it is composed of four modules.

- KSManager is responsible for creating HTTPrequest for MOHSE Knowledge

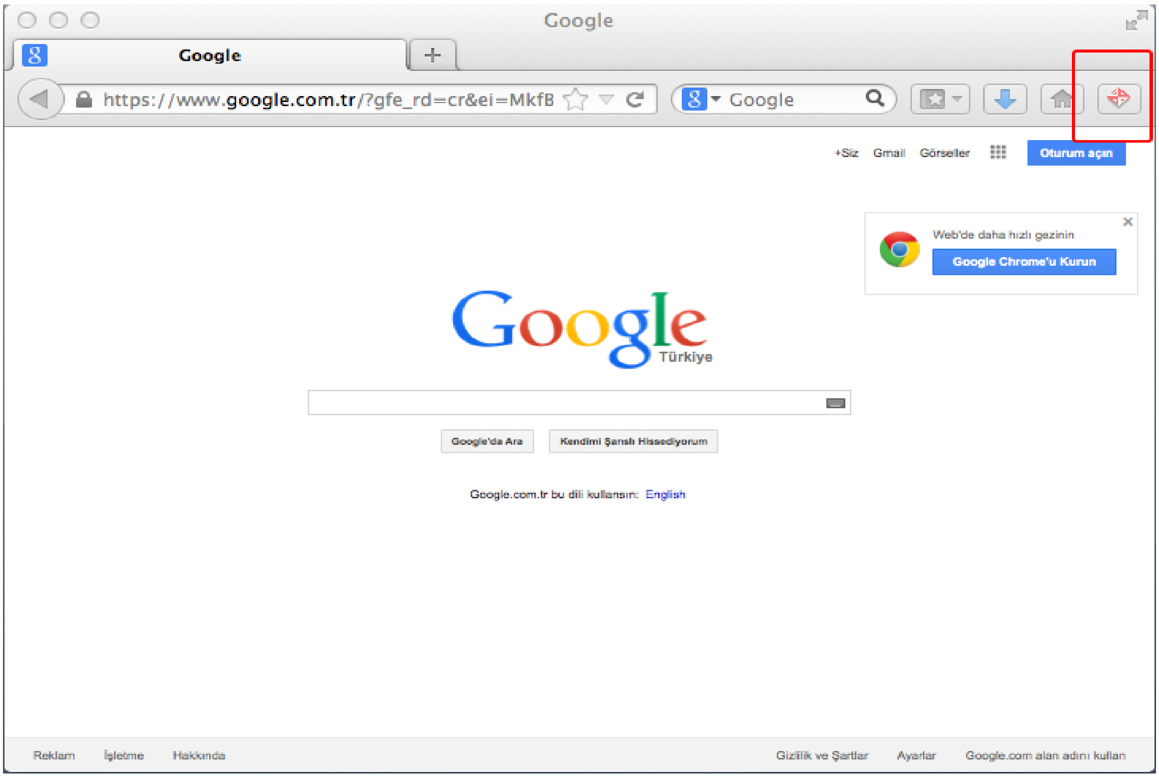

Google
Google
https://www.google.com.tr/?gfe_rd=cr&ei=MkfB
Google
+Siz
Gmail
Görseller
Oturum açın
Web'de daha hızlı gezinin
Google Chrome'u Kurun
Google
Türkiye
Google'da Ara
Kendimi Şanslı Hissediyorum
Google.com.tr bu dili kullanısın: English
Reklam
İşletme
Hakkında
Gizlilik ve Şartlar
Ayarlar
Google.com alan adını kullan

Figure 4.5: MOHSE Toolbar Button

Server and getting returned results.

- RSManager creates HTTPRequest for Solr server.
- Parsers parse the results from MOHSE Knowledge Server and MOHSE Resource Service and create objects from the results.
- Servlet interacts with MOHSE Firefox Extension via HTTPRequests.

KSManager and RSManager creates a HTTPRequest URL according to given parameters. Then they make a HTTPRequest to MOHSE Knowledge service or MOHSE Resource Service. The services returns the result in XML format. These results are parsed by parser module. Parser module uses SAX parser[7].

Servlet module provides all the logic. It takes the URL and context from MOHSE Browser Extension. By using JSoup [8], it converts the URL to DOM Objects [9]. It also

[7] http://www.saxproject.org/

[8] http://jsoup.org/

[9] http://www.w3.org/DOM/

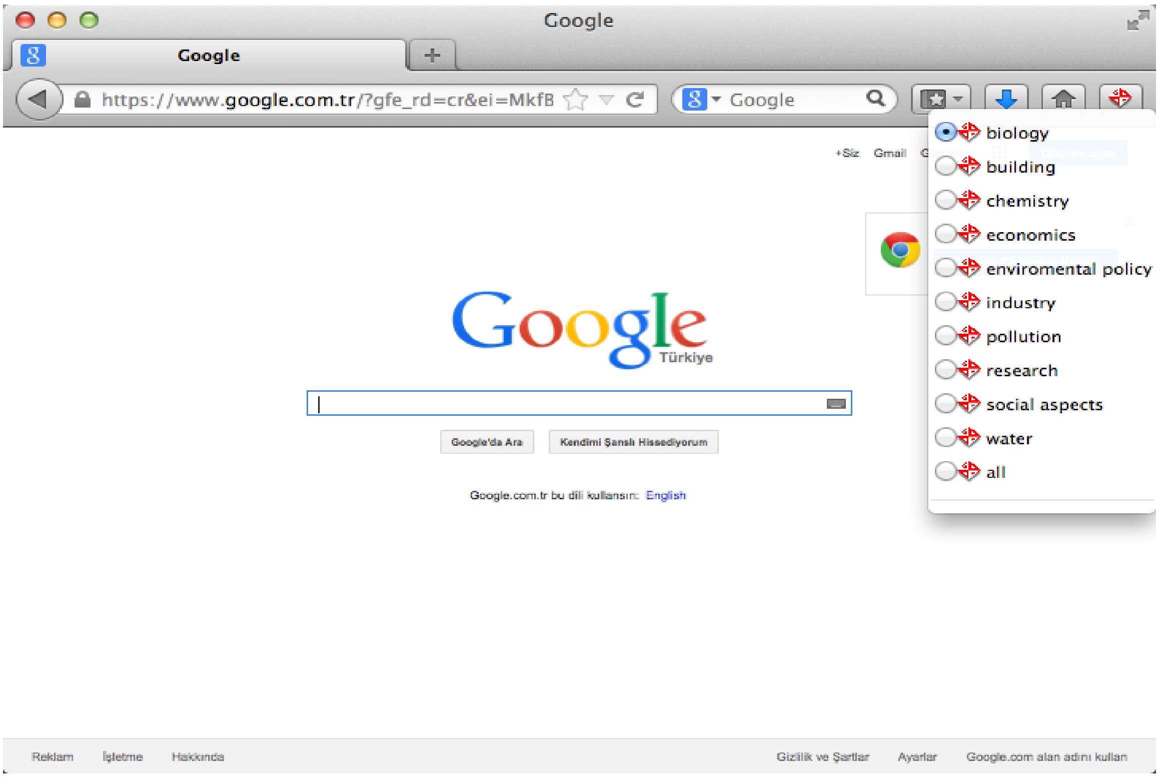
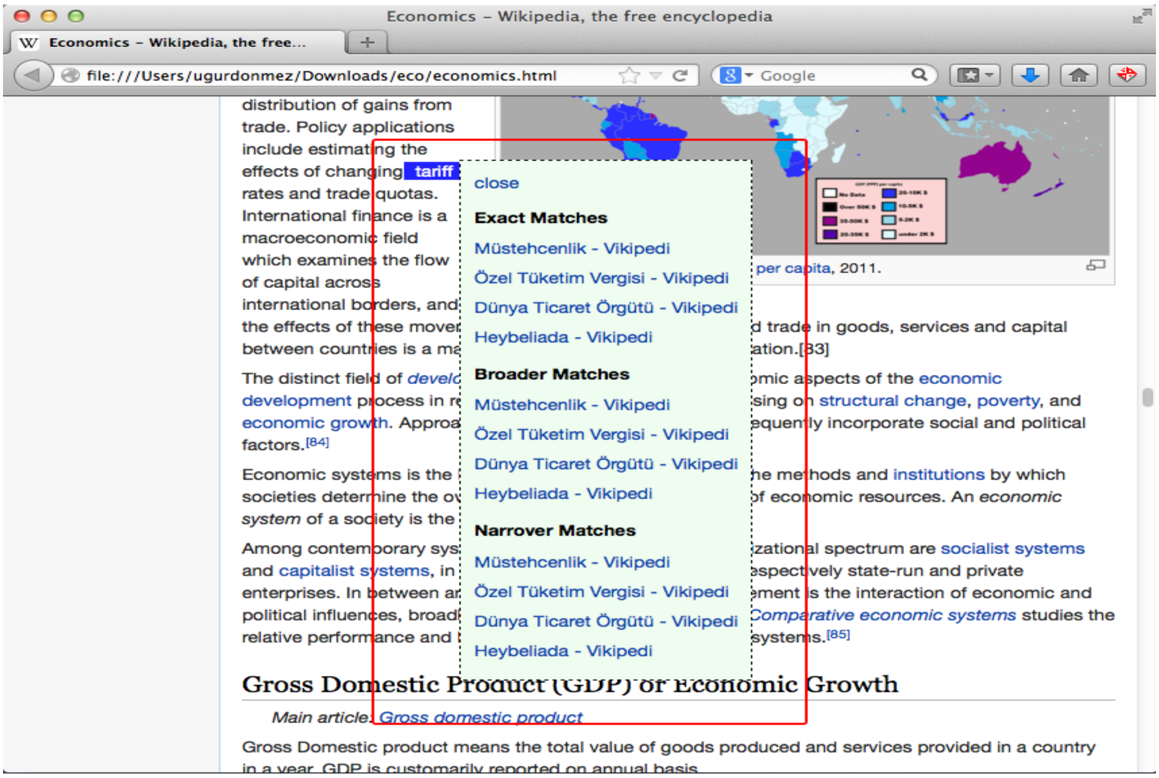


Figure 4.6: MOHSE Toolbar Button Ontology List

recognizes possible anchors from the returning results from KSManager according to a given context. If one of the anchor appear in DOM objects, sources for anchor taken from RSManager are added to DOM objects. After all the work is done, DOM Objects are sent via HTTPResponse to MOHSE Browser Extension in HTML format.


Economics – Wikipedia, the free encyclopedia
Economics – Wikipedia, the free…
file:///Users/ugurdonmez/Downloads/eco/economics.html
Google

distribution of gains from trade. Policy applications include estimating the effects of changing tariff rates and trade quotas. International finance is a macroeconomic field which examines the flow of capital across international borders, and the effects of these mover between countries is a ma

per capita, 2011.

d trade in goods, services and capital ation.[83]

The distinct field of develo development process in re economic growth. Approa factors.[84]

omic aspects of the economic sing on structural change, poverty, and equently incorporate social and political

Economic systems is the societies determine the ov system of a society is the

he methods and institutions by which of economic resources. An economic

Among contemporary sys and capitalist systems, in enterprises. In between ar political influences, broad relative performance and

zational spectrum are socialist systems espectively state-run and private ement is the interaction of economic and Comparative economic systems studies the systems.[85]

close

Exact Matches

Müstehcenlik - Vikipedi

Özel Tüketim Vergisi - Vikipedi

Dünya Ticaret Örgütü - Vikipedi

Heybeliada - Vikipedi

Broader Matches

Müstehcenlik - Vikipedi

Özel Tüketim Vergisi - Vikipedi

Dünya Ticaret Örgütü - Vikipedi

Heybeliada - Vikipedi

Narrover Matches

Müstehcenlik - Vikipedi

Özel Tüketim Vergisi - Vikipedi

Dünya Ticaret Örgütü - Vikipedi

Heybeliada - Vikipedi

Gross Domestic Product (GDP) or Economic Growth

Main article: Gross domestic product

Gross Domestic product means the total value of goods produced and services provided in a country in a year. GDP is customarily reported on annual basis

