TermPedia for Interactive Document Enrichment:

Using Technical Terms (TT) to Provide Relevant Contextual Information

Proscovia Olango

October 30, 2008
Revised March 20, 2009
Abstract

Technical Terms (TTs) and/or jargon embedded within technical documents can make it difficult or impossible to understand a document. This is why we would like to investigate a possibility of providing information for the TTs by linking them to relevant lexicon or encyclopedia pages. In this way, additional contextual information relating to the TTs shall be readily available and hopefully make reading and understanding easier.
# Contents

List of Tables iii

List of Figures iv

1 Introduction 1
   1.1 Document Enrichment 2
   1.2 The Role of Document Enrichment in eLearning 3
   1.3 Definition and Explanation of Terms 4
   1.4 Research Questions and Hypothesis 6
   1.5 Scientific Challenges 6
   1.6 Application 7

2 Related works 7
   2.1 Automatic Term Extraction 8
   2.2 Automatic Term Definition 8
   2.3 Word Sense Disambiguation 9
   2.4 Automatic Link Generation 10
   2.5 Using Knowledge Bases for Document Enrichment 11

3 Method to Accomplish Project Objectives 14
   3.1 Initial Approach 14
      3.1.1 Recognition of Technical Terms 15
      3.1.2 Blindfold Term Recognition (BTR) 15
      3.1.3 Longest String Based Term Recognition (LTR) 16
      3.1.4 Frequency-Based Term Recognition (FTR) 16
      3.1.5 Sense-Based Term Recognition (STR) 16
      3.1.6 Automatic Hypertext Generation 18
   3.2 Expected Challenges 18
   3.3 Assumptions 19

4 Pilot Experiment and Evaluation 19
   4.1 Collecting Medical Anchor Text 20
   4.2 Document Enrichment Using Anchor Text 20
   4.3 Preliminary Results and Evaluation 22
   4.4 Discussion of Results 24

5 Future Work 25
   5.1 Evaluating the systems on External Data 25
   5.2 Integrating the BTR and the FTR Methods 25
   5.3 Improving the Term Recognition Process 26
   5.4 Improving the Systems Coverage 26
   5.5 Application of Machine Learning Techniques 26
   5.6 Application of Statistical Modules 26
   5.7 User Evaluation 26

References 29

Appendix 30
List of Tables

1 Frequency list of the top 10 most frequent words in Wikipedia . 17
2 Recognizing and linking TT to Wikipedia by BTR, LTR, and STR 22
3 Evaluation of Technical Term Recognition . . . . . . . . . . . . . . 23
4 Statistics of TTs predicted by STR against gold standard . . . . 23
5 Evaluation of Link (TT and Target) Prediction . . . . . . . . . . 24
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Screen shot of using Google “define:” to define the TT <em>dura mater</em></td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>Screen shot of using Google “define:” to define the TT <em>cranium</em></td>
<td>13</td>
</tr>
<tr>
<td>3</td>
<td>An example of ambiguous anchor text.</td>
<td>21</td>
</tr>
</tbody>
</table>
1 Introduction

Technical documents are often written for a particular purpose and they are usually geared towards a specific subject [See Markel, 2004]. From a general point of view, technical documentation is intended to link ideas, technologies, processes, and products with people. We can therefore say that the primary objective for technical writing is education since the technical writers relate explanatory notes about various subjects. Take for example technical manuals written for different purposes from fixing minor computer software malfunctions to those explaining difficult operations by doctors in major hospital theaters. Information contained in such manuals may be complicated but, technical writing is beneficial if the written document is clearly understood by the people for whom it is intended. On the other hand, it would also be helpful if a technical document can be easily understood by anyone who is not necessarily part of the interested audience.

Although the primary goal of a technical writer is to write documents that are easy to understand, sometimes this goal is not achieved, especially when the writer addresses experts of a certain subject or knowledge domain. In such cases, technical terms (TTs) or jargon may be introduced without any definition. Such technical documents may also include abbreviations and acronyms that are not explained. If such technical terms, abbreviations, or acronyms are met for the first time by a particular reader who is not familiar with them, they may hinder the reader’s ability to understand the context of a document. To clearly illustrate the problem of technical terms, let us look at two sentence examples which may be difficult to understand for a first time reader who is not familiar with the medical knowledge domain.

Epidural hematoma (EDH) is a rapidly accumulating hematoma between the dura mater and the cranium.

Example 1: Sentence heavily loaded with technical terms

To an ordinary reader this sentence example presents the following technical terms or jargon:

i. epidural hematoma,

ii. hematoma,

iii. dura mater, and

iv. cranium

The example also presents the abbreviation or acronym “EDH” and it is clear that this abbreviation refers to epidural hematoma, but what exactly does this mean? Although the technical term epidural hematoma is explained in the sentence example, the explanation still makes no sense if the terms hematoma, dura mater, and cranium do not “ring a bell” in the reader’s mind.

1Sentence examples were taken from, http://en.wikipedia.org/wiki/Head_injury, in Aug., 2008
We therefore see that sometimes even if the technical terms are explained within the document, additional contextual information would be necessary for the reader to understand it. In the next example we cite a scenario where an abbreviation is written without prior explanation.

**Lens-shaped extracerebral hemorrhage will likely be visible on a CT scan of the head.**

Example 2: Sentence showing unexplained abbreviation

In Example 2 above, the writer assumes that the reader knows the meaning of the abbreviation *CT*. In fact *CT* stands for *Computed Tomography*.

For readers who are not familiar with the medical knowledge domain, the above examples show that unexplained technical terms may hinder the understanding of technical documents. Fortunately we can employ the technique of document enrichment to improve understanding of documents by providing additional contextual information about technical terms.

### 1.1 Document Enrichment

Document enrichment is a relatively new research area in information science that employs the techniques of information extraction and text mining. The difference in objectives is that, while information extraction and text mining concentrate on deducing information and knowledge from existing text, document enrichment emphasizes adding information and knowledge to existing text.

A major technique in document enrichment is that of generating interactive text. One way to generate interactive text is by the creation of hypertext. This involves the process of tagging text with anchors that lead to external or internal information resources of a document [See Domingue et al., 2001]. In order to accomplish the task of document enrichment in this project, we shall use technical terms as anchors. The anchors are important as hypertext links to additional encyclopedia information. The project will assimilate work done by Csomai and Mihalcea [2007] where they try to link educational materials to encyclopedic knowledge.

Encyclopedias are a large source of authentic information but to make good use of this information, a user has to switch between interfaces, whether on the World Wide Web or even while using hard copies of these references. We would like to minimize the time and distance between available information and text by integrating encyclopedic knowledge and other relevant references like educational material into the text. In a pilot experiment, we investigate the possibility of using Wikipedia as a knowledge base for enriching text documents.

Wikipedia is a free-content online encyclopedia which is a product of the continuous collaborative effort of many volunteer contributors. Although a few critics have questioned the credibility and coverage of Wikipedia, in the year 2005 a special report on science articles indicated that Wikipedia is similar to...
Encyclopedia Britannica in both coverage and accuracy [Giles, 2005]. Questions about the authenticity of Wikipedia arise because of the collaborative nature through which the encyclopedia grows, because many of the contributors are not accredited authors. This is apparently seen as an advantage by the co-founders of wikipedia as they anticipate that any error noticed on the content pages shall be boldly corrected by the people who notice them.

One common knowledge domain that affects people of all categories and carries with it a lot of technical terms is the medical domain. It is therefore our interest to try and experiment with the medical knowledge domain after which the successful results shall be applied for general use in all other knowledge domains of both arts and sciences.

Since the main aim of enriching documents in this research is to help readers comprehend text, the research can be looked at in a broader sense as an aid in computer assisted learning or eLearning. It is therefore important to briefly discuss the role of document enrichment in eLearning. The following section therefore gives a brief overview about the role of document enrichment in eLearning.

1.2 The Role of Document Enrichment in eLearning

The role that document enrichment is expected to play in this research is that of bridging the gap between encyclopedic knowledge and text. This means that the research will no doubt develop a user interface so that it’s findings can be implemented in an educational setting. An implementation oriented research is beneficial because it does not only stop at reporting the findings but also endeavors to see to it that the findings are put into use. Therefore it is obligatory that this research develops a user interface that enables the availability of the resources resulting from its findings. A lot of implementation oriented research has been done in the area of computational linguistics (CL) and natural language processing (NLP).

A common example of an implementation oriented research may be given as the one continually being done by Google Incorporation. Google implements a search engine that allows users to search the world wide web through the use of keywords for images, maps, news, videos, and more. Google also implements the Google Language Tools that enables people around the world to translate text and web pages for 41 languages including Arabic, Dutch, Hindi, and Zulu, to mention but a few. Although TermPedia is modeled for only the English language, one major projection is that the result of this research shall be made available to users through an interface that enables them to define technical terms. Users shall also be able to get background information about the terms as TermPedia links the terms to encyclopedias.

Many information scientists are also thinking about how the achievements of CL and NLP research can be implemented. An example of such scientists can be given as Monachesi and Westerhout [2008] who are interested in showing

that the current achievements of NLP and the Semantic Web can play a relevant role in improving the functionality of existing Learning Management Systems (LMS) through the Language Technology for eLearning (LT4eL) project. LT4eL provide content creators with a keyword extractor which allows for semi-automated metadata annotation of learning objects as one of its features. Similar to TermPedia, this feature of LT4eL uses keyword extraction (or information extraction) which as been widely explored in NLP and IR community. The results of LT4eL are adapted to the eLearning context by using statistical measures in combination with linguistic processing to detect salient words which are good key candidates, a technique which is related to the one discussed in section 3.1.5 of this report.

Keogh et al. [2004] also considered the implementation of achievements from CL and NLP by participating in a workshop for eLearning for computational linguistics and computational linguistics for eLearning. The workshop was organized by COLING - The International Conference on Computational Linguistics. Keogh et al. [2004] look at how CL and NLP resources can be deployed in Computer Assisted Language Learning (CALL). Their work concentrates on language learning, but the important thing is that they use computer technology to do this and they are also interested in availing the results of their research to users. They agree that CL and NLP have a lot to offer to eLearning as they focus on asynchronous eLearning for natural languages use in primary schools for learning Irish and German.

In their conclusion, [Keogh et al., 2004] admit that it was difficult to integrate CL and NLP resources into CALL because these resources are not generally designed with CALL in mind. We hope that TermPedia will overcome this difficulty because it is designed with the aim of integrating CL and NLP in an eLearning environment. Encouragingly, [Keogh et al., 2004] report that there are environments where CL and NLP can be successfully integrated and used proactively. We are optimistic that TermPedia will be used fruitfully in an eLearning environment as it provides the results from a CL and NLP research to users through education systems and semantic web technologies.

With the importance of document comprehension in mind, we feel that it is good to define and explain a few terms that have been used so far at this point.

### 1.3 Definition and Explanation of Terms

For clarification, it is important at this point to define and explain some of the vocabulary used in this document.

**TERMPEDEIA**

A good place to start is by explaining what TermPedia means and how this name was conceived. As you may have noticed, the name TermPedia comes from two words *term* and *pedia*. A *term* is a word or an expression that has a precise meaning in some uses or is peculiar to a science, art, profession, or subject.\(^3\)

This definition was retrieved from Merriam-Webster’s Online Dictionary which

is America’s foremost publisher of language-related reference works. The word \textit{pedia} is a plural form of the word \textit{pedion} which is a noun that refers to a crystal form having only a single face, without a symmetrical equivalent.\textsuperscript{4} TermPedia therefore refers to the precise meaning of a word or an expression as used within a given text context without a symmetric equivalence of the meaning of the term in that context. This is because the system is designed to interpret as accurately as possible the meaning of technical terms in the context where they may occur.

\textbf{TECHNICAL TERM (TT)}

A technical term is a newly introduced or uncommon word or combination of words within a particular knowledge domain. Technical term may also refer to common words or combination of words used with a special meaning in the context of a particular technical document. Technical terms carry with them the inherent feature of ambiguity from natural language since the same term may have a different meaning in various knowledge domains [Stolle et al., 2003]. Take for example the technical term \textit{ontology}:

WordNet\textsuperscript{5} defines the word \textit{ontology} as \textit{“a rigorous and exhaustive organization of some knowledge domain that is usually hierarchical and contains all the relevant entities and their relations”}, with reference to the field of computer science or \textit{“the metaphysical study of the nature of being and existence”}, with reference to the field of philosophy.

The word ontology refers to different concepts when considered within computer science or philosophy knowledge domains. This therefore exemplifies an ambiguous technical term that may occur in some technical documents. For humans, it is normally possible to distinguish the interpretation of an ambiguous term intuitively when it occurs in a certain context, but not for machines (computers). It is the challenge of correct interpretation of the meaning of a technical term for machines that we plan to deal with in this project.

\textbf{ACRONYM}

An acronym is a word made up of the initial letters or syllables of other words.\textsuperscript{6} For example PC is a common acronym for a \texttt{PERSONAL COMPUTER} or a \texttt{DESKTOP COMPUTER}.

\textbf{ABBREVIATION}

An abbreviation is a shortened or contracted form of a word or phrase, used to represent the whole word or phrase, as Dr. for Doctor, and U.S. or US for United States.\textsuperscript{7}

For ease of writing, throughout this paper we collectively refer to technical terms, acronyms and abbreviations as technical terms (TTs).