Figure 4.7: MOHSE Link Box

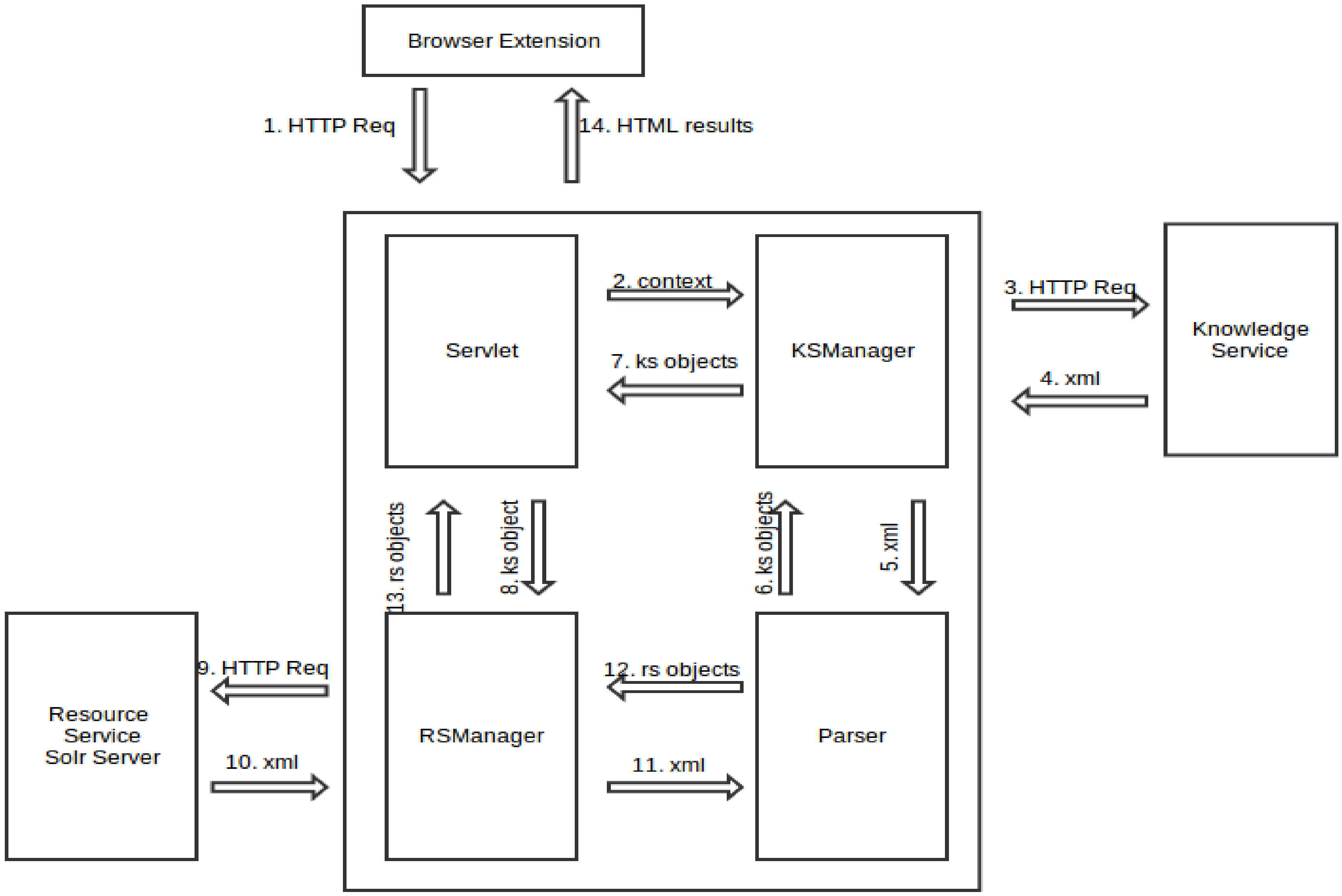


Figure 4.8: MOHSE Server Module Flow Chart

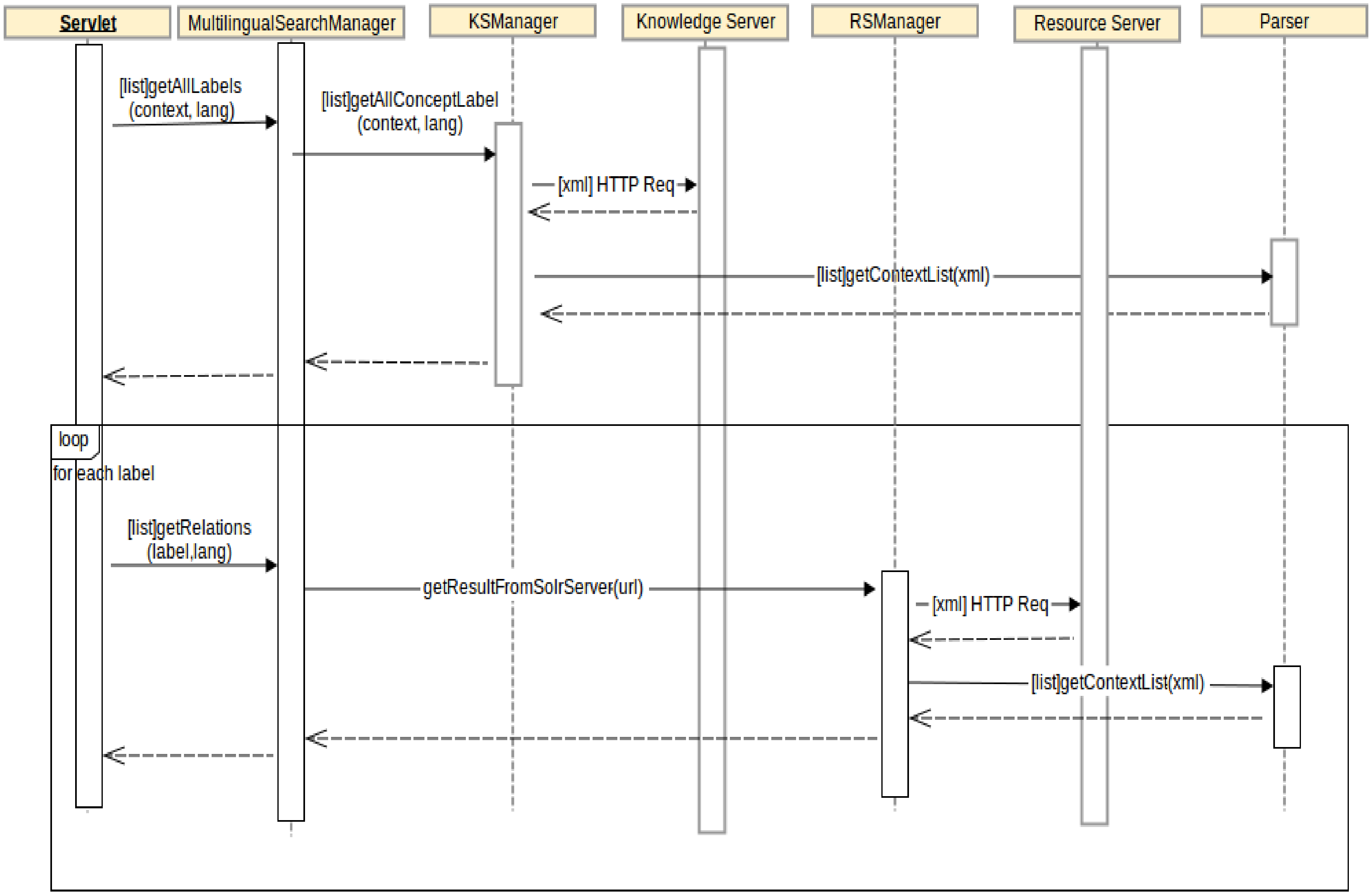


Figure 4.9: MOHSE Server Sequence Diagram

Tauscher and Greenberg (1997) [68] found that 58% of an individual's pages are revisits, and that users continually add new web pages into their repertoire of visited pages. People tend to revisit, access only a few pages frequently, browse in very small clusters of related pages and generate only short sequence of repeated URL paths. According to these statistics, we implemented a cache mechanism for MOHSE Server KSManager and RSManager. HTTPRequest URL and results are stored in a hashmap. Before sending request to MOHSE Knowledge Service or MOHSE Resource Manager, the hashmap is checked and if the result is found here, it is directly sent from the hashmap. Hashmap also store the last time that the object is used. The object is deleted which is not used longer. Therefore, it let the system work with recent data.

4.8 Summary

In this chapter we described detailed architecture of MOHSE. MOHSE's architecture is an impressed and extended version of the COHSE's architecture. Therefore, we summarized architecture of COHSE. Then we gave information about our design requirement. After that detailed architecture of MOHSE and each of the MOHSE's modules were defined. MOHSE has four components: (i) Knowledge Service is responsible for interaction with ontologies, (ii) Resource Service is an web search agent and crawler which crawls web resources continuously and maps terms to these resources, (iii) Firefox Extension controls the user's interaction with MOHSE, and (iv) MOHSE Server integrates all the components.

CHAPTER 5

EVALUATION

MOHSE evaluation should be conducted after studying usability in details. According to ISO standard, usability is divided into effectiveness, efficiency and satisfaction. *Effectiveness* is the ”accuracy and completeness with which users achieve specified goals”. In other words, a tool is effective if it helps users accomplish particular tasks. The most common way to measure effectiveness is using traditional performance measures, such as precision and recall tests. *Efficiency* is the ”resource expended in relation to the accuracy and completeness with which users achieve goals”. A tool is efficient if it helps users complete their tasks with minimum waste, expense or effort. The most common way to measure efficiency is to detect the time it takes a subject to complete a task. Ease of use and ease of learning are other indicators for measuring efficiency. If a tool is not easy to use or easy to learn, then it is inefficient for using. *Satisfaction* is the ”freedom from discomfort, and positive attitudes of the user to the product”. One of the most well-known instruments for measuring satisfaction is the Questionnaire for user interface satisfaction. To assess satisfaction a general question about satisfaction (e.g., how satisfied are you with your performance?) should be placed as a questionnaire item [41].

MOHSE evaluation process should cover these terms. Therefore, MOHSE should be tested for relevance of added links, user experience, and system performance to make sure that each modules of MOHSE works efficiently. Relevance of added links measures the performance of Knowledge Service and Resource Manager. User evaluation will test the MOHSE Firefox extension. System performance tests are necessary for testing integration of all modules.

The remainder of the chapter organized as follows: User evaluation and results are placed in Section 5.1. Section 5.2 covers link evaluation process and results. Section 5.3 covers system performance evaluation.

5.1 User Evaluation

In order to evaluate our work on multilingual linking of Web resources, we have conducted a number of tests for user evaluation. Our aim is to assess MOHSE for effectiveness, efficiency and satisfaction.

5.1.1 Procedure

We have designed and conducted the user evaluation which includes the following three main parts:

Introduction: The participants read the information sheet and signed the consent form (See Appendix A - B). Next, participants were given a short questionnaire to get contextual measures [41]. The contextual measures include those which can be used to characterize the subject: such as age, sex, search experience, personality-type, and those which can be used to characterize information-seeking situation: such as task-type and subjects familiarity with topic (See Appendix **??**).

Lastly, participants were informed about MOHSE and how to use it. MOHSE toolbar was shown. In order to teach participant for usage of MOHSE, an example Web page was enriched with MOHSE. The example Web page would not be used in further tasks.

Main Part: The participants were asked to answer 12 questions about economy and biology with using MOHSE (See Appendix D) to assess effectiveness and efficiency.

Conclusion: At the end, the participants were given a short questionnaire to evaluate user satisfaction (See Appendix E).

5.1.2 Materials

For conducting the user survey, two contexts of MOHSE were chosen out of 10. These contexts are ”economy” and ”biology”. We chose these ones because these include more terms than others. After selecting them, we selected two Wikipedia pages for testing these contexts. These pages include more anchor than other pages. The pages are:
http://en.wikipedia.org/wiki/Economics
http://en.wikipedia.org/wiki/Biology

In order to prepare questions, those pages were given to teaching assistants from the department of economy and biology in METU NCC. In order to demonstrate an objective approach, the questions were prepared by them. You can see the questions in Appendix D.

Each question were answered in four ways. In the first way, the questions were answered in English with the help of MOHSE. In second way, they were answered in English only. In the third way, they were answered in Turkish with the help of MOHSE. In last way, they were answered in Turkish only. Therefore, four task types were prepared according to these ways. You can see the distribution of the questions in Table 5.1. Moreover, Appendix D includes these tasks.

Table 5.1: Question Distribution for Each Task

Question	Task A	Task B	Task C	Task D
Economy 1	Eng-Mohse	Eng	Tr-Mohse	Tr
Economy 2	Eng-Mohse	Eng	Tr-Mohse	Tr
Economy 3	Eng-Mohse	Eng	Tr-Mohse	Tr
Economy 4	Tr-Mohse	Tr	Eng-Mohse	Eng
Economy 5	Tr-Mohse	Tr	Eng-Mohse	Eng
Economy 6	Tr-Mohse	Tr	Eng-Mohse	Eng
Biology 1	Eng	Eng-Mohse	Tr	Tr-Mohse
Biology 2	Eng	Eng-Mohse	Tr	Tr-Mohse
Biology 3	Eng	Eng-Mohse	Tr	Tr-Mohse
Biology 4	Tr	Tr-Mohse	Eng	Eng-Mohse
Biology 5	Tr	Tr-Mohse	Eng	Eng-Mohse
Biology 6	Tr	Tr-Mohse	Eng	Eng-Mohse

5.1.3 The Scientific Questions

Semantic Web Browsers are shared by users however, little attention has been given for evaluation with real users to figure out the enhancements and obtain valuable feedback [57]. Nevertheless, for testing Sealife SWBs a user-centered evaluation framework is developed which use three source of data: (i) web server logs; (ii) user questionnaires; and (iii) semi-structured interviews. These sources are analyzed and comparisons are performed between each browser and a control system.

According to this framework we developed following hypothesis to evaluate MOHSE:

H1: MOHSE help users to find information or complete task.
H2: Users understand how to use MOHSE to find such information or complete task.
H3: MOHSE help users to understand English Web sites.
H4: Users find MOHSE easy to use.
H5: Where semantic links are available, users will follow them instead of nonsemantic links.

To prove or disprove the hypotheses, the following questions were considered to test

MOHSE:
Q1: Does having question in English or Turkish affect the correctness of the results?
Q2: Does using MOHSE affect the correctness of the results?
Q3: Do participants find MOHSE easy to use?
Q4: Does MOHSE help participants to complete their tasks?
Q5: Is information found by MOHSE relevant?

For each hypotheses we included the following details:

- What do we try to find with this scientific question?
- What is the context of the data which we have related with this scientific question?
- Which techniques are used to evaluate the result of this scientific question?
- What are the results and how do we interpret them?

5.1.4 Participants

For manual assessment volunteers were found in graduate and undergraduate students in METU and METU NCC which are comfortable in both Turkish and English. These students make good assessors for two reason: (i) they are well educated, and can understand Wikipedia articles; and (ii) they are comfortable in both language because Turkish is their native language and English is medium of instruction in METU.

The survey results includes many participants with different profiles. Of our 40 participants, 7 were female and 33 were male. Age average of participants was 22.25 with 2.97 standard deviation. Thirty participants completed high/secondary school, seven completed bachelors degree, and three completed master degree. All the participants use Internet daily. Seven participants use Wikipedia daily, 22 of them use weekly, 6 of them use monthly, and 5 of them access to the Wikipedia less than once every month. Seven participants prefer Turkish while searching for something on the Web, 8 prefer English, and 25 prefer both. Demographic summary of the participants is given in the Table 5.2.

Table 5.2: Demographic summary of participants

Criteria	Value	Count	Percent (%)
Gender	Female	7	17.50
	Male	33	82.50
Age	Under 18	0	0.00
	18-24	33	82.50
	25-34	7	17.50
	35-54	0	0.00
	55+	0	0.00
Web Usage	Daily	40	100.00
	Weekly	0	0.00
	Montly	0	0.00
	Less than once a month	0	0.00
	Never	0	0.00
Education	Grade/Primary School	0	0.00
	High/Secondary School	30	75.00
	Associates Degree	0	0.00
	Bachelors Degree	7	17.50
	Masters Degree	3	7.50
	Doctorate	0	0.00
	Other	0	0.00
Wikipedia Usage	Daily	7	17.50
	Weekly	22	55.00
	Montly	6	15.00
	Less than a month	5	12.50
	Never	0	0.00
Preferred Web Language	English	8	20.00
	Turkish	7	17.50
	Both	25	62.50

5.1.5 Results

Table 5.3 shows the overall results of each task according to participants answers.