\textsuperscript{4}Retrieved on Feb. 10, 2009, from, \texttt{http://dictionary.reference.com/browse/pedia}
\textsuperscript{5}WordNet is a semantic lexicon for the English language
\textsuperscript{6}See, \texttt{Oxford English Dictionary}, 2005
\textsuperscript{7}Retrieved in Aug., 2008, from, \texttt{http://dictionary.reference.com/abbreviation}
From the definitions we realize that TTs are always words, a combination of words or parts of words (or phrases) that carry special meaning whenever they occur. These special features are almost always responsible for the contextual interpretation of a sentence or a document as a whole, not forgetting that they may also be ambiguous. Therefore, the task of disambiguating the sense of TTs in context will be a useful part of this research.

1.4 Research Questions and Hypothesis

The following are the research questions guiding this research:

i. How can technical terms be identified in text?

ii. How can technical terms be linked accurately to encyclopedic resources?

iii. Does providing definitions of technical terms improve understanding of technical documents?

iv. If information embedded in a knowledge base is seamlessly integrated into technical documents, does this reduce the time required to acquire knowledge and understanding?

The first hypothesis of this research is that providing definitions of technical terms can improve understanding of technical documents. We saw in Example 1 that these definitions may not always be adequate in the goal of providing aid to the understanding of context. This is what brings us to a second hypothesis where we believe that linking technical terms to contextually related information can increase availability of relevant information, thereby improving understanding and making the learning experience easier and more interesting. A more specific hypothesis is that knowledge bases contain an enormous amount of clustered information that can easily be utilized to shorten the time needed to search for relevant information. This should also reduce the time required for acquiring knowledge.

1.5 Scientific Challenges

A major scientific challenge that we would like to investigate is whether the provision of term definitions suffice to facilitate comprehension of technical documents. In the cases where term definitions are not sufficient for document comprehension, we would like to find out if providing relevant encyclopedic information facilitates document comprehension. We are also interested in finding out what else might be needed in order to facilitate the comprehension of documents.

It shall also be a challenge to develop a system that is able to identify people at different reading levels, so that relevant and at the same time adequate information is provided for the different reading levels. This challenge is in relation to how much a reader is acquainted with a certain knowledge domain.

There may also be a problem of coverage of technical terms by existing knowledge bases. This is to say, encyclopedias or glossaries may not include a technical term that exists within a certain document. The challenge would then be to
evaluate the accuracy of using the existing knowledge bases as a reference to
term lists for term extraction from technical documents. Hoste et al. [2007]
discuss a machine learning approach that extracts technical terms from patient
information without the aid of external term lists. It would be interesting to
compare the results of this project with those found in that of Hoste et al. [2007]
which reports an 80% F-score for the machine learning based approach.

1.6 Application

As indicated in section 1.2 of this report, result of this project could be applied
in e-learning as most countries invest a significant part of their education budget
on information and communication technology (ICT). Unfortunately this may
not be very feasible for a developing country like Uganda considering expenses
related to ICT facilities and the knowledge for using and maintaining them [Far-
rell and Shafika, 2007].

Daily Monitor, a local English newspaper in Uganda, reported on the 16th of
July, 2008 that “The level of use of the Internet in Uganda has been constrained
by the high connectivity costs”. The report also says that, according to Google,
as of September 2007, 750,000 Ugandans (about 2.6 per cent of the total pop-
ulation) were accessing the Internet. This shows that it is a meager portion
of the population in a third world country that would benefit from e-learning
requiring Internet connectivity, presumably also benefiting from our project.
However, we are optimistic that this percentage will rapidly grow considering
the large amount of investments different governments put in ICT resources.

The semantic web could also use the automatic hypertext generation feature
of our document enrichment process since this results in anchored text that is
linked to web pages. Also, the term extraction feature of our project could be
used in knowledge processing and information extraction applications for im-
proved annotations.

In practice, the resultant system from this project shall be integrated with
web browsers in order to assist interested persons to automatically access the
definition and explanation of technical terms. The system is anticipated to be
beneficial to researchers, students, and technicians among others.

2 Related works

In the methodology section we shall see that document enrichment is an inte-
gration of techniques such as term extraction, automatic term definition, word
sense disambiguation, and automatic hypertext generation. In this section, we
shall look at some of the work that has been done in the different component
techniques and explain how we can apply them in our project from similar or
different perspectives.

8Mark Kirumira, “Education ministry in meager ICT plans”, Daily Monitor, 16th July,
2008
2.1 Automatic Term Extraction

[Fahmi et al., 2007] focus on the use of existing terms from glossaries, thesaurus, or ontologies to extract new terms from a domain specific text. Their baseline system combines a linguistic pattern for extracting candidate noun phrases with a statistical method for ranking candidate phrases according to their association strength in a domain specific corpus. They developed a method for ranking candidate terms, extracted from Dutch medical corpora, with the help of the Unified Medical Language Systems (UMLS) as an external knowledge source. They concentrated on extraction of phrasal terms and their method combines frequency of occurrence of candidate terms in a corpus with information on how the candidate terms are computed from existing multilingual terms.

Noting that it is only phrasal terms that were considered, it could be a good idea to provide their definitions since the terms are in fact extracted from existing knowledge bases. It would also be interesting to extend the term extraction for other domains of knowledge by using other external knowledge sources in addition to UMLS. Later in our project, we would like to investigate the combination of linguistic features and statistical methods for identifying domain specific terms.

The method that [Fahmi et al., 2007] used could be referred to as a combination of linguistic and term frequency methods. Other methods that are used in term extraction include tf-idf weight (term frequency-inverse document frequency), term co-occurrence, and concept identification using Wikipedia. The latest method considers Wikipedia article page titles as terms and these are in turn used to recognize terms in plain text documents. Jones [2007] made an evaluation of these methods and found out that the Wikipedia technique was seen as significantly more effective than the other techniques. For this reason, we are motivated to use Wikipedia article titles and links embedded within the articles as a baseline for our technical term extraction technique.

2.2 Automatic Term Definition

[Torralbo et al., 2005] present a similar piece of work from the perspective of document summarization rather than document enrichment as proposed by this research. [Torralbo et al., 2005] also exclusively use knowledge bases to find summaries for already existing technical terms (domain-specific terms). The purpose of their proposed system was as follows: given a list of domain-specific terms, to generate a definition of each of the terms. It can be thought of as a complement of term identification procedures, so they are able to provide a small definition of each of the new terms identified. In this approach, they use the World Wide Web as the source of documents, so the module can be applied to every domain for which there is information on the Web.