Table 5.3: Result of each task

Questions		English				Turkish			
		Mohse		Without MOHSE		MOHSE		Without MOHSE	
		Correct	Wrong	Correct	Wrong	Correct	Wrong	Correct	Wrong
Economy	Question 1	2	8	2	8	5	5	4	6
	Question 2	6	4	4	6	8	2	8	2
	Question 3	8	2	5	5	9	1	10	0
	Question 4	9	1	5	5	8	2	7	3
	Question 5	5	5	8	2	10	0	8	2
	Question 6	5	5	6	4	6	4	4	6
Biology	Question 1	8	2	9	1	10	0	10	0
	Question 2	9	1	9	1	8	2	9	1
	Question 3	8	2	7	3	10	0	10	0
	Question 4	9	1	10	0	9	1	6	4
	Question 5	5	5	7	3	5	5	4	6
	Question 6	8	2	3	7	6	4	0	10

The Scientific Question I: Does having question in English or Turkish affect the correctness of the results?

What we try to find: In this question, we try to find effects of the question's language on the answers.

Context of the data: Our data are participant's answers of the questions.

Technique of evaluation: We have applied Pearson's Chi Square Test to find effects of the language. To get effects of language only, we applied test the questions answered with using MOHSE or without MOHSE, separately. Our significance level is 0.05. Our null hypothesis is that language does not affect the answers. Our alternative hypothesis is that language affects participants answers.

Results: Result of the participants answers according to language is shown in the Table 5.4.

Discussion: When confidence level is taken as 0.05, language affect only 3 questions answers. On the other hand, when confidence is level 0.10, language affect 6 of the answers out of 20. Therefore, we can assume that our null hypothesis is true except from 3 questions with our significance level 0.05. As stated in Section 5.1.4, participants are comfortable in both Turkish and English. Results do not conflict with this.

Table 5.4: Pearson’s Chi Square Test results for participant answers according to language of the questions

Questions	MOHSE					Without Mohse				
	English		Turkish		Asymp. Sig.	English		Turkish		Asymp. Sig.
	Correct	Wrong	Correct	Wrong	2-sided	Correct	Wrong	Correct	Wrong	2-sided
Economy 1	2	8	5	5	0.160	2	8	4	6	0.329
Economy 2	6	4	8	2	0.329	4	6	8	2	0.068
Economy 3	8	2	9	1	0.531	5	5	10	0	**0.010**
Economy 4	9	1	8	2	0.531	5	5	7	3	0.361
Economy 5	5	5	10	0	**0.010**	8	2	8	2	1.000
Economy 6	5	5	6	4	0.633	6	4	4	6	0.371
Biology 1	8	2	10	0	0.136	9	1	10	0	0.305
Biology 2	9	1	8	2	0.531	9	1	9	1	1.000
Biology 3	8	2	10	0	0.136	7	3	10	0	0.060
Biology 4	9	1	9	1	1.000	10	0	6	4	**0.025**
Biology 5	5	5	5	5	1.000	7	3	4	6	0.178
Biology 6	8	2	6	4	0.329	3	7	10	0	0.060

The Scientific Question II: Does using MOHSE increase the number of correct answers?

What we try to find: In this question, we try to find effects of the MOHSE on the answers.

Context of the data: Our data are participant's answers of the questions.

Technique of evaluation: We have applied Pearson's Chi Square Test to find effects of the language. Our significance level is 0.05. Our null hypothesis is that MOHSE does not affect the answers. Our alternative hypothesis is that MOHSE affects participants answers.

Results: Result of the participants answers according to language is shown in the Table 5.5.

Discussion: According to confidence level 0.05, MOHSE affects only 2 questions' answers. Moreover, according to confidence level 0.10, MOHSE affects 3 questions. Therefore, our null hypotheses is true for 18 questions with our significance level 0.05. However, it is clear that participants who used MOHSE have better result than others because their number of correct answers is better than others'. It shows that MOHSE has some effect on the answers.

Table 5.5: Pearson's Chi Square Test results for participant answers according to usage of MOHSE

Questions	English					Turkish				
	MOHSE		W. MOHSE		Asymp. Sig.	MOHSE		W. MOHSE		Asymp. Sig.
	Correct	Wrong	Correct	Wrong	2-sided	Correct	Wrong	Correct	Wrong	2-sided
Economy 1	2	8	2	8	1.000	5	5	4	6	0.653
Economy 2	6	4	4	6	0.371	8	2	8	2	1.000
Economy 3	8	2	5	5	0.160	9	1	10	0	0.305
Economy 4	9	1	5	5	0.051	8	2	7	3	0.606
Economy 5	5	5	8	2	0.160	10	0	8	2	0.136
Economy 6	5	5	6	4	0.633	6	4	4	6	0.371
Biology 1	8	2	9	1	0.531	10	0	10	0	1.000
Biology 2	9	1	9	1	1.000	8	2	9	1	0.531
Biology 3	8	2	7	3	0.606	10	0	10	0	1.000
Biology 4	9	1	10	0	0.305	9	1	6	4	0.121
Biology 5	5	5	7	3	0.361	5	5	4	6	0.653
Biology 6	8	2	3	7	**0.025**	6	4	0	10	**0.003**

The Scientific Question III: Is there a correlation between participants' profiles and participants' answers to questions.

What we try to find: At the beginning of the survey, we collected some demographic information of the participants.This information includes gender, age, education, Wikipedia usage, and preferred search language on the web. Using these demographic data, we can group the participants in different level. In this question, we try to find whether some trends exists for some participant groups on user answers.

Context of the data: Data consists of the participant profile for all criteria and their answers of the questions.

Technique of evaluation: For each criteria, we have applied Pearson's Chi Square Test to find whether a correlation exists between participant groups and their responses. Our significance level is 0.05. Our null hypothesis is that there is no relationship between participant profiles and their responses. Our alternative hypothesis is that there is a relationship between participant profiles and their responses.

Results: Result of the participants answers according to language is shown in the Table 5.6 and Table 5.7. The result which satisfy p-value <= 0.05 indicate that , null hypothesis can be rejected in their cases.

Discussion: According to confidence level 0.05 participants' groups do not affect on participants' answers. Therefore, our null hypotheses is true.

Table 5.6: Pearson’s Chi Square Test results for participant groups to economy questions

Participant Profile		N	Economy 1			Economy 2			Economy 3			Economy 4			Economy 5			Economy 6		
			C	W	p-value	C	W	p-value	C	W	p-value	C	W	p-value	C	W	p-value	C	W	p-value
Gender	Female	7	3	4	0.519	5	2	0.695	5	2	0.533	6	1	0.389	4	3	0.156	3	4	0.574
	Male	33	10	23		21	12		27	6		23	10		27	6		18	15	
Age	18-24	31	10	21	0.952	20	11	0.905	25	6	0.850	21	10	0.211	24	7	0.982	17	14	0.583
	25-34	9	3	6		6	3		7	2		8	1		7	2		4	5	
Education	High/ Secondary School	30	8	22	0.301	19	11	0.920	24	6	0.788	20	10	0.200	23	7	0.597	17	13	0.634
	Bachelors Degree	7	4	3		5	2		6	1		7	0		5	2		3	4	
	Masters Degree	3	1	2		2	1		2	1		2	1		3	0		1	2	
Wikipedia Usage	Daily	7	2	5	0.798	5	2	0.745	6	1	0.836	5	2	0.861	6	1	0.875	5	2	0.365
	Weekly	22	8	14		14	8		18	4		17	5		16	6		12	10	
	Monthly	6	1	5		3	3		4	2		4	2		5	1		2	3	
	Less than once a month	5	2	3		4	1		4	1		3	1		4	1		1	4	
Preferred Search Language	Turkish	7	5	2	0.530	3	4	0.375	6	1	0.875	6	1	0.601	5	2	0.875	4	3	0.218
	English	8	2	6		6	2		6	2		5	3		5	3		2	6	
	Both	25	6	19		17	8		20	5		18	7		20	5		15	10	

Table 5.7: Pearson’s Chi Square Test results for participant groups to biology questions

Participant Profile		N	Biology 1			Biology 2			Biology 3			Biology 4			Biology 5			Biology 6		
			C	W	p-value	C	W	p-value	C	W	p-value	C	W	p-value	C	W	p-value	C	W	p-value
Gender	Female	7	6	1	0.453	7	0	0.271	7	0	0.271	7	0	0.221	4	3	0.787	1	6	0.096
	Male	33	31	2		28	5		28	5		27	6		17	16		16	17	
Age	18-24	31	30	1	0.057	26	5	0.198	28	3	0.316	27	4	0.491	18	13	0.191	14	17	0.527
	25-34	9	7	2		9	0		7	2		7	2		3	6		3	6	
Education	High/ Secondary School	30	29	1	0.129	25	5	0.386	27	3	0.10	26	4	0.651	16	14	0.775	15	15	0.177
	Bachelors Degree	7	6	1		7	0		7	0		6	1		3	4		2	5	
	Masters Degree	3	2	1		3	0		1	2		2	1		2	1		0	3	
Wikipedia Usage	Daily	7	7	0	0.188	7	0	0.188	6	1	0.895	6	1	0.784	6	1	0.105	4	3	0.613
	Weekly	22	20	2		20	2		20	2		18	4		12	10		9	13	
	Monthly	6	5	1		5	1		5	1		5	1		2	4		3	3	
	Less than once a month	5	3	2		3	2		4	1		5	0		1	4		1	4	
Preferred Search Language	Turkish	7	6	1	0.486	6	1	0.486	7	0	0.342	6	1	0.970	5	2	0.359	2	5	0.289
	English	8	7	1		8	0		6	2		7	1		5	3		2	6	
	Both	25	24	1		21	4		22	3		21	4		11	14		13	12	

The Scientific Question IV: Is there a correlation between participants' groups and participants' rating on post questionnaire.

What we try to find: In this question, we try to find whether some trends exists for some participants' groups on participants' ratings.

Context of the data: Data consists of the participants' profile for all criteria and their answers of the questions in post questionnaire.

Technique of evaluation: We simply calculate the mean and standard deviation values for participants' ratings.

Results: Result of the participants answers according to language is shown in the Table 5.8 and Table 5.9.

Discussion: Participants' gender has no effect on rating of post questionnaire as mean and standard deviation of ratings are very close to each other. When the level of education is examined, the higher the participants' education level, the lower his/her ratings are. Therefore, it can be assumed that education has some effects on participant rating on post questionnaire. Wikipedia usage has similar effects to those of education. If usage frequency of Wikipedia increases, the ratings decrease slightly. However, it is clear that preferred search language does not affect participants ratings. It verifies our assumption on Section 5.1.4 and Scientific Question I.

Table 5.8: Rating results for post questionnaire’s first four question with respect to participants’ profiles

Participant Profile		N	Question 1		Question 2		Question 3		Question 4	
			Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Gender	Female	7	4.14	0.900	1.71	0.951	4.43	1.134	3.71	1.113
	Male	33	4.06	1.171	1.45	0.711	4.42	1.001	3.88	1.219
Education	High/ Secondary School	30	4.13	1.074	1.40	0.724	4.47	1.008	4.03	1.098
	Bachelors Degree	7	4.43	0.787	1.86	0.900	4.43	1.134	3.71	1.113
	Masters Degree	3	2.67	1.528	1.67	0.577	4.00	1.000	2.33	1.528
Wikipedia Usage	Daily	7	3.71	1.890	1.29	0.488	4.00	1.155	3.86	1.574
	Weekly	22	4.32	0.839	1.41	0.590	4.77	0.612	3.95	0.999
	Monthly	6	4.00	1.095	1.67	0.816	4.17	1.169	3.67	1.506
	Less than once a month	5	3.60	0.894	2.00	1.414	3.80	1.643	3.60	1.342
Preferred Search Language	Turkish	7	4.29	0.756	2.29	0.756	4.29	1.113	3.57	1.134
	English	8	4.13	1.458	1.25	0.463	4.63	0.744	3.88	1.553
	Both	25	4.00	1.118	1.36	0.700	4.40	1.080	3.92	1.115

Table 5.9: Rating results for post questionnaire's last four question with respect to participants' profiles

Participant Profile		N	Question 5		Question 6		Question 7		Question 8	
			Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Gender	Female	7	3.86	1.069	4.29	1.254	4.29	0.951	4.29	0.756
	Male	33	4.03	1.334	4.55	0.833	4.39	0.899	4.27	0.977
Education	High/ Secondary School	30	4.13	1.279	4.60	0.724	4.50	0.682	4.37	0.809
	Bachelors Degree	7	3.86	1.069	4.57	0.787	4.29	0.951	4.29	0.756
	Masters Degree	3	3.00	1.732	3.33	2.082	3.33	2.082	3.33	2.082
Wikipedia Usage	Daily	7	3.43	1.718	4.00	1.414	3.86	1.464	3.29	1.496
	Weekly	22	4.09	1.109	4.68	0.716	4.55	0.739	4.55	0.510
	Monthly	6	4.50	0.837	4.33	1.033	4.33	0.816	4.33	0.816
	Less than once a month	5	3.80	1.789	4.60	0.548	4.40	0.548	4.40	0.894
Preferred Search Language	Turkish	7	3.86	0.900	4.29	0.756	3.86	0.900	4.00	0.816
	English	8	4.00	1.414	4.38	1.408	4.38	1.408	4.25	1.389
	Both	25	4.04	1.369	4.60	0.764	4.52	0.653	4.36	0.810

The Scientific Question V: Do participants find the system easy to use?

What we try to find: In this question, we try to check whether MOHSE is easy to use or not.

Context of the data: Data consists of the participant 5 scale rating results to the first three questions of the post questionnaire.

Technique of evaluation: We simply calculate the mean and standard deviation values for participants' ratings.

Results: The results are given in Table 5.10.

Discussion: For the first question, significant of the participants (%50) state that they think they would like to use the system frequently, whereas only %5 of the participants do not share the same view as they stated above. Moreover, mean of the participants' rating is 4.075 out of 5. It seems very satisfying. In case of second question, an excessive part of the participants (%62.5) found the system uncomplicated. Moreover, none of the participants found MOHSE hard to use as nobody chose totally agree option. Also, mean of the ratings is too low (1.500) with a few standard deviation. The third question asks to participants whether the system is easy to use or not. It is nearly opposite of the second question. According to results of the third question, there is no contradiction in ratings of second and third question. %67.5 of the participants found system easy to use and nobody indicated that it was hard. To sum up, participants found the system easy to use with respect to information indicated above.