Torralbo et al.’s work is similar to our intended research but with a different approach. Their approach crucially depends on an existing list of technical terms and here, we would like to also devise a means of identifying the technical terms without a predefined term list. We like to do this because it is probable that new terms are developed which may not be included in the list of available terms.
This would also be a realistic approach of acquiring relevant definitions for each term as used in a particular document. The scope of the World Wide Web is too huge and it is likely that Torralbo and Alfonsaca notice this as well when they write, “... this source is not fully reliable, as it may contain inexact or erroneous information...” This is why we would like to work with authentic documents and try to find both the technical term and their definitions from within the same document in a way to improve the accuracy of automatic cross-references.

2.3 Word Sense Disambiguation

Word sense disambiguation is the process of accurately and automatically identifying the sense in which a word is used in context. Many techniques have been used in this process including the use of machine readable dictionaries. In this approach word sense is guessed by counting overlaps between dictionary definitions of various word senses and the context where a word appears in text [See Lesk, 1986]. Michael Lesk who is known for introducing the Lesk algorithm which is a classical algorithm for word sense disambiguation in 1986, reports that his system disambiguates word sense with an accuracy rage between 50% to 70%, on text from Pride and Prejudice and selected papers of the Associated Press. Notice that only single words are disambiguated in Lesk’s paper and yet the majority of technical terms are compound nouns or word phrases. Nakagawa and Moris [2000], mentions that “85% of domain specific terms are said to be compound nouns”. Therefore as we apply the technique of word sense disambiguation for disambiguating the meaning of technical terms, we take into consideration that most of these terms are a result of compound nouns or a combination of word phrases.

We shall investigate the possibility of extending the Lesk algorithm for technical term disambiguation so that we may be able to automatically provide their accurate definitions in context. Dictionaries may not be regularly updated whereas new terms spring up very often in technical documents. For this reason, it is relevant to use an up-to-date source like Wikipedia during the process of disambiguating technical terms.

“Google and WordNet Based Word Sense Disambiguation” was presented by Klapaftis and Manandhar as an unsupervised method for automatic disambiguation of noun terms found in domain-specific unrestricted corpora. The method used Google to find contextually relevant terms, which in turn helped in assigning the correct WordNet sense to each term under disambiguation. Klapaftis and Manandhar start with a list of terms extracted from a collection of domain specific documents. Each sentence containing the domain specific term is then sent to Google and the first four resultant documents are used in combination with WordNet synset to disambiguate each term. The system evaluation revealed a disambiguation accuracy of 58.90%. It is important to note that this system disambiguates both single nouns and compound nouns but limiting the disambiguation process to the list of words found in WordNet is quite restrictive. In some cases the term to be disambiguated may not be listed in WordNet and we hope to overcome this problem by using Wikipedia which we expect has a larger coverage of terms compared to WordNet.
A recent discussion presenting various word sense disambiguation techniques that use Wikipedia as a resource can be found in section 3.2 of the paper written by Medelyan et al., under the title “Mining Meaning From Wikipedia”. They cover techniques for disambiguating phrases and named entities and show how disambiguation is used to map manually created knowledge structures to Wikipedia. In particular, we wish to explore the method of [Milne and Witten, 2008] that extends the approach of [Medelyan and Milne, 2008] where they use machine learning techniques. Other than using semantic similarity, we would like to collect features relating to the TTs like its frequency, preceding and following words and its position in text for calculating a conditional probability of mapping its resolved sense to Wikipedia articles.

2.4 Automatic Link Generation

Hyperwords is a project that extensively enriches the World Wide Web with information from on-line knowledge bases.

“Hyperwords are interactive words; words you can issue commands on. (This concept is very broad and particularly designed to help knowledge workers move more effectively through their information environment, gather useful information and have access to better tools)”.

It is definitely necessary to have relevant and sufficient information but, making every word interactive would avail too much information and most likely obstruct the reader’s ability to follow the documents by drawing attention away from important concepts. Our project however has a similar goal to that of Hyperwords that is, to shorten the distance between curiosity and knowledge, and between intent and result, but we aim only to provide relevant additional information for important concepts in relation to context.

It is also motivating to consider the work done by [Rotard et al., 2007] under the title, “Semantic lenses: Seamless augmentation of web pages with context information from implicit queries”, who say:

“In analytical processes huge amounts of complex data have to be collected, filtered, and provided to the users in an appropriate way”.

Notably Rotard et al. [2007], emphasize the representation of appropriate data, but we would like to concentrate on implicit information availability by providing explanations and definitions of technical terms. The relevance of the Rotard et al.’s work is that they indicate that on demand, available contextual information can be seamlessly integrated and visualized. It is also in the interest of this project to seamlessly integrate contextual information with available knowledge bases.

Another similar work in relation to automatic link generation is that done by [Goffinet and Noirhomme-Fraiture, 1995]. This work deals with the problem

---

of automatically generating cross-reference links when converting text to hypertext. They use a statistical approach, based on techniques commonly used in Information Retrieval. They also use complementary probabilistic methods in order to create more relevant links than the ones generated without any knowledge. In a similar effort, we shall explore the methods used by Goffinet & Noirhomme-Fratrie and tailor them for generating automatic links from technical terms to on-line knowledge bases. These links will help in the provision of relevant contextual information.

2.5 Using Knowledge Bases for Document Enrichment

The recent years have seen an enthusiastic growth of research in the area of using existing knowledge bases to enrich documents. This is a logical development that arises from the need to comprehend technical documents. Knowledge bases already have information that can improve the comprehension of technical documents if such information is accurately linked to the technical documents. We shall now examine two related works that use existing knowledge base to enrich documents.

A real motivation for our project is the work done by Mihalcea and Csomai [2007] who introduce the use of Wikipedia as a resource for automatic keyword (technical term) extraction and word sense disambiguation. The two methods were combined into a system called Wikify! which automatically performs the annotation task following Wikipedia guidelines. If a document is given as input to Wikify!, the system has the ability to identify the important terms in the text and link them to corresponding Wikipedia pages. Eventually, these links provide users with quick access to additional information. We have generalized this process as document enrichment.

Mihalcea and Csomai do not take into consideration the importance of providing definitions for the terms identified. Nevertheless, these terms are linked to contextually related Wikipedia pages through a word sense disambiguation process. We find that it is important to provide definitions for the terms because in some cases these definitions are sufficient to quench a reader’s thirst for comprehension. Wikipedia may not contain all existing terms, which is why it may be an improvement to use other knowledge bases like WordNet in such cases or the Google “define” function. An impressive Turing test evaluation showed that Wikify! generates Wikipedia annotations that are hardly distinguishable from the ones that are human-generated.

A mini version of Wikify! was tailored for linking educational materials to encyclopedic Knowledge with a hypothesis that the system can facilitate the access to information by students, and consequently it can improve the learning process [Csomai and Mihalcea, 2007]. Results from this experiment indicated that:

“providing students with information relevant to the topic of study, and bringing such information within easy reach through hyperlinking, are both successful strategies for increased effectiveness in pedagogical tasks”.

11
We also share the above vision and hope to construct a system that eases comprehension of documents specifically those of a technical nature.

It was enjoyable to read the related work of [Elhadad, 2006] under the title, “Comprehending Technical Texts: Predicting and Defining Unfamiliar Terms”. In this work Noémie Elhadda investigates a way of improving access to medical literature for health consumers by using an unsupervised method that identifies unfamiliar medical terms. The terms are identified depending on how frequently they occur in the text and their definitions mined from the World Wide Web. She says that the provision of both the term and its web definition improve the comprehension of sentences that contain these technical terms. For defining terms, she uses Google “define:” functionality, this is advantageous because the work of mining definitions from multiple glossaries and web pages is already done.

The figures below show snap shots of using the Google “define:” function to define the terms dura mater and cranium. These figures are used to highlight the work done by the Google “define:” function in mining definitions from multiple knowledge bases that are available on the World Wide Web. To illustrate the benefit of using the Google “define:”, please see figures 1 and 2.

![Google define: dura mater](image.png)

**Figure 1:** Screen shot of using Google “define:” to define the term *dura mater*.

From figure 1 we notice that the Google “define:” function was able to mine definitions of the term *dura mater* from WordNet and an encyclopedia maintained by the University of Rochester Medical Center among other knowledge bases and online educational materials. The Google “define:” function was also
able to mine the definitions of the technical term *cranium* from Wikipedia, and WordNet including other encyclopedias and educational materials on the World Wide Web, as shown in figure 2 below.

![Google search for define: cranium](image)

**Figure 2:** Screen shot of using Google “define:” to define the TT *cranium*

In Noémie’s research, technical terms are defined independently of the context in which they occur because she believes that the context does not influence the fact that a term is technical or unfamiliar to a certain reader but his reading level does. In the event that the assumption about the reader’s level is true, would it be possible to automatically predict a reader’s level? A drawback in this method is that for each sentence, only the first mention of a complex term is defined. Therefore other complex terms that may occur within the same sentence do not get defined and these may still hinder the comprehension of technical documents. We believe that defining all the complex terms within a sentence in consideration of reader level would be an improvement to Noémie’s method.

Generally, most of the related works concentrate on using a list of available technical terms in order to identify technical terms in various documents, we shall also follow this trend.
3 Method to Accomplish Project Objectives

This section provides an overall description of our approach including materials, and procedures that shall be used in the research project. More specifically we shall talk about the data collection and how it was analyzed.

Currently, we are using the entire English uva XML Wikipedia dump of November 2006 which contains more than 5,000,000 articles with over 3,000,000 non-redirect articles as our main corpus. uva is an abbreviation for University of Amsterdam whereas XML is an abbreviation for EXtensible Markup Language. The XML corpus was created by the Information and Language Processing Systems (ILPS) department, Informatics Institute, University of Amsterdam. This XML version of Wikipedia was developed to serve as a multi-lingual text collection for experiments in Information Retrieval and Natural Language Processing, in particular, in the context of Cross-Language Evaluation Forum (CLEF). We use the corpus for extraction of technical terms which is in line with the information retrieval intentions for its creation, but only for the English language. We also use the English XML Wikipedia dump of January 2008 which was compiled by Wikimedia Foundation Incorporation. From this dump, only 4,347 articles that belong to the medical category are considered.

These two sets of data come with all the XML Wikipedia formatting information which we remove to generate a clean text version of the Wikipedia articles. This process was done to create a medical corpus from the November 2006 Wikipedia dump and the cleaning script was written by Geert Kloosterman.

3.1 Initial Approach

In order to accomplish our goal of document enrichment, we need to identify technical terms and provide definitions for these. Keeping in mind that some of the terms are ambiguous, we need to perform sense disambiguation in order to reference a technical term to its contextually relevant definition. During the process of linking technical terms to their definitions, hypertext will be generated. This is an important step for integrating technical documents with available knowledge bases. On the other hand, we are not only interested in the definition and explanation of technical terms because we are doubtful that this will totally satisfy the reader’s quest for content comprehension. Therefore, we also link the terms to appropriate Wikipedia articles depending on their contextual disambiguation.

The final system is an application of integrated techniques including technical term extraction, automatic term definition, technical term sense disambiguation, and automatic hypertext generation, which shall ultimately result into a system of document enrichment. We hope that eventually the enriched documents shall play an important role in computer assisted learning or eLearning in general. Let us take a look at some of the techniques that we use in the integrated application in order to build a document enrichment system.

3.1.1 Recognition of Technical Terms

The technical term recognition process explores a supervised method that uses string look-up. String look-up is the process where we try to find pieces of text in a document that are the same in characters and length to the technical terms that we have in an existing term list. In order to facilitate the string look-up process two databases were created, one containing technical terms and their definitions and the other containing technical terms and Wikipedia articles to which the terms are linked. Each anchor text from the XML dump of Wikipedia is considered as a technical term regardless of how many n-grams it contains. An anchor text is a string of characters that occur between the “<a>” tag of a hypertext markup language (html). XQuery, a language designed to query a collection of XML documents was used to extract the anchor texts together with the Wikipedia articles to which they are linked. XQuery uses XPath to traverse the XML version of Wikipedia in order to retrieve this information. XPath is a language for finding information in an XML document.\(^\text{12}\)

In order to provide definitions of technical terms we carry out a target look-up task for the Wikipedia anchor text. Taking a close look at Wikipedia pages we noticed that the first paragraph is often a definition of the title of that page. Conventionally a Wikipedia anchor text is linked to a Wikipedia page (the target), therefore the definition of the anchor text is the first paragraph of the target page. For this task we do not take into consideration the disambiguation problem but trust that each anchor text is linked to a contextually correct Wikipedia page by the page authors.

Once we have the list of technical terms extracted from Wikipedia, we then use them to recognize TTS in the medical corpus that we generated from Wikipedia. The term recognition process is done through four different methods, which we later evaluate to see which one performs better. The four methods used for term recognition are blindfold term recognition (BTR), longest string based term recognition (LTR), sense-based term recognition (STR), and frequency-based term recognition (FTR). BTR and LTR methods do not take into consideration the fact that the sense of a term may be ambiguous in relation to context. We expected the STR to have the best recall since it was designed to link terms to Wikipedia after considering their meaning in context.

3.1.2 Blindfold Term Recognition (BTR)

The BTR method uses a string look-up technique that marks all possible strings within the text that match an anchor text from the technical terms list. The queue head concept borrowed from the a stack data structure is used for each sentence within the plain text. In other words, we believe that a sentence contains a stack of technical terms and the first term to be recognized in that sentence is then linked to a Wikipedia page. In general, there is also a first come first serve idea for the terms and we expect this method to have the highest precision because it links practically every string that it is able to identify as a term within a sentence to a Wikipedia article. Remember also that a term could be linked to more than one Wikipedia page, but BTR only assigns one

link per term. This method is expensive because for each line of running text, a possible match is looked up for all the technical terms in the created list of terms.