Table 5.10: Participants' rating results for system usage

Question	N	Mean	Std. Dev.
I think that I would like to use this system frequently	40	4.075	1.10425
I found the system unnecessarily complex	40	1.500	0.74161
I thought the system easy to use	40	4.425	0.99718

The Scientific Question VI: Does MOHSE help participants for completing task?

What we try to find: In this question, we try to find that use of MOHSE help participants to complete task.

Context of the data: Data consists of the participant 5 scale rating results to the forth, seventh and eighth question of the post questionnaire.

Technique of evaluation: We simply calculate the mean and standard deviation values for participants' ratings.

Results: The results are given in Table 5.11.

Discussion: Most of the participants (%42.5) were able to find the answers to the task in the information found by MOHSE however, the mean of the question is 3.85 and it is the lowest rating between post questionnaire questions. Nevertheless, the ratings of other questions are satisfactory. %67.5 of the participants found anchors helpful and %50 of the participants stated that semantic links help them to complete task. Only one participant (%2.5) did not agree with others' decisions.

Table 5.11: Participants' rating results for helps of MOHSE to complete tasks

Question	N	Mean	Std. Dev.
I was able to find the answers to the task in the information found by the system(anchors and the links added to these anchors)	40	3.850	1.17366
Did you find the highlighting the ontology terms helpful?	40	4.375	0.88564
Did you find the semantic links helpful?	40	4.275	0.92161

The Scientific Question VII: Is information found by MOHSE relevant?

What we try to find: In this question, we try to find that most of the information found by MOHSE is relevant.

Context of the data: Data consists of the participants 5 scale rating results to the fifth question of the post questionnaire.

Technique of evaluation: We simply calculate the mean and standard deviation values for participants' ratings.

Results: The results are given in Table 5.12.

Discussion: The mean of the participants' ratings for the question is 4. It shows that participants think that information found by MOHSE is relevant. Moreover, a half of the participants chose the highest rating (5, totally agree) for the question.

Table 5.12: Participants' rating results for information found by MOHSE

Question	N	Mean	Std. Dev.
The most information I found was relevant.	40	4.000	1.17366

5.1.6 Summary

User evaluation of MOHSE was conducted to investigate the usability of MOHSE. Participants were given a short initial questionnaire to get contextual measures. Then, participants were asked to answer 12 questions with the help of MOHSE to measure effectiveness and efficiency. At the end of evaluation, participants were answered post questionnaire which evaluate satisfaction of participants. After the evaluation, we found following results:

- Having questions in English or Turkish does not effect the correctness of the results.

- There are no correlations between participants' profiles and participants' answers to the questions.

- There are no correlations between participants' profiles and participants' rating on the post questionnaire.

- MOHSE help participants to complete their tasks. Results show that participants who use MOHSE, scored better than others. It shows that MOHSE is efficient and effective.

- According to outcomes of post questionnaire, MOHSE satisfied participants.

5.2 Link Evaluation

An evaluation framework for cross-lingual link discovery (CLLD) is created to automatically identify meaningful and relevant hypertext links between documents in different languages and used to assess runs submitted to the NTCIR (National Institute of Informatics Testbeds and Community for Information Access Research) [67]. This is particularly helpful in knowledge discovery if a multi-lingual knowledge base is sparse in one language or another, or the topical coverage in each language is different; such as the case with Wikipedia. This section of the chapter gives a brief information about the framework.

Evaluation methodology includes four parts: *Input, System, Output and Evaluation*.

Inputs is a test collection consisting of topics and target documents set. In our study test topics are small set of English Wikipedia pages and target documents are Turkish Wikipedia pages. Existing links in the topics pages are removed. The MOHSE system identifies the anchors in the source pages and tries to recommend related pages in target documents set.

System is our initial MOHSE implementation.

Output is an enriched topic page with Turkish links added by MOHSE system.

Evaluation includes assessment method and evaluation metrics. According to the framework there are two kinds of assessments: (i) automatic assessment and (ii) manual assessment. Under automatic assessment, the links that are already presented in the Wikipedia (existing link before link remove operation) are considered to be the ground-truth gold standard. Under the manual assessment, all links in all runs are pooled and assessed by a human judge.

Links in the Wikipedia

A link is a navigation entity within a single document that consists of two parts: an anchor and a target.

An anchor is a snippet of text associated with the link and it is relevant the topic of target article. Wikipedia anchors are often manually added and target only one destination. There are four types of links in Wikipedia:

- mono-lingual article-to-article ("see also") links;
- mono-lingual anchor-to-article links;
- cross-lingual article-to-article ("language") links;
- cross-lingual anchor-to-article links.

For automatic assessment, links should be separated according to their relevance. There are two types of links that are relevant in Wikipedia: (i) all the mono-lingual links from the target language version of the source article are considered relevant, and (ii) all the cross-lingual links from the mono-lingual links from the source article are considered as relevant.

Cross-lingual link discovery consist of two phases: (i) detecting prospective anchor in the source article; and (ii) identifying relevant articles in the target language. For evaluating this task, links on the pages are represented as:
$a_i \rightarrow (d_1, d_2 \cdots d_j); i < M; j < N$
where a_i is the *i*th anchor in the source document; d_j is target document *j* for the anchor; M is the number of anchors that are identified; and N is the number of of target links identified for each anchor.

Evaluation Metrics

Precision and recall are the two underlying key metrics used to measure system performance in Information Retrieval. For evaluating CLLD systems traditional definitions are extended for both anchors and targets.

File-to-file evaluation

$$Precision_{f2f} = \frac{foundandrelevant}{found}$$

$$Recall_{f2f} = \frac{foundandrelevant}{relevant}$$

Anchor-to-file evaluation

$$fanchor = \begin{cases} 1, & \text{if anchor is relevant} \\ 0, & \text{otherwise} \end{cases}$$

$$ftarget = \begin{cases} 1, & \text{if target is relevant} \\ 0, & \text{otherwise} \end{cases}$$

$$Precision_{link}(anchor) = f_{anchor} \times \frac{\sum_{j=1}^{k} f_{target_j}}{k}$$

$$Recall_{link}(anchor) = f_{anchor} \times \frac{\sum_{j=1}^{k} f_{target_j}}{N}$$

where N is the number of known relevant targets for the anchor. The precision for at some point, after n anchors, in the results lists is given by:

$$Precision_{a2f} = \frac{\sum_{i=1}^{n} Precision_{link}(i)}{n}$$

$$Recall_{a2f} = \frac{\sum_{i=1}^{n} Recall_{link}(i)}{M}$$

where M is the number of known relevant anchors.

5.2.1 Procedure

Links added by MOHSE system should be tested for detecting effectiveness of the system. In order to detect system effectiveness precision tests were done according the given framework [67].

Although cross-lingual link discovery consist of two phases: (i) detecting prospective anchor in the source article; and (ii) identifying relevant articles in the target language, we will not perform evaluation for detecting prospective anchor for MOHSE system because anchors are directly chosen by GEMET themes according to user choice.

For automatic assessment, internal links of the source pages should be removed. Relevant links should be determined according to relevance link structure of the Wikipedia,. Then evaluation tool should be implemented which calculates precision of the links which are added by MOHSE.

For manual assessment, process consist of two steps. Firstly, participants were informed about MOHSE and how to use it. Secondly, participants were requested to browse each page and enrich the page. After that participants assesses first five link for each anchor. Participants in turn filled the form in the Appendix F.

5.2.2 Materials

Our source language is English. Therefore, Web source from English Wikipedia should be determined for testing. First we chose the following 10 context main pages and 10 random term page for each main context for automatic and manual assessment.

http://en.wikipedia.org/wiki/Biology
http://en.wikipedia.org/wiki/Botany
http://en.wikipedia.org/wiki/Building
http://en.wikipedia.org/wiki/Dwelling
http://en.wikipedia.org/wiki/Chemistry
http://en.wikipedia.org/wiki/Electrolysis
http://en.wikipedia.org/wiki/Economics
http://en.wikipedia.org/wiki/Labor_force

http://en.wikipedia.org/wiki/Environmental_policy
http://en.wikipedia.org/wiki/Telemetry
http://en.wikipedia.org/wiki/Industry
http://en.wikipedia.org/wiki/Fermentation
http://en.wikipedia.org/wiki/Pollution
http://en.wikipedia.org/wiki/Distillation
http://en.wikipedia.org/wiki/Research
http://en.wikipedia.org/wiki/Climatology
http://en.wikipedia.org/wiki/Population
http://en.wikipedia.org/wiki/Primary_education
http://en.wikipedia.org/wiki/Water
http://en.wikipedia.org/wiki/Lagoon

However most of them are not suitable for assessment since they do not include enough anchor or link for testing. We choose the following Wikipedia pages for automatic and manual assessment because these pages include more anchor and link than others. Therefore, these are suitable for testing.

http://en.wikipedia.org/wiki/Economics
http://en.wikipedia.org/wiki/Water
http://en.wikipedia.org/wiki/Chemistry
http://en.wikipedia.org/wiki/Electrolysis
http://en.wikipedia.org/wiki/Biology
http://en.wikipedia.org/wiki/Botany

All Turkish Wikipedi pages should be crawled as target sources.

5.2.3 The Scientific Questions

Link evaluation conducted to examine the following hypotheses:

1. Most information found by MOHSE was relevant.
2. Manual and automatic assessment should give similar results.

For each hypotheses we included the following details:

- What do we try to find with this scientific question?
- What is the context of the data which we have related with this scientific question?
- Which techniques are used to evaluate the results of this scientific question?
- What are the results and how do we interpret them?

5.2.4 Participants

First ten participants of user evaluation evaluated the links as well.

5.2.5 Results

The Scientific Question I: Are links found by MOHSE relevant?
What we try to find: In this question, we try to show that most of the links found by MOHSE are relevant.
Context of the data: Data is links which are added to the Web pages by MOHSE.
Technique of evaluation: We simply calculate precision according to given framework.
Results: The results are given in Table 5.13, 5.14, 5.15, 5.16, 5.17 and 5.18.
Discussion: For economy page, most of the links added by MOHSE are relevant according the framework's precision calculation. Although two of the precision values (income, allocation) are calculated 0, two of them (interest, macroeconomics) are calculated 1 means %100 of the links are relevant for this anchor and context. Moreover, 4 of them are higher than 0.500 for both automatic and manual assessment. If water page is examined, results do not look good. Only 2 of the precision values are higher than 0.500 for both automatic and manual assessment. Moreover, 2 of them totally

are irrelevant (well, dam) for automatic assessment. Also, file to file assessment is not very successful. For chemistry page, automatic assessment seem not very successful however, manual assessment give good results. Half of the precision values higher than 0.500 and none of them are 0. According to electrolysis page's results, the links performance is average. Two of the calculated precision values are 0 (chemistry, platinum) for automatic assessment however, there is only 0 (platinum) for manual assessment. The links added for platinum anchor are related to platinum but they are not related to chemistry. Therefore, they were assumed irrelevant and then the precision was calculated 0. The other anchors give reasonable results. Biology page precision values are very well except from biology anchor. All the values are higher than 0.500 for manual assessment. It can be seen that MOHSE works well for biology page. Botany page gives fine result like biology page. Both automatic and manual assessment results are high. Moreover, for automatic assessment 8 of the results are higher than 0.500. To conclude, MOHSE was successful in adding links for all pages apart from water page. The big part of the precision values are higher than 0.500.

Table 5.13: Precision of links added to Economy page by MOHSE

Economy	Automatic Assessment	Manual Assessment
file to file	0.520	0.472
goods	0.600	0.400
interest	1.000	0.760
economy	0.200	0.340
income	0.000	0.000
cost	0.400	0.320
accounting	0.600	0.540
macroeconomics	1.000	0.920
price	0.600	0.580
monopoly	0.800	0.860
allocation	0.000	0.000

The Scientific Question II: Does manual assessment and automatic assessment give similar results?

What we try to find: In this question, we try to show that automatic assessment and manual assessment gives similar results.

Context of the data: Data is links which are added to the Web pages by MOHSE.

Table 5.14: Precision of links added to Water page by MOHSE

Water	Automatic Assessment	Manual Assessment
file to file	0.300	0.368
ice	0.400	0.160
limnology	0.200	0.480
hydrology	0.600	0.600
sea	0.600	0.480
river	0.400	0.220
irrigation	0.400	0.500
ocean	0.200	0.600
filtration	0.200	0.380
well	0.000	0.040
dam	0.000	0.220

Table 5.15: Precision of links added to Chemistry page by MOHSE

Chemistry	Automatic Assessment	Manual Assessment
file to file	0.220	0.512
chemistry	0.000	0.440
carbon	0.400	0.700
hydrogen	0.200	0.460
thermodynamics	0.600	0.660
nitrogen	0.200	0.660
spectroscopy	0.200	0.460
oxidation	0.200	0.540
nitrate	0.000	0.400
mineral	0.400	0.640
iron	0.000	0.200

Technique of evaluation: We simply calculate precision according to given framework and compare them.

Results: The results are given in Table 5.13, 5.14, 5.15, 5.16, 5.17 and 5.18.

Discussion: The economy page shows that automatic and manual assessment give similar results according to Table 5.13 as precision values are very close to each other. For water page, when the results of automatic and manual assessment are compared, it can be seen that manual assessment gives better results than other because some of

Table 5.16: Precision of links added to Electrolysis page by MOHSE

Electrolysis	Automatic Assessment	Manual Assessment
file to file	0.280	0.498
electrolysis	0.800	0.880
chemistry	0.000	0.440
hydrogen	0.800	0.460
acid	0.000	0.520
oxygen	0.200	0.620
salt	0.200	0.820
platinum	0.000	0.000
metal	0.200	0.180
oxidation	0.200	0.580
bromine	0.400	0.480

Table 5.17: Precision of links added to Biology page by MOHSE

Biology	Automatic Assessment	Manual Assessment
file to file	0.560	0.666
biology	0.000	0.000
evolution	0.800	0.740
botany	0.400	0.600
ecology	0.600	0.760
anatomy	0.400	0.560
animal	0.400	0.700
DNA	0.600	0.580
gene	1.000	0.880
ecosystem	0.600	0.920
photosynthesis	0.800	0.920

the links added by MOHSE do not make sense for English. However, they are meaningful in Turkish. To illustrate, for dam anchor, some of the dam placed in Turkey added to the anchor and automatic assessment is unable to recognize them. Thus, manual assessment gives better results. In chemistry page, all the precision values of manual assessment are higher than the values of automatic assessment. Electrolysis, biology, and botany pages also give similar results to those of chemistry page. All the manual assessment values are higher than automatic assessment values except few of

Table 5.18: Precision of links added to Botany page by MOHSE

Botany	Automatic Assessment	Manual Assessment
file to file	0.520	0.536
botany	0.400	0.620
DNA	0.800	0.580
genetics	0.800	0.540
morphology	0.400	0.640
anatomy	0.400	0.560
biology	0.000	0.000
ecology	0.600	0.760
leaf	0.800	0.740
photosynthesis	1.000	0.920
behavior	0.000	0.000

them. Therefore, we can conclude that manual assessment gives better results than automatic assessment. Thus we should refuse that automatic and manual assessment give similar results.