3.1.3 Longest String Based Term Recognition (LTR)

The only difference between BTR and LTR is that LTR first provides Wikipedia links to anchor text (technical terms) that consist of the longest match in a line of running text. The term list is thus sorted in such a way that the ones having greater length are at the queue head considering that the list of terms is a stack. This criterion is adopted in order to avoid splitting up terms that are made of compound nouns or noun phrases. For example, if the term *Epidural hematoma* is not extracted first, it may be split into *Epidural* and *hematoma*. The first come first serve idea here is biased towards the longest string in a sentence that can be recognized as a technical term. Similar to BTR, LTR does not take into consideration that the contextual sense of a term may be ambiguous. Thus the term senses are not disambiguated in this method either.

3.1.4 Frequency-Based Term Recognition (FTR)

For the FTR approach we developed a criterion based on the keyword ranking method called *Keyphraseness*, which was presented by Mihalcea and Csomai [2007]. In this approach all possible n-grams in a document that are present in the list of terms are identified and ranked according to their likelihood of being selected as a TT. If a term is most of the time selected as a TT among its total number of occurrence, it is most likely that it will again be selected in a new document as a TT. Therefore the probability \( P \) that a term \( X \) is selected as a TT in a new document is calculated as the total number of documents where the term was already selected as a TT \( \text{count}(D_{TT}) \) divided by the total number of documents where the term appeared \( \text{count}(D_X) \).

\[
P(\text{TT}|X) \approx \frac{\text{count}(D_{TT})}{\text{count}(D_X)}
\]

The counts are generated from all the articles in the Wikipedia 2006 dump. Given a list of TTs in a document, we select the top 10% although Mihalcea and Csomai [2007] note that on average, 6% of the TTs in a Wikipedia article are actually linked to another page. Note that in a way, the FTR method also does sense disambiguation by picking the most frequent sense for a term since it is linked to the most frequent Wikipedia target page.

3.1.5 Sense-Based Term Recognition (STR)

In the FTR method, we do not take deliberate consideration about the contextual meaning of a term before it is linked to a Wikipedia article. On the other hand, STR takes into consideration the contextual meaning of a term before it is selected as a TT by overlapping the words in the paragraph where the term occurs with the words in the term’s definition. If there is high overlap of words between the two sets of text, then the term is selected as a TT. In cases when a term is ambiguous and has more than one definition, the term is assigned the definition with the highest overlap rank. Terms are only selected as TTs if their
context is similar to the term definition, which process we refer to as TT sense disambiguation.

We use an intersection computation algorithm that works with smoothened percentage scores in order to disambiguate context of the technical terms. Given two paragraphs, we find out which words are common to both paragraphs (intersection). Each paragraph is tokenized and sorted in such a way that they have unduplicated items.

A list of the most common words used in Wikipedia was created to form a stop word list for the this research. The stop word list was created by getting a frequency of all the words that occur in the entire Wikipedia dump using the plain text version. The most frequent 1000 words were then used as a list of stop words. Table 1 shows us that the word “the” occurs “68,730,054” times in the entire uva plain text Wikipedia dump of November 2006 that we created.

<table>
<thead>
<tr>
<th>Common Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>68730054</td>
</tr>
<tr>
<td>of</td>
<td>37253050</td>
</tr>
<tr>
<td>and</td>
<td>28039006</td>
</tr>
<tr>
<td>in</td>
<td>25464434</td>
</tr>
<tr>
<td>a</td>
<td>20977980</td>
</tr>
<tr>
<td>to</td>
<td>20042836</td>
</tr>
<tr>
<td>is</td>
<td>11389541</td>
</tr>
<tr>
<td>was</td>
<td>10265507</td>
</tr>
<tr>
<td>for</td>
<td>08428993</td>
</tr>
<tr>
<td>as</td>
<td>07597870</td>
</tr>
</tbody>
</table>

Table 1: Frequency list of the top 10 most frequent words in Wikipedia

These common words were removed from the tokenized paragraphs to form a pair of word lists \( P_1 \) and \( P_2 \) that contain only the important words within the paragraphs. \( P_1 \) is a list of words derived from the paragraph containing the term definition and \( P_2 \) is the word list derived from the paragraph where the term occurs in a new text document. To find out a percentage overlap rank between the two paragraphs \( \text{POR}_{P_1 \text{and } P_2} \), a count of the words in their intersection \( \text{count}(P_1 \cap P_2) \) is divided by the number of words in \( P_1 \), \( \text{count}(P_1) \) and multiplied by 100.

\[
\text{POR}_{P_1 \text{and } P_2} \approx \frac{\text{count}(P_1 \cap P_2)}{\text{count}(P_1)} \times 100
\]

All ambiguous terms are ranked using this weighted score and the paragraph \( P_1 \) that has the highest rank is given as the term definition and the term marked as a TT.
3.1.6 Automatic Hypertext Generation

After technical terms have been extracted, disambiguated, and defined then each term is used as an anchor text to the appropriate Wikipedia article. An HTML `<a>` (anchor) tag is inserted around it so as to create a link to the respective Wikipedia article. The Uniform Resource Locator (url) of the wikipedia article serves as the hypertext reference (href) attribute of the `<a>` tag. For example if the TT *Epidural hematoma* is recognized by the system, the term will become a hypertext after the following transformation:

```html
```

Example 3: The Effect of document enrichment on a technical term

The hypertext generation completes the document enrichment process because once users click onto the hypertext, they are presented with a Wikipedia article that defines and explains the TT in question.

For a more user-friendly environment, a Java script hover function may be provided to pop-up a window with only the definition of the technical term (anchor text) that includes a link to more information from Wikipedia. If a user is not satisfied with the definition, he is then at liberty to link to the encyclopedia for more explanation about the technical term in question.

3.2 Expected Challenges

The biggest challenge of this project is that of ambiguous terms which is inherent to natural language. The ambiguity is that many terms have several meanings or senses. That is to say, for each term given in context there is an ambiguity of how it could be interpreted.

The project therefore seeks to develop a competent disambiguation engine that is able to predict the interpretation of a term in context. In effect, terms will always be linked to a correct definition or explanation.

In the case where we plan to use existing term lists, it is possible that the term lists do not have all the technical terms used in a specific document. Another limitation could be that the term lists are not up to date, thus lacking in their coverage of available terminology. This challenge shall be reduced by using a combination of multiple term lists.

Inter-annotation disagreements is another challenge that this research faces. Experiments have shown that given the same piece of text, different humans annotate different technical terms. Although this disagreement is controlled by the peoples’ different reading levels as demonstrated by Elhadad [2006] and Csomai and Mihalcea [2007], it is likely that certain Wikipedia contributors overlook technical terms in particular articles. People at the same reading level are expected to have acquired the same level of knowledge too. Therefore, they are likely to meet the same difficulty while reading a document and have more
agreement on what part of the document should be considered as technical terms. However, if people at higher reading levels overlook certain technical terms, additional information shall not be provided for the overlooked terms and probably these terms may still hinder the understanding of a document by someone at a lower reading level.

3.3 Assumptions

We assume that there exist a huge number of technical documents in various fields of knowledge that have been published and that these documents exist electronically.

We also assume that these technical documents contain technical terms that may not be easily understood by students or researchers and other users in general.

Another assumption is that for the various knowledge domains, there exist knowledge bases like dictionaries, encyclopedias, glossaries, thesauruses, and other lists of terms that would provide definitions and explanations for the technical terms.

The fourth assumption therefore is that there is a possibility of linking the knowledge bases to the technical documents in an effort to provide definitions and explanations for each of the technical terms.

4 Pilot Experiment and Evaluation

In order to have a manageable scope, we used article from the medical category of Wikipedia because this data had already been collected by John Kizito. From John’s data we collected ids for medical pages. By using the medical page ids, we were able to generate a medical corpus from the uva XML Wikipedia dump of March 2006. For more information about the uva XML Wikipedia dump please see section 3 of this report.

The medical corpus was created by extracting Wikipedia pages with the same ids as those in John’s medical corpus from the 2006 dump. A total of 1,166 articles constitute the medical sub-corpus that we worked with for training and evaluating TermPedia.

The main aim of the experiment is to link terms existing within a plain text document to a contextually relevant Wikipedia page in reference to the terms’ context. This experiment was inspired by the work done by [Mihalcea and Csomai, 2007], who link documents to Wikipedia with the help of anchor text. For more information about anchor text, please see section 3.1.1 of this report. Therefore as a “kick-off” point we carried out a supervised experiment to predict which terms in a Wikipedia lemma get tagged with links to other Wikipedia pages.

\[^{13}\text{John Kizito (2008). PhD Student, Computational Linguistics, Department of Information Science, University of Groningen.}\]
4.1 Collecting Medical Anchor Text

From the Wikipedia dump of March 2006 that consisted of 1,166 articles, we collected a total of 26,440 page-titles and 30,528 anchor texts. Each anchor text was treated as a medical technical term. We use the anchor text to refer to term lemmas because they reflect the exact form in which terms are written within plain text. For each anchor text we collected information about the Wikipedia page-id, page-name, and page-title to which that anchor refers.

The anchor text collection consisted of single words, groups of words, abbreviations, and/or acronyms, among other forms. The page-titles collected alongside these anchor texts tell us which Wikipedia page the anchor text is linked to. Below, you can see a sample output showing the collection of medical anchor texts from Wikipedia. The output contains various records of four fields, each field within a record was separated by a tab.

<table>
<thead>
<tr>
<th>Page-Id</th>
<th>Page-Name</th>
<th>Page-Title (Target)</th>
<th>Anchor Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>353792</td>
<td>Folk_medicine</td>
<td>Herb</td>
<td>herbs</td>
</tr>
<tr>
<td>310484</td>
<td>Patent_medicine</td>
<td>Herb</td>
<td>herbal</td>
</tr>
<tr>
<td>2142761</td>
<td>Leukapheresis</td>
<td>Blood plasma</td>
<td>Plasma</td>
</tr>
</tbody>
</table>

Sample 1: Output of collected medical anchor texts and targets

The first field is the page-id, the second field the page-name, the third field the page-title of the link, and the forth field the anchor text of the link. So the page with id 2142761 on “Leukapheresis” has a link to the Wikipedia page “Blood plasma” with anchor text “Plasma”.

4.2 Document Enrichment Using Anchor Text

By now, we have a list of medical anchor texts (or term lemmas) and page-titles of the Wikipedia pages to which the anchor texts are to be linked. This makes it possible for us to enrich plain text documents with information from Wikipedia by automatically generating hypertext using the existing list of anchor texts. Some anchor text refer to more than just one Wikipedia page-title. For example, the anchor text Avicenna refers to both Avicenna and Avicenna (crater) page-titles as shown in Figure 3.

Finding Anchor Text in Plain Text

We retrieved all possible Wikipedia page targets for each anchor text and each page-title was separated from the other using the pipe "|" character. We called these multiple page targets, alternative links. The following sentence is presented to demonstrate the process of enriching plain text by creating hypertext that links to Wikipedia pages. Our challenge during the disambiguation process for example, was to make sure that the anchor text Avicenna is linked to the Wikipedia page-title Avicenna and not Avicenna (crater). Certainly the anchor text Avicenna in this context refers to a person and not a crater as shown by Example 4 below. The stages of enriching the sentence are shown in examples 5 to 7 so that we can see the transformation process the sentence goes through as it is enriched with information from Wikipedia.
Figure 3: An example of ambiguous anchor text.

Apparently for the sentence below, there were three technical terms recognized by the methods BTR, LTR and STR. The methods also linked the recognized technical terms to the same Wikipedia articles (targets). The recognized technical terms were *Avicenna*, *medical*, and *herbs*.

Avicenna also introduced medical herbs.

Example 4: Sentence containing ambiguous anchor texts

Example 5 below shows the final state of linking the ambiguous TTs to Wikipedia articles. The ambiguous TTs in the enriched text document were only linked to one Wikipedia page as decided by the methods.

```html
```

Example 5: Adding HTML `<a>` tag to plain text by BTR, LTR, and STR

Table 2 below shows the technical terms that were recognized by these methods and the Wikipedia anchors to which the anchors were linked. A web browser view of the link enriched sentence is shown in example 6 below.

[Avicenna] also introduced [medical] [herbs].

Example 6: Web browser view of BTR, LTR and STR sentence enrichment
Recognized TT | Recognized Target (Name of Wikipedia article)
i. Avicenna | Avicenna
ii. medical | Herbalism
iii. herbs | Medical care

Table 2: Recognizing and linking TT to Wikipedia by BTR, LTR, and STR

On the other hand, if FTR is used to enrich the sentence, only one anchor text is recognized. Similarly for the technical term recognized by FTR, it is linked to the same Wikipedia articles as that provided by BTR, LTR, and STR methods. FTR recognizes Avicenna as a technical term and links it to the Wikipedia article with the name Avicenna as shown in example 7 below.