5.3 System Performance Evaluation

System's response time is one of the important aspect which affect user satisfaction [41]. Therefore, processing time of MOHSE should be measured. The concept's term size and anchor number affect the response time performance. In this section, response time and memory usage of MOHSE are examined.

5.3.1 Procedure

In order to measure the MOHSE's response time and memory usage according to anchor and term size, we designed two scenarios:

- In the first scenario, we created an HTML page where anchor size is fixed for each of the contexts. MOHSE processed the HTML page for each context;

therefore, we measured memory usage and response time with respect to term size of each context.

- In second scenario, we created 11 HTML pages. For first page MOHSE adds 0 anchor, for second page MOHSE adds 1 anchor, and for last page MOHSE adds 10 anchors. Then, we measured memory usage and response time with respect to number of anchor.

5.3.2 Materials

HTML test pages were created according to specific scenarios. We have 12 HTML pages in total. For measuring response time, we used Java Timestamp class [1]. For measuring memory usage, we used Apache's server status page. Our server runs under Ubuntu 13.10 operating system with Intel core i7-3770 CPU 3.40 Gz x 8 [2] and 16 GB memory configuration.

5.3.3 The Scientific Questions

System performance evaluation was conducted to examine the following hypotheses:

- MOHSE cache decreases response time.
- If context's term size increases, response time of MOHSE also increases.
- If context's term size increases, memory usage of MOHSE also increases.
- If anchor size increases, response time of MOHSE also increases.
- If anchor size increases, memory usage of MOHSE also increases.

[1] Java Class Timestamp,http://docs.oracle.com/javase/7/docs/api/java/sql/Timestamp.html

[2] Intel® Core™ i7-3770 Processor, http://ark.intel.com/tr/products/65719/Intel-Core-i7-3770-Processor-8M-Cache-up-to-3-90_GHz

5.3.4 Results

The Scientific Question I: Does MOHSE cache decrease response time?
What we try to find: In this question, we try to show that MOHSE cache decreases response time.
Context of the data: Data is response time of MOHSE for each scenario.
Technique of evaluation: We simply measure response time of MOHSE for processing page with using cache and without using cache in milliseconds.
Results: The results are given in Table 5.20 and 5.19.
Discussion: It seems clear that MOHSE cache decreases response times significantly.

Table 5.19: Response time and memory usage for each context

Context	Term size	First response time	Cache response time	Memory usage
Biology	632	1840	21	70.23
Building	311	1890	30	59.68
Chemistry	574	1863	19	65.33
Economics	475	1927	25	59.64
Environmental Policy	441	1863	24	58.89
Industry	560	1906	24	61.71
Pollution	529	1836	19	61.25
Research	417	1893	19	57.08
Social Aspects	466	1900	17	59.14
Water	543	1772	18	61.19

The Scientific Question II: If context's term size increases, does response time of MOHSE increase?
What we try to find: In this question, we try to show that if context's term size increases, response time of MOHSE also increases.
Context of the data: Our data is response time of MOHSE for second scenario which context's term size increases for each run.
Technique of evaluation: We simply measure response time of MOHSE for processing page with using cache and without using cache in milliseconds.

Results: The results are given in Table 5.19, Figure 5.1 and 5.2.

Discussion: Although both of the response time fluctuates according to context's term size, we can not say that term size affect the both response time. The response times are very close the each other, although term size changes.

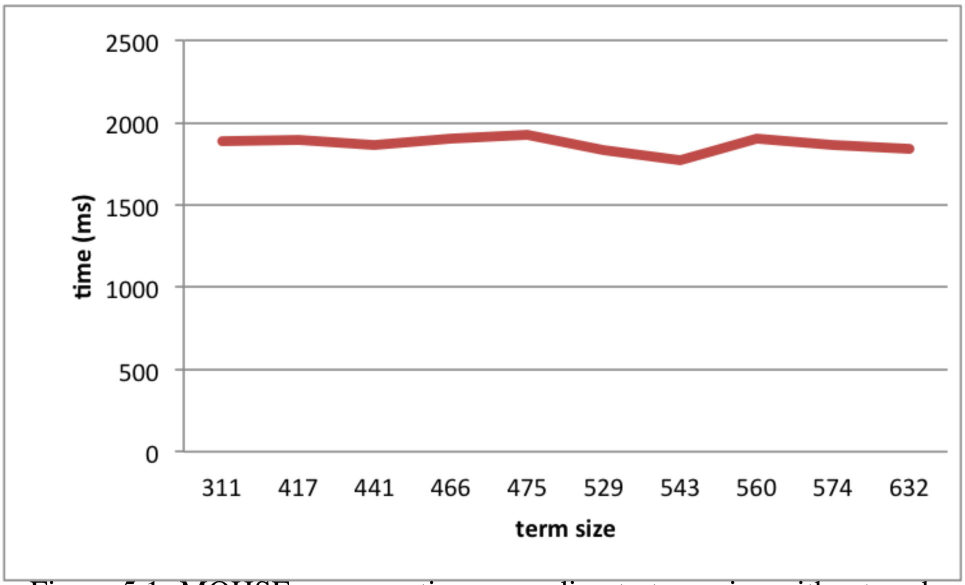


Figure 5.1: MOHSE response time according to term size without cache

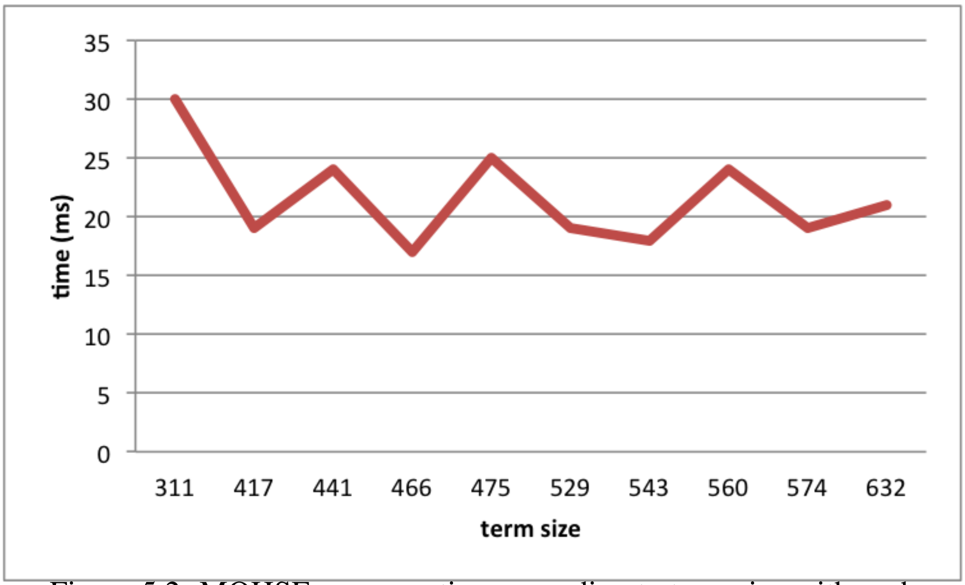


Figure 5.2: MOHSE response time according to term size with cache

The Scientific Question III: If context's term size increases, does memory usage of MOHSE increase?

What we try to find: In this question, we try to find that if context's term size increases, memory usage of MOHSE also increase.

Context of the data: Our data is memory usage of MOHSE for second scenario which context's term size increases for each run.

Technique of evaluation: We simply measure memory usage of MOHSE for processing page with using cache and without using cache in megabytes.

Results: The results are given in Table 5.19 and Figure 5.3.

Discussion: According the results, it is clear that memory usage of MOHSE increases according to context's term size, however it is too slightly.

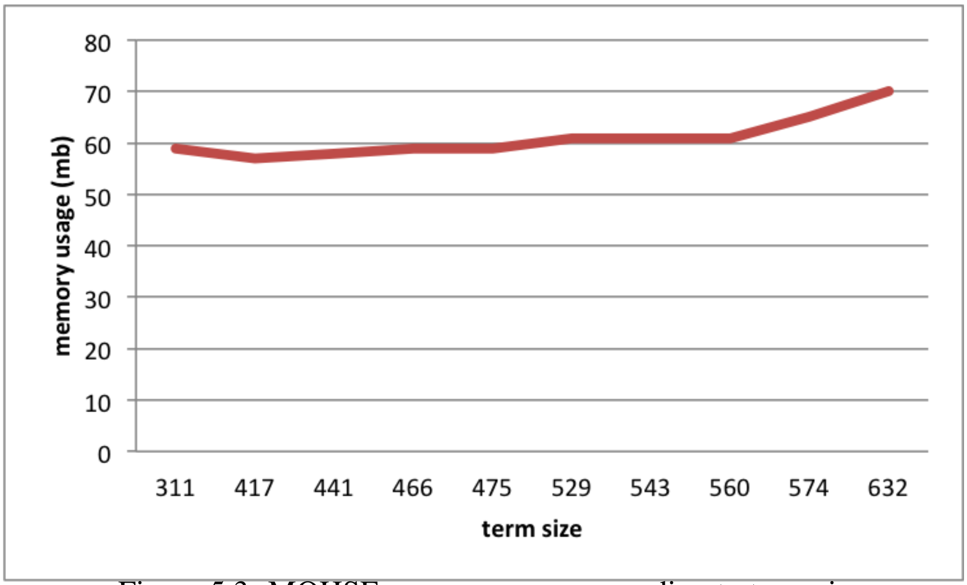


Figure 5.3: MOHSE memory usage according to term size

Table 5.20: Response time and memory usage according to anchor size for Biology context

Anchor	First response time	Cache Response Time	Memory Usage
0	1840	21	51.89
1	1955	53	70.46
2	2072	83	92.58
3	2139	76	109.40
4	2455	82	147.21
5	2413	112	187.18
6	2283	134	200.41
7	2292	150	209.96
8	2335	153	215.21
9	2381	164	217.33
10	2401	171	235.54

The Scientific Question IV: If anchor size increases, does response time of MOHSE increase?

What we try to find: In this question, we try to find that if anchor size increases, response time of MOHSE also increases.

Context of the data: Our data is response time of MOHSE for first scenario which anchor size increases for each run.

Technique of evaluation: We simply measure response time of MOHSE for processing page with using cache and without using cache in milliseconds.

Results: The results are given in Table 5.20, Figure 5.4 and 5.5.

Discussion: Response time with cache and response time without cache increase according to anchor size. The increases of response time is more evident for systems with cache when compared with systems without cache as number of anchors increase.

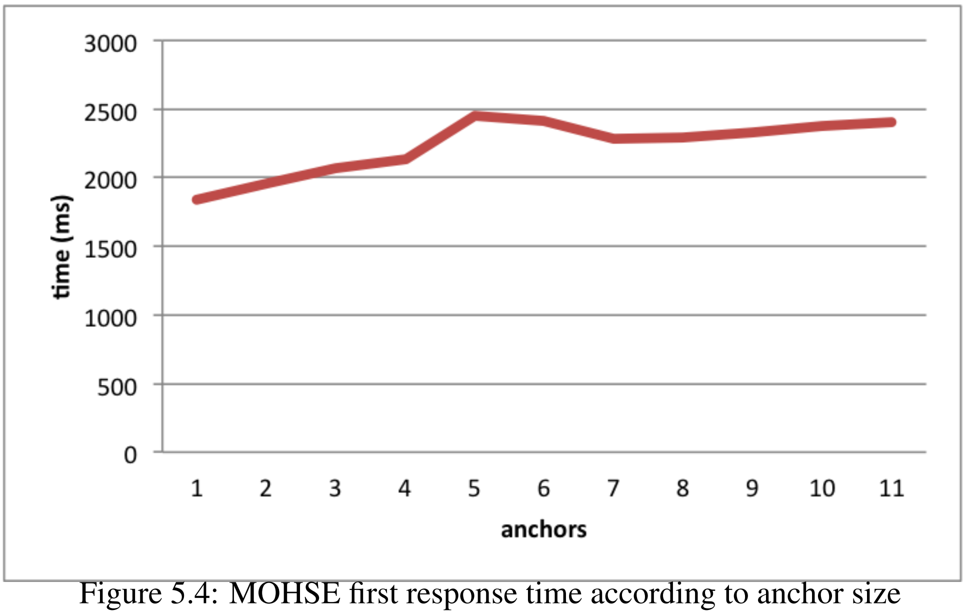


Figure 5.4: MOHSE first response time according to anchor size

The Scientific Question V: If anchor size increases, does memory usage of MOHSE increase?

What we try to find: In this question, we try to find that if context's term size increases, memory usage of MOHSE also increase.

Context of the data: Our data is memory usage of MOHSE for second scenario which context's term size increases for each run.

Technique of evaluation: We simply measure memory usage of MOHSE for pro-

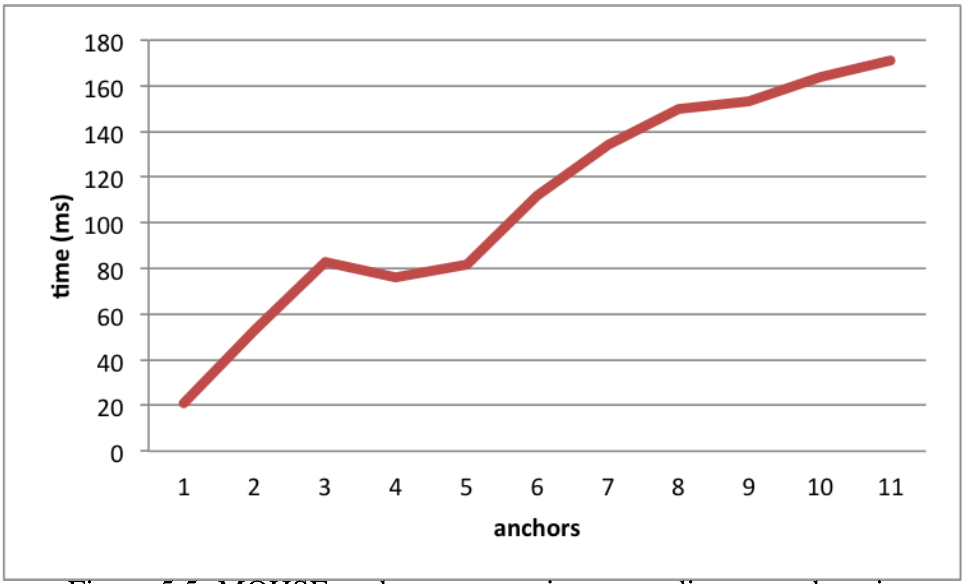


Figure 5.5: MOHSE cache response time according to anchor size

cessing page with using cache and without using cache in megabytes.

Results: The results are given in Table 5.20 and Figure 5.6.

Discussion: Although memory usage decreased in some run, we can say that anchor size increase memory usage. If we multiply term size to ten, memory usage doubled according to our test cases.

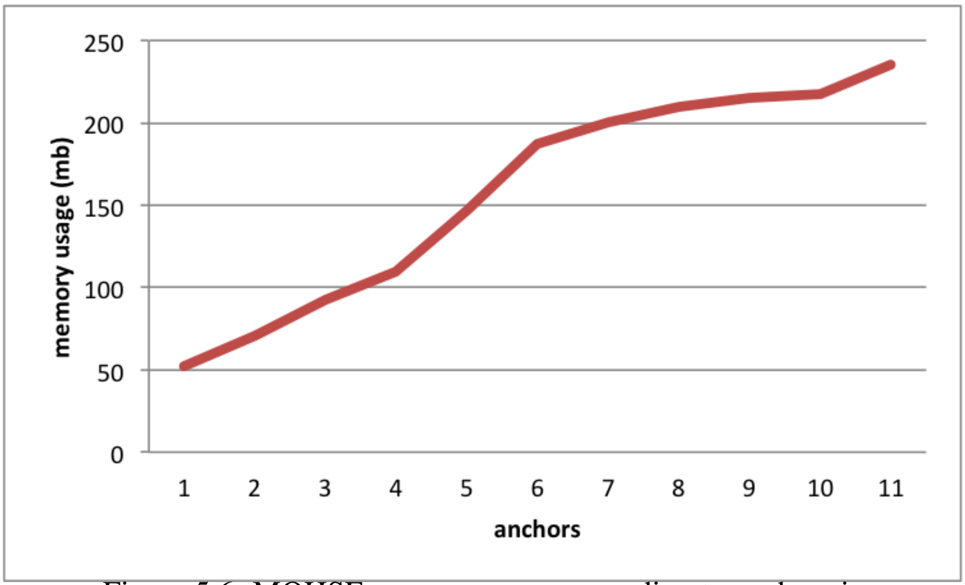


Figure 5.6: MOHSE memory usage according to anchor size

5.4 Summary

There are three step considered for evaluation process which are:

- **User Evaluation** is aiming to test system effectiveness, efficiency, and satisfaction. To achieve these, we conducted a user survey which includes several steps with 40 people. First step is introduction. In this step we informed user about MOHSE. Then we characterized users. Second step short test which user supposed to solve by using MOHSE. Lastly, we gave user a short questionnaire to detect user satisfaction. According to our experiments, MOHSE increases participants' Web usage performance. Participants' answers show that MOHSE works efficiently, and satisfy them.
- **Link Evaluation** is designed to test quality of links added by MOHSE. We used a framework which enable us to test relevance of links according to Wikipedia. In most of the cases, MOHSE added relevant links according to the context of the resources.
- **System Performance Evaluation** test the memory usage and response time of the system according to term and anchor size. Results show that MOHSE's performance is acceptable.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 Conclusion

This thesis presents the Msc project called Multilingual Open Hypermedia Service (MOHSE). The overall aim of this thesis is multilingual dynamic linking of Web resources. To achieve this aim, we had the following objectives:

1. Research the state of the art multilingual controlled vocabularies, their representation, and their use in dynamic linking of Web resources.
2. Investigate existing multilingual controlled vocabularies that include terms from both Turkish and English.
3. Extend COHSE infrastructure to dynamically link Web resources in Turkish and English.
4. Experiment with this extended infrastructure to demonstrate how multilingual dynamic linking can be achieved.

Firstly, we described what is Semantic Web. With the advances of the Semantic Web, the embedded link structure of the Web, which is one of the limitations of the current Web, can be improved and extended. Web resources are annotated with semantic mark-up, using knowledge representation languages, such as RDF or OWL, we showed that Web resources can be linked dynamically.

Secondly, we focused on researching the existing literature on dynamic multilingual linking of Web resources, and multilingual controlled vocabularies. The overall method was bases on the following two task: (i) literature review of multilingual dynamic linking and (ii) literature review of multilingual controlled vocabularies. We showed that there are no other systems that aim to do multilingual dynamic linking, and showed that there are a number of manually built large multilingual controlled vocabularies. However, we realized that not many of these international multilingual vocabularies include Turkish terms. One of the largest controlled vocabulary which include support to Turkish is GEMET. Latest version of GEMET includes more than 6,000 terms in 27 languages. GEMET is also available in SKOS format; therefore, it can easily be used with MOHSE.

Thirdly, we described MOHSE design requirement. We analyzed COHSE infrastructure to extend it for supporting multilingualism. MOHSE's architecture is an impressed and extended version of the COHSE's architecture. MOHSE has four components: (i) Knowledge Service is responsible for interaction with ontologies, (ii) Resource Service is an web search agent and crawler which crawls web resources continuously and maps terms to these resources, (iii) Firefox Extension controls the user's interaction with MOHSE, and (iv) MOHSE Server integrates all the components.

Lastly, we described MOHSE evaluation process which includes three parts. In first experiment, we conducted a detailed user survey with 40 participant, and then we concluded that MOHSE increased participants Web usage performance. Second experiment measured quality of links added by MOHSE. In most of the cases, MOHSE added relevant links according to the context of the resources. Last experiment includes tests which evaluate system performance. We measured memory usage and response time of MOHSE. Outcomes show that MOHSE works in an acceptable performance.

6.2 Future Work

MOHSE initial application was completed and deployed as a browser plugin. However it has some problems. Firstly, MOHSE adds Turkish links to English web pages in its current implementation. It could be extended for both way. This can be easily done in current architecture. If resource service crawls English web pages, this extension could be completed. Another problem occurs when users want to add new ontology to existing ones. Although MOHSE Knowledge Service support this feature, browser plugin could be extended for supporting it.

MOHSE offers a single Resource Service now. It could be extended according to context. For example if users select to enrich web pages according to biology context, only the links about biology can be added. For doing this, multiple resource service could be offered like selecting the context.

Another extension point of MOHSE is that, users could be offered a context according to visiting page. Then users are prevented from selecting irrelevant contexts.

REFERENCES

[1] Alper Aksac, Orkun Ozturk, and Erdogan Dogdu. A novel semantic web browser for user centric information retrieval: Person. *Expert Systems with Applications*, 39(15):12001–12013, 2012.

[2] Şerbetçi Ayşe, Orhan Zeynep, and Pehlivan İlknur. Extraction of semantic word relations in turkish from dictionary definitions. *ACL HLT 2011*, page 11, 2011.

[3] Sean Bechhofer, Robert Stevens, Phillip Lord, et al. Ontology driven dynamic linking of biology resources. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, page 79, 2005.

[4] Sean Bechhofer, Frank Van Harmelen, Jim Hendler, Ian Horrocks, Deborah L McGuinness, Peter F Patel-Schneider, Lynn Andrea Stein, et al. Owl web ontology language reference. *W3C recommendation*, 10:2006–01, 2004.

[5] Sean Bechhofer, Yeliz Yesilada, Bernard Horan, and Carole Goble. Knowledge-driven hyperlinks: Linking in the wild. In *Adaptive Hypermedia and Adaptive Web-Based Systems*, pages 1–10. Springer, 2006.

[6] Sean Bechhofer, Yeliz Yesilada, Robert Stevens, Simon Jupp, and Bernard Horan. Using ontologies and vocabularies for dynamic linking. *Internet Computing, IEEE*, 12(3):32–39, 2008.

[7] Richard Benjamins, Jesus Contreras, Oscar Corcho, and Asuncion Gomez-Perez. The six challenges of the semantic web. *AIS SIGSEMIS Bulletin*, 1, 2002.

[8] Tim Berners-Lee, James Hendler, Ora Lassila, et al. The semantic web. *Scientific american*, 284(5):28–37, 2001.

[9] Orhan Bilgin, Özlem Çetinoğlu, and Kemal Oflazer. Building a wordnet for turkish. *Romanian Journal of Information Science and Technology*, 7(1-2):163–172, 2004.

[10] Niels Olof Bouvin. Unifying strategies for web augmentation. In *Proceedings of the tenth ACM Conference on Hypertext and hypermedia: returning to our diverse roots: returning to our diverse roots*, pages 91–100. ACM, 1999.

[11] Dan Brickley, Ramanathan V Guha, and Brian McBride. Rdf vocabulary description language 1.0: Rdf schema. w3c recommendation (2004), 2004.

[12] Paul Buitelaar, Thomas Eigner, and Stefania Racioppa. Semantic navigation with views. In *UserSWeb: Workshop on User Aspects of the Semantic Web, Crete*. Citeseer, 2005.

[13] Fazli Can, Seyit Kocberber, Erman Balcik, Cihan Kaynak, H Cagdas Ocalan, and Onur M Vursavas. Information retrieval on turkish texts. *Journal of the American Society for Information Science and Technology*, 59(3):407–421, 2008.

[14] Leslie Carr, Wendy Hall, Sean Bechhofer, and Carole Goble. Conceptual linking: ontology-based open hypermedia. In *Proceedings of the 10th international conference on World Wide Web*, pages 334–342. ACM, 2001.

[15] Leslie Carr, Simon Kampa, Wendy Hall, Sean Bechhofer, and Carole Goble. Handbook on ontologies, chapter cohse: Conceptual open hypermedia service, 2004.

[16] Erbug Celebi, Baturman Sen, and Burak Gunel. Turkish—english cross language information retrieval using lsi. In *Computer and Information Sciences, 2009. ISCIS 2009. 24th International Symposium on*, pages 634–638. IEEE, 2009.

[17] Chang Chun and Lu Wenlin. The translation of agricultural multilingual thesaurus. In *Proceedings of the Third Asian Conference for Information Technology in Agriculture*, 2002.

[18] W Bruce Croft, John Broglio, and Hideo Fujii. Applications of multilingual text retrieval. In *System Sciences, 1996., Proceedings of the Twenty-Ninth Hawaii International Conference on,*, volume 5, pages 98–107. IEEE, 1996.

[19] Daniel Cunliffe, Carl Taylor, and Douglas Tudhope. Query-based navigation in semantically indexed hypermedia. In *Proceedings of the eighth ACM conference on Hypertext*, pages 87–95. ACM, 1997.

[20] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. Gate: an architecture for development of robust hlt applications. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 168–175, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[21] Mathieu d'Aquin, Marta Sabou, Martin Dzbor, Claudio Baldassarre, Laurian Gridinoc, Sofia Angeletou, and Enrico Motta. Watson: a gateway for the semantic web. 2007.

[22] Hugh Davis, Wendy Hall, Ian Heath, Gary Hill, and Rob Wilkins. Towards an integrated information environment with open hypermedia systems. In *Proceedings of the ACM conference on Hypertext*, pages 181–190. ACM, 1992.

[23] Gerard De Melo and Gerhard Weikum. Towards a universal wordnet by learning from combined evidence. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 513–522. ACM, 2009.

[24] Gayo Diallo, Khaled Khelif, Olivier Corby, Patty Kostkova, and Gemma Madle. Semantic browsing of a domain specific resources: The corese-neli framework. In *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on*, volume 3, pages 50–54. IEEE, 2008.

[25] Martin Dzbor, Enrico Motta, and John Domingue. Magpie: Experiences in supporting semantic web browsing. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(3):204 – 222, 2007.

[26] İlknur Durgar El-Kahlout and Kemal Oflazer. Initial explorations in english to turkish statistical machine translation. In *Proceedings of the Workshop on Statistical Machine Translation*, StatMT '06, pages 7–14, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[27] Mauricio Espinoza, Asunción Gómez-Pérez, and Elena Montiel-Ponsoda. Multilingual and localization support for ontologies. In *The Semantic Web: Research and Applications*, pages 821–825. Springer, 2009.

[28] Enrico Francesconi, Sebastiano Faro, and Elisabetta Marinai. A framework for semantic mapping between thesauri. In *Proceedings of the 2nd international conference on Theory and practice of electronic governance*, pages 251–257. ACM, 2008.

[29] John H Gennari, Mark A Musen, Ray W Fergerson, William E Grosso, Monica Crubézy, Henrik Eriksson, Natalya F Noy, and Samson W Tu. The evolution of protégé: an environment for knowledge-based systems development. *International Journal of Human-computer studies*, 58(1):89–123, 2003.

[30] Carole A Goble, Sean Bechhofer, Les Carr, David De Roure, and Wendy Hall. Conceptual open hypermedia= the semantic web? In *SemWeb*, 2001.

[31] Laurian Gridinoc, Mathieu d'Aquin, Davide Guidi, Martin Dzbor, and Enrico Motta. Powermagpie: A semantic web browser–v1. 2007.

[32] Catherine Havasi, Robert Speer, and Jason Alonso. Conceptnet 3: a flexible, multilingual semantic network for common sense knowledge. In *Recent Advances in Natural Language Processing*, pages 27–29, 2007.

[33] Mathew Hillier. The role of cultural context in multilingual website usability. *Electronic Commerce Research and Applications*, 2(1):2–14, 2003.

[34] Ian Horrocks and Sean Bechhofer. Semantic web. Human-Computer Interaction Series, chapter 19, pages 315–330. Springer, London, 1st edition, September 2008.

[35] Shihong Huang and Scott Tilley. Issues of content and structure for a multilingual web site. In *Proceedings of the 19th annual international conference on Computer documentation*, pages 103–110. ACM, 2001.

[36] David Huynh, Stefano Mazzocchi, and David Karger. Piggy bank: Experience the semantic web inside your web browser. In *The Semantic Web–ISWC 2005*, pages 413–430. Springer, 2005.

[37] Jason J Jung, Anne Håkansson, and Ronald Hartung. Indirect alignment between multilingual ontologies: A case study of korean and swedish ontologies. In *Agent and Multi-Agent Systems: Technologies and Applications*, pages 233–241. Springer, 2009.

[38] Simon Jupp, Sean Bechhofer, Patty Kostkova, Robert Stevens, and Yeliz Yesilada. Document navigation: Ontologies or knowledge organisation systems?

[39] Simon Jupp, Robert Stevens, Sean Bechhofer, Yeliz Yesilada, and Patty Kostkova. Knowledge representation for web navigation. In *Semantic Web Applications and Tools for the Life Sciences (SWAT4LS 2008) Workshop*, 2008.

[40] Hiroyuki Kaji and Toshiko Aizono. Extracting word correspondences from bilingual corpora based on word co-occurrences information. In *Proceedings of the 16th conference on Computational linguistics-Volume 1*, pages 23–28. Association for Computational Linguistics, 1996.

[41] Diane Kelly. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3:1–224, January 2009.

[42] Atanas Kiryakov, Borislav Popov, Ivan Terziev, Dimitar Manov, and Damyan Ognyanoff. Semantic annotation, indexing, and retrieval. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2(1):49–79, 2004.

[43] Kazuaki Kishida. Technical issues of cross-language information retrieval: a review. *Information Processing & Management*, 41(3):433–455, 2005.

[44] Ray R Larson, Fredric Gey, and Aitao Chen. Harvesting translingual vocabulary mappings for multilingual digital libraries. In *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, pages 185–190. ACM, 2002.

[45] Anita C Liang and Margherita Sini. Mapping agrovoc and the chinese agricultural thesaurus: definitions, tools, procedures. *New Review of Hypermedia and Multimedia*, 12(1):51–62, 2006.

[46] David Lowe and Wendy Hall. *Hypermedia & the Web.* Wiley Chichester, 1999.

[47] Alistair Miles and Sean Bechhofer. Skos simple knowledge organization system reference. Technical report, Technical report, W3C, 2009.

[48] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[49] Teruko Mitamura. Controlled language for multilingual machine translation. In *Proceedings of Machine Translation Summit VII, Singapore*, pages 46–52, 1999.

[50] Elena Montiel-Ponsoda, G Aguado de Cea, Asunción Gómez-Pérez, and Wim Peters. Enriching ontologies with multilingual information. *Natural language engineering*, 17(03):283–309, 2011.

[51] Elena Montiel-Ponsoda, Jorge Gracia del Río, Guadalupe Aguado de Cea, and Asunción Gómez-Pérez. Representing translations on the semantic web. 2011.

[52] Terri Morgan, Carol Luttrell, and Yuzeng Liu. Designing multilingual web sites: applied authoring techniques. In *Proceedings of the 19th annual international conference on Computer documentation*, pages 230–231. ACM, 2001.

[53] Jocelyne Nanard and Marc Nanard. Using structured types to incorporate knowledge in hypertext. In *Proceedings of the third annual ACM conference on Hypertext*, pages 329–343. ACM, 1991.

[54] Roberto Navigli and Simone Paolo Ponzetto. Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225. Association for Computational Linguistics, 2010.

[55] Douglas W Oard and Bonnie J Dorr. A survey of multilingual text retrieval. 1998.

[56] Kemal Oflazer and Ilknur Durgar El-Kahlout. Exploring different representational units in english-to-turkish statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 25–32, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

[57] Helen Oliver, Gayo Diallo, Ed de Quincey, Dimitra Alexopoulou, Bianca Habermann, Patty Kostkova, Michael Schroeder, Simon Jupp, Khaled Khelif, Robert Stevens, et al. A user-centred evaluation framework for the sealife semantic web browsers. *BMC bioinformatics*, 10(Suppl 10):S14, 2009.

[58] Kasper Østerbye and Uffe Kock Wiil. The flag taxonomy of open hypermedia systems. In *Proceedings of the the seventh ACM conference on Hypertext*, pages 129–139. ACM, 1996.

[59] Borislav Popov, Atanas Kiryakov, Angel Kirilov, Dimitar Manov, Damyan Ognyanoff, and Miroslav Goranov. Kim–semantic annotation platform. In *The Semantic Web-ISWC 2003*, pages 834–849. Springer, 2003.

[60] Erhard Rahm and Philip A Bernstein. A survey of approaches to automatic schema matching. *the VLDB Journal*, 10(4):334–350, 2001.

[61] Melike Şah, Wendy Hall, and David C De Roure. Dynamic linking and personalization on web. In *Proceedings of the 2010 ACM Symposium on Applied Computing*, pages 1404–1410. ACM, 2010.

[62] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, Inc., New York, NY, USA, 1986.

[63] Michael Schroeder, A Burger, Patty Kostkova, Robert Stevens, Bianca Habermann, and Rose Dieng-Kuntz. From a services-based escience infrastructure to a semantic web for the life sciences: The sealife project. In *Proceedings of the Sixth International Workshop NETTAB 2006 on "Distributed Applications, Web Services, Tools and GRID Infrastructures for Bioinformatics*, 2006.

[64] Push Singh, Thomas Lin, Erik T Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. Open mind common sense: Knowledge acquisition from the general public. In *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE*, pages 1223–1237. Springer, 2002.

[65] Dagobert Soergel, Boris Lauser, Anita Liang, Frehiwot Fisseha, Johannes Keizer, and Stephen Katz. Reengineering thesauri for new applications: the agrovoc example. *Journal of digital information*, 4(4), 2006.

[66] Huatong Sun. Building a culturally-competent corporate web site: an exploratory study of cultural markers in multilingual web design. In *Proceedings of the 19th annual international conference on Computer documentation*, pages 95–102. ACM, 2001.

[67] Ling-Xiang Tang, Shlomo Geva, Andrew Trotman, Yue Xu, and Kelly Y Itakura. An evaluation framework for cross-lingual link discovery. *Information Processing & Management*, 50(1):1–23, 2014.

[68] Linda Tauscher and Saul Greenberg. How people revisit web pages: Empirical findings and implications for the design of history systems. *International Journal of Human-Computer Studies*, 47(1):97–137, 1997.

[69] Paolo Tonella, Filippo Ricca, Emanuele Pianta, and Christian Girardi. Restructuring multilingual web sites. In *Software Maintenance, 2002. Proceedings. International Conference on*, pages 290–299. IEEE, 2002.

[70] Paolo Tonella, Filippo Ricca, Emanuele Pianta, and Christian Girardi. Automatic support for the alignment of multilingual web sites. *Journal of Software Maintenance and Evolution: Research and Practice*, 18(3):153–179, 2006.

[71] Cássia Trojahn, Paulo Quaresma, and Renata Vieira. A framework for multilingual ontology mapping. 2008.

[72] Dan Tufis, Dan Cristea, and Sofia Stamou. Balkanet: Aims, methods, results and perspectives. a general overview. *Romanian Journal of Information science and technology*, 7(1-2):9–43, 2004.

[73] Cigdem Keyder Turhan. An english to turkish machine translation system using structural mapping. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, ANLC '97, pages 320–323, Stroudsburg, PA, USA, 1997. Association for Computational Linguistics.

[74] Victoria Uren, Enrico Motta, Martin Dzbor, and Philipp Cimiano. Browsing for information by highlighting automatically generated annotations: a user study and evaluation. In *Proceedings of the 3rd international conference on Knowledge capture*, pages 75–82. ACM, 2005.

[75] Piek Vossen. *EuroWordNet: a multilingual database with lexical semantic networks*. Kluwer Academic Boston, 1998.

[76] Reyyan Yeniterzi and Kemal Oflazer. Syntax-to-morphology mapping in factored phrase-based statistical machine translation from english to turkish. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 454–464, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[77] Yeliz Yesilada, Sean Bechhofer, and Bernard Horan. Personalised dynamic links on theweb. In *Semantic Media Adaptation and Personalization, 2006. SMAP'06. First International Workshop on*, pages 7–12. IEEE, 2006.

[78] Yeliz Yesilada, Sean Bechhofer, and Bernard Horan. Dynamic linking of web resources: Customisation and personalisation. In *Advances in Semantic Media Adaptation and Personalization*, pages 1–24. Springer, 2008.

APPENDIX A

INFORMATION SHEET

Middle East Technical University

Computer Engineering Department

Multilingual Dynamic Linking of Web Resources

Participant Information Sheet

Introduction

This study evaluates the results of the MOHSE project, which enriches English Web resources with external Turkish links semantically.

Why have I been chosen?

I am inviting anyone who is computer literate and between the ages of 18 and 35, and know Turkish and English to take part in the evaluation if they want.

What would I be asked to do if I take part?

You will be asked to fill a short questionnaire about your demographic information and Web experience. Next, you will be required to complete a simple task, which includes 6 simple questions about the economy and 6 simple questions about the biology fields. You will use Wikipedia and MOHSE to find answers of these questions. At the end, your opinions about to MOHSE will be asked.

How is confidentiality maintained?

Data will be made anonymous (names and any other information that may identify an individual will not be included), so no one will be able to recognize whom the data belongs to.

Do I have to take part?

You do not have to take part in the study. If you decide to take part and then later change your mind, either before you start the study or during it, you can withdraw without giving your reasons, and, if you wish, your data will be destroyed.

Will the outcomes of the research be published?

The outcomes of the research will be published in my thesis and conference proceedings and journal articles.

Contact

For further information, please contact:
Ugur Donmez
e-mail: donmez@metu.edu.tr

APPENDIX B

CONSENT FORM

Middle East Technical University

Computer Engineering Department

Multilingual Dynamic Linking of Web Resources

Consent Form

If you are happy to participate please complete and sign the consent form below

	Please Initial Box
1. I confirm that I have read the attached information sheet on the above project and have had the opportunity to consider the information and ask questions and had these answered satisfactory.	□
2. I understand that my participation in the study is voluntary and that I am free to withdraw at any time without giving a reason.	□
3. I agree to use of anonymous quotes.	□
4. I understand that my performance will be entirely for research purposes and will not affect in any way my grades.	□

I agree to take part in the above project.

Name of participant	Date	Signature
□	□	□

Name of person taking consent	Date	Signature
□	□	□

APPENDIX C

INITIAL QUESTIONNAIRE

Middle East Technical University

Computer Engineering Department

Multilingual Dynamic Linking of Web Resources

Initial Questionnaire

1. What is your gender?
   - ○ Female
   - ○ Male

2. What is your age?

3. How often do you use the Web?
   - ○ Daily
   - ○ Weekly
   - ○ Monthly
   - ○ Less than once a month
   - ○ Never

4. Highest level of education you have completed:
   - ○ Grade / Primary School
   - ○ High / Secondary School
   - ○ Associates Degree
   - ○ Bachelors Degree
   - ○ Masters Degree
   - ○ Doctorate
   - ○ Other

5. How often do you use the Wikipedia?
   - ○ Daily
   - ○ Weekly
   - ○ Monthly
   - ○ Less than once a month
   - ○ Never

6. When you searching for something on the Web, do you prefer English or Turkish?
   - ○ English
   - ○ Turkish
   - ○ Both

7. What is your department? What is your year?

APPENDIX D

TASKS

Middle East Technical University

Computer Engineering Department

Multilingual Dynamic Linking of Web Resources

Task A

Answer the following questions according to economy page in Wikipedia with help of MOHSE.

1. How does a country measure domestic aggregate production of goods and services?
   a. Gross domestic product
   b. Gross national income
   c. Gross domestic input
   d. Aggregate household income

2. When tourists from developed countries visit developing countries, they discover identical goods and services cheaper. Why is that the case?
   a. Gross national income
   b. Gross domestic product
   c. Consumer preferences
   d. Purchasing power parity

3. How do we refer to an enterprise that is the only seller of a goods and services?
   a. Free market
   b. Oligopoly
   c. Monopoly
   d. Duopoly

4. Üretim maliyeti arttığında, tüketim pazarındaki genel fiyatlar da artar. Fiyatlardaki bu artışın iktisadi tanımı nedir?
   a. Deflasyon
   b. Enflasyon
   c. Katma deger vergisi
   d. Devaluasyon

5. Bu yıl, geçen yılki ile aynı maaşın (kıdem vs ayarlanmadan, nominal olarak) alınmasına rağmen yaşam giderleri artmıştır. Neden?
    a. Deflasyon
    b. Enflasyon
    c. Katma deger vergisi
    d. Devaluasyon

6. Reel ve nominal faiz arasındaki fark nedir?
    a. Hiçbir fark yoktur.
    b. Reel faiz enflasyonu hesaplarken, nominal faiz hesaplamaz.
    c. Nominal faiz enflasyonu hesaplarken, reel faiz hesaplamaz.
    d. Nominal faiz oranı yalnızca ders kitaplarında bulunur.

Answer the following questions according to biology page in Wikipedia.

1. What is the name of organelle, which is responsible for photosynthesis in plants?
    a. Mitochondria
    b. Chromoplast
    c. Chloroplast
    d. Leucoplast

2. What is the study of plants?
    a. Morphology
    b. Botany
    c. Zoology
    d. Ecology

3. What is the process of organism change?
    a. Evoluation
    b. Organism
    c. Growth
    d. Genetics

4. Organizmalar ____________ ‘ları yakından ilişkili ise aynı grupta sınıflanır.
    a. Renk
    b. Yeme aliskanligi
    c. DNA
    d. Boyut

5. Hangisi bir ekosistem tipi değildir?
    a. Kayalik
    b. Mera
    c. Tundra
    d. Savana

6. Zatürree, gıda zehirlenmesi ve kan zehirlenmesi (septisemi)’ne _______________ sebep olur.
    a. Hayvanlar
    b. Bitkiler
    c. Bakteriler
    d. Virusler

Middle East Technical University

Computer Engineering Department

Multilingual Dynamic Linking of Web Resources

Task B

Answer the following questions according to economy page in Wikipedia with help of MOHSE.

1. Bir ülke toplam mal ve hizmet üretimini nasıl ölçer?
   a. Gayrısafi yurt içi hasıla
   b. Toplam milli gelir
   c. Gayrısafi yurt içi girdi
   d. Toplam hane geliri

2. Bir malin veya hizmetin tek saticinisin oldugu girisimi nasil adlandiririz?
   a. Serbest piyasa
   b. Oligopol
   c. Monopol
   d. Duopol

3. Gelişmiş ülkelerden gelen turistler gelişmekte olan ülkeleri ziyarete geldiklerinde aynı mal ve hizmetleri daha ucuza bulurlar. Bu durumun sebebi nedir?
   a. Satınalma gücü paritesi
   b. Toplam milli gelir
   c. Gayrısafi yurt içi hasıla
   d. Tüketici tercihi

4. When the cost of production increases, the general price levels in the consume market go up. What is an economic term for increase in price levels?
   a. Inflation
   b. Deflation
   c. Value added tax (vat)
   d. Devaluation

5. What is the difference between real and nominal interest?
    a. There is no difference.
    b. Real account for inflation but nominal does not.
    c. Nominal account for inflation but real does not.
    d. Nominal interest rate exists only in textbooks.

6. Assuming that you earn identical (unadjusted, nominal terms) salary this year from the previous year, the cost of living has increased. Why is that case?
    a. Inflation
    b. Deflation
    c. Value added tax (vat)
    d. Devaluation

Answer the following questions according to biology page in Wikipedia.

1. Bitkilerde fotosentezden sorumlu organelin adı nedir?
    a. Kloroplast
    b. Kromoplast
    c. Lökoplast
    d. Mitokondri

2. Organizmanın değişim sürecine ne denir?
    a. Evrim
    b. Organizma
    c. Buyume
    d. Genetik

3. Bitki bilimi nedir?
    a. Botanik
    b. Zooloji
    c. Ekoloji
    d. Morfoloji

4. Pneumonia, food poisoning and blood poisoning (sepsis) is caused by ....
    a. Animals
    b. Plants
    c. Bacterium
    d. Viruses

5. Organisms are classified under the same group if their..... is closely related.
    a. Color
    b. Size
    c. DNA
    d. Food

6. Which one of the following is not an ecosystem type?
    a. Tundra
    b. Rocky
    c. Savanna
    d. Grassland

Middle East Technical University

Computer Engineering Department

Multilingual Dynamic Linking of Web Resources

Task C

Answer the following questions according to economy page in Wikipedia.

1. When tourists from developed countries visit developing countries, they discover identical goods and services cheaper. Why is that the case?
   a. Gross national income
   b. Gross domestic product
   c. Consumer preferences
   d. Purchasing power parity

2. How does a country measure domestic aggregate production of goods and services?
   a. Gross domestic product
   b. Gross national incomes
   c. Gross domestic input
   d. Aggregate household income

3. How do we refer to an enterprise that is the only seller of a goods and services?
   a. Free market
   b. Oligopoly
   c. Monopoly
   d. Duopoly

4. Reel ve nominal faiz arasındaki fark nedir?
   a. Hiçbir fark yoktur.
   b. Reel faiz enflasyonu hesaplarken, nominal faiz hesaplamaz.
   c. Nominal faiz enflasyonu hesaplarken, reel faiz hesaplamaz.
   d. Nominal faiz oranı yalnızca ders kitaplarında bulunur.

5. Üretim maliyeti arttığında, tüketim pazarındaki genel fiyatlar da artar. Fiyatlardaki bu artışın iktisadi tanımı nedir?
    a. Deflasyon
    b. Enflasyon
    c. Katma deger vergisi
    d. Devaluasyon

6. Bu yıl, geçen yılki ile aynı maaşın (kıdem vs ayarlanmadan, nominal olarak) alınmasına rağmen yaşam giderleri artmıştır. Neden?
    a. Deflasyon
    b. Enflasyon
    c. Katma deger vergisi
    d. Devaluasyon

Answer the following questions according to biology page in Wikipedia with help of MOHSE.

1. What is the name of organelle, which is responsible for photosynthesis in plants?
    a. Mitochondria
    b. Chromoplast
    c. Chloroplast
    d. Leucoplast

2. What is the process of organism change?
    a. Evoluation
    b. Organism
    c. Growth
    d. Genetics

3. What is the study of plants?
    a. Morphology
    b. Botany
    c. Zoology
    d. Ecology

4. Hangisi bir ekosistem tipi değildir?
    a. Kayalik
    b. Mera
    c. Tundra
    d. Savana

5. Zatürree, gıda zehirlenmesi ve kan zehirlenmesi (septisemi)'ne _______________ sebep olur.
    a. Hayvanlar
    b. Bitkiler
    c. Bakteriler
    d. Virusler

6. Organizmalar ____________ 'ları yakından ilişkili ise aynı grupta sınıflanır.
    a. Renk
    b. Yeme aliskanligi
    c. DNA
    d. Boyut

Middle East Technical University

Computer Engineering Department

Multilingual Dynamic Linking of Web Resources

Task D

Answer the following questions according to economy page in Wikipedia.

1. Bir malin veya hizmetin tek saticinisin oldugu girisimi nasil adlandiririz?
   a. Serbest piyasa
   b. Oligopol
   c. Monopol
   d. Duopol

2. Bir ülke toplam mal ve hizmet üretimini nasıl ölçer?
   a. Gayrısafi yurt içi hasıla
   b. Toplam milli gelir
   c. Gayrısafi yurt içi girdi
   d. Toplam hane geliri

3. Gelişmiş ülkelerden gelen turistler gelişmekte olan ülkeleri ziyarete geldiklerinde aynı mal ve hizmetleri daha ucuza bulurlar. Bu durumun sebebi nedir?
   a. Satınalma gücü paritesi
   b. Toplam milli gelir
   c. Gayrısafi yurt içi hasıla
   d. Tüketici tercihi

4. What is the difference between real and nominal interest?
   a. There is no difference.
   b. Real account for inflation but nominal does not.
   c. Nominal account for inflation but real does not.
   d. Nominal interest rate exists only in textbooks.

5. Assuming that you earn identical (unadjusted, nominal terms) salary this year from the previous year, the cost of living has increased. Why is that case?
    a. Inflation
    b. Deflation
    c. Value added tax (vat)
    d. Devaluation

6. When the cost of production increases, the general price levels in the consume market go up. What is an economic term for increase in price levels?
    a. Inflation
    b. Deflation
    c. Value added tax (vat)
    d. Devaluation

Answer the following questions according to biology page in Wikipedia with help of MOHSE.

1. Bitki bilimi nedir?
    a. Botanik
    b. Zooloji
    c. Ekoloji
    d. Morfoloji

2. Bitkilerde fotosentezden sorumlu organelin adı nedir?
    a. Kloroplast
    b. Kromoplast
    c. Lökoplast
    d. Mitokondri

3. Organizmanın değişim sürecine ne denir?
    a. Evrim
    b. Organizma
    c. Buyume
    d. Genetik

4. Which one of the following is not an ecosystem type?
   a. Tundra
   b. Rocky
   c. Savanna
   d. Grassland

5. Pneumonia, food poisoning and blood poisoning (sepsis) is caused by ....
   a. Animals
   b. Plants
   c. Bacterium
   d. Viruses

6. Organisms are classified under the same group if their..... is closely related.
   a. Color
   b. Size
   c. DNA
   d. Food

APPENDIX E

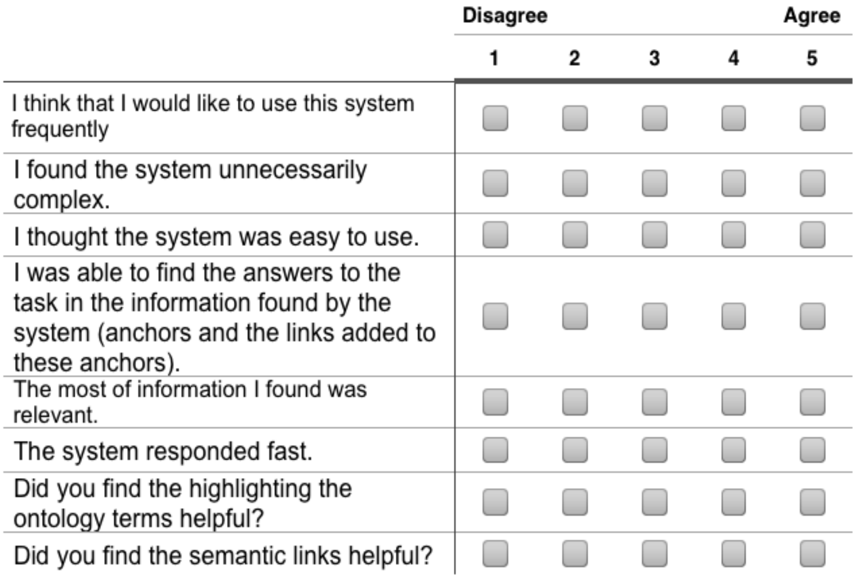
SATISFACTION QUESTIONNAIRE

Middle East Technical University

Computer Engineering Department

Multilingual Dynamic Linking of Web Resources

Post Questionnaire



	Disagree				Agree
	1	2	3	4	5
I think that I would like to use this system frequently	□	□	□	□	□
I found the system unnecessarily complex.	□	□	□	□	□
I thought the system was easy to use.	□	□	□	□	□
I was able to find the answers to the task in the information found by the system (anchors and the links added to these anchors).	□	□	□	□	□
The most of information I found was relevant.	□	□	□	□	□
The system responded fast.	□	□	□	□	□
Did you find the highlighting the ontology terms helpful?	□	□	□	□	□
Did you find the semantic links helpful?	□	□	□	□	□

APPENDIX F

LINK EVALUATION

Middle East Technical University

Computer Engineering Department

Multilingual Dynamic Linking of Web Resources

Link Evaluation

	Economics					
		1	2	3	4	5
1	goods					
2	interest					
3	economy					
4	income					
5	cost					
6	accounting					
7	macroeconomics					
8	price					
9	monopoly					
10	allocation					

	Water					
		1	2	3	4	5
1	ice					
2	limnology					
3	hydrology					
4	sea					
5	river					
6	irrigation					
7	ocean					
8	filtration					
9	well					
10	dam					

	Chemistry					
		1	2	3	4	5
1	chemistry					
2	carbon					
3	hydrogen					
4	thermodynamics					
5	nitrogen					
6	spectroscopy					
7	oxidation					
8	nitrate					
9	mineral					
10	iron					

	Electrolysis					
		1	2	3	4	5
1	electrolysis					
2	chemistry					
3	hydrogen					
4	acid					
5	oxygen					
6	salt					
7	platinum					
8	metal					
9	oxidation					
10	bromine					

Biology						
		1	2	3	4	5
1	biology					
2	evolution					
3	botany					
4	ecology					
5	anatomy					
6	animal					
7	DNA					
8	gene					
9	ecosystem					
10	photosynthesis					

Botany						
		1	2	3	4	5
1	botany					
2	DNA					
3	genetics					
4	morphology					
5	anatomy					
6	biology					
7	ecology					
8	leaf					
9	photosynthesis					
10	behavior					

APPENDIX G

KNOWLEDGE SERVICE OPERATIONS

- **load**

  Loads SKOS or OWL ontologies. Takes context URL and type of ontology as an argument.

  Example usage of service:

  http://localhost:8080/ks/KnowledgeService?service= load&context= http://localhost:8080/rdf/water.rdf&arg=skos

- **unload**

  Unloads the given ontologies as an argument.

  Example usage of service:

  http://localhost:8080/ks/KnowledgeService?service= unload&context= http://localhost:8080/rdf/water.rdf

- **allConcepts**

  Returns the collection of classes in the ontology. Takes context URL as an argument.

  Example usage of service:

  http://localhost:8080/ks/KnowledgeService?service= allConcepts&context= http://localhost:8080/rdf/water.rdf

- **allTerms**

  Returns the collection of terms in the given ontology according to language.

  Example usage of service:

  http://localhost:8080/ks/KnowledgeService?service= allTerms&context= http://localhost:8080/rdf/water.rdf&arg=tr

- **preferredConceptLabelAll**

  Returns the collection of terms and URL of concept in the given ontology according to language.

  Example usage of service:

  http://localhost:8080/ks/KnowledgeService?service= preferredConceptLabelAll&context= http://localhost:8080/rdf/water.rdf&arg=tr

- **preferredConceptLabel**

  Returns the concept and URL of concept in the given ontology and given concept according to language.

  Example usage of service:

  http://localhost:8080/ks/KnowledgeService?service= preferredConceptLabel&context= http://localhost:8080/rdf/water.rdf&arg=http://www.eionet.europa.eu/gemet/concept/419,tr

- **labelledConcepts**

  Returns concepts labeled with the given term in the given ontology.

  Example usage of service:

  http://localhost:8080/ks/KnowledgeService?service= labelledConcepts&context= http://localhost:8080/rdf/water.rdf&arg=arg=water

- **linkedConcepts**

  Returns concepts linked to given concept via given property. Expects an RDF description in the body of the request.

  Example usage of service:

  http://localhost:8080/ks/KnowledgeService?service= linkedConcepts&context= http://localhost:8080/rdf/water.rdf&arg=http://www.eionet.europa.eu/gemet/concept/419, http://www.w3.org/2004/02/skos/core#broader

- **linkedConceptsLang**

  Returns concepts and labels according to given language linked to given concept via given property. Expects an RDF description in the body of the request.

  Example usage of service:

  http://localhost:8080/ks/KnowledgeService?service= linkedConceptsLang&context= http://localhost:8080/rdf/water.rdf&arg=http://www.eionet.europa.eu/gemet/concept/419, http://www.w3.org/2004/02/skos/core#broader,tr

APPENDIX H

CRAWLING AND INDEXING

1. Apache Solr is installed and started.

    (a) Apache Solr is download from web page.

    (b) Extract compressed file and navigate to /solr/example/.

    (c) Start Solr

    ```
    $ java -jar start.jar
    ```

2. Apache Nutch is installed. Download compressed file and extract it.

3. A folder named urls is created under Nutch home. A file named seeds.txt created under urls folder. Web Urls which Nutch will crawl are written seeds.txt file.

4. Agent name is added to nutch-site.xml file.

5. Crawling process is started

    ```
    $ bin/nutch crawl urls -dir crawl -depth 3 -topN 50
    ```

6. Nutch scheme.xml is copied under Solr conf folder.

7. Crawled Web sources are added to Solr.

    ```
    $ bin/nutch solrindex http://127.0.0.1:8983/solr/ crawl/crawldb crawl/linkdb crawl/segments
    ```