Example 7: Adding HTML <a> tag to plain text by FTR

4.3 Preliminary Results and Evaluation

We evaluated the performance of the four methods used for recognizing TTs from text and linking them to Wikipedia articles. The original Wikipedia pages came in eight zipped files with the file names wikipedia-en0.txt.gz to wikipedia-en7.txt.gz. A random number of 151 Medical articles were selected from the wikipedia-en6.txt.gz and wikipedia-en7.txt.gz zipped files for the evaluation purpose. Considering that each system automatically adds hypertext links to Wikipedia-lemmas, the idea for the methods evaluation was therefore to extract all automatically added links by these methods and compare these links to the links in the original Wikipedia page.

Evaluation of term recognition and automatic link generation for all the three methods was done by calculating precision, recall and F-score for each method. F-score in particular measures the system’s accuracy and reaches its best value at 1 and worst at 0. The overall F-scores which are presented in Tables 3 and 5 below are the F-scores of the average precision and recall for each system. In the case of term recognition, precision is the number of correctly recognized terms divided by the number of all terms that were recognized by the methods and recall is the number of correctly recognized terms divided by the number of terms that exist in the original Wikipedia articles.

Similarly, for automatic link generation, precision is the number of links correctly generated by the methods divided by the number of all links that were generated by the methods. Recall is the number of links that were correctly generated by the methods divided by the number of links that exist in the original Wikipedia articles.

LTR method has the best recall at recognizing terms but since it is using a list of known anchor text its performance should be as close to 100% recall as possible.
<table>
<thead>
<tr>
<th>Method</th>
<th>#docs</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>BTR</td>
<td>151</td>
<td>0.186</td>
<td>0.679</td>
<td>0.292</td>
</tr>
<tr>
<td>LTR</td>
<td>151</td>
<td>0.237</td>
<td>0.864</td>
<td>0.372</td>
</tr>
<tr>
<td>FTR</td>
<td>151</td>
<td>0.332</td>
<td>0.742</td>
<td>0.458</td>
</tr>
<tr>
<td>STR</td>
<td>151</td>
<td>0.214</td>
<td>0.797</td>
<td>0.338</td>
</tr>
</tbody>
</table>

Table 3: Evaluation of Technical Term Recognition

Thus, although a 86.40% recall is acceptable it is not good enough for the resource and method in use. We suspect that the longest match look-up criteria partly counts for the 20% lost terms because a term can only be recognized once and if it is seen in the longest string then it will not be seen again. Interestingly, all the three methods have poor precision. Although FTR has the best precision of 33.20% it reflects a much lower standard result than that presented by [Mihalcea and Csomai, 2007] in their method of keyphraseness which had a precision of 53.37%. Unfortunately we cannot compare these results literally because Mihalcea and Csomai [2007] use the entire Wikipedia and we only consider articles from the medical category.

<table>
<thead>
<tr>
<th>Article Id</th>
<th>Gold</th>
<th>Overlap</th>
<th>Missed</th>
<th>New TTs</th>
<th>All TTs</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>6021903</td>
<td>6</td>
<td>4</td>
<td>2</td>
<td>30</td>
<td>34</td>
<td>0.118</td>
<td>0.667</td>
<td>0.200</td>
</tr>
<tr>
<td>6027168</td>
<td>5</td>
<td>5</td>
<td>0</td>
<td>6</td>
<td>11</td>
<td>0.455</td>
<td>1.000</td>
<td>0.625</td>
</tr>
<tr>
<td>6029299</td>
<td>7</td>
<td>7</td>
<td>0</td>
<td>33</td>
<td>40</td>
<td>0.175</td>
<td>1.000</td>
<td>0.298</td>
</tr>
<tr>
<td>6037441</td>
<td>11</td>
<td>11</td>
<td>0</td>
<td>73</td>
<td>84</td>
<td>0.131</td>
<td>1.000</td>
<td>0.232</td>
</tr>
<tr>
<td>6044476</td>
<td>17</td>
<td>16</td>
<td>1</td>
<td>136</td>
<td>152</td>
<td>0.105</td>
<td>0.941</td>
<td>0.189</td>
</tr>
<tr>
<td>6045326</td>
<td>22</td>
<td>19</td>
<td>3</td>
<td>28</td>
<td>47</td>
<td>0.404</td>
<td>0.864</td>
<td>0.551</td>
</tr>
<tr>
<td>6086338</td>
<td>20</td>
<td>16</td>
<td>4</td>
<td>55</td>
<td>71</td>
<td>0.225</td>
<td>0.800</td>
<td>0.352</td>
</tr>
<tr>
<td>6090423</td>
<td>32</td>
<td>26</td>
<td>6</td>
<td>155</td>
<td>181</td>
<td>0.155</td>
<td>0.875</td>
<td>0.263</td>
</tr>
<tr>
<td>6099668</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>42</td>
<td>46</td>
<td>0.087</td>
<td>1.000</td>
<td>0.160</td>
</tr>
<tr>
<td>6104845</td>
<td>31</td>
<td>28</td>
<td>3</td>
<td>59</td>
<td>87</td>
<td>0.345</td>
<td>0.968</td>
<td>0.508</td>
</tr>
<tr>
<td>Totals</td>
<td>135</td>
<td>136</td>
<td>19</td>
<td>617</td>
<td>754</td>
<td>2.260</td>
<td>9.115</td>
<td>3.378</td>
</tr>
<tr>
<td>Averages</td>
<td>13.5</td>
<td>13.6</td>
<td>1.9</td>
<td>61.7</td>
<td>75.4</td>
<td>0.220</td>
<td>0.912</td>
<td>0.338</td>
</tr>
</tbody>
</table>

Table 4: Statistics of TTs predicted by STR against gold standard

Key for table 4

- Article Id: Identification number of the Wikipedia article.
- Gold: Total number of TTs that exist in an original Wikipedia article.
- Overlap: Total number of TTs that exist in an original Wikipedia article that were predicted by STR.
- Missed: Total number of TTs that exist in an original Wikipedia article that were not predicted by STR.
- New: Total number of TTs that were predicted by STR but did not exist in an original Wikipedia article.
- All: Total number of TTs that were predicted by STR in the text version of a Wikipedia article.
Taking a close look at a few documents, it was noticed that TermPedia methods predicted many more TTs for the articles than the ones that were indicated by the contributing authors. This is thus the main reason for the general poor precision results by the methods regarding TTs prediction. In table 4 we can see that for 10 randomly selected articles, a total of 617 new terms were predicted by the STR method as compared to a total 135 terms that originally existed in these Wikipedia articles. The total number of newly predicted TTs is well over 50% consequently producing low precision. The statistics in this table were produced as a result of running STR on the plain text version of the 10 random articles. For evaluation only links that occurred in the text section of an article were considered. From the table we can clearly see that the total number of terms that exist in an original Wikipedia article are far less compared to the ones predicted by TermPedia Method STR. This comparison is also similar for all the other methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>#docs</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>BTR</td>
<td>151</td>
<td>0.153</td>
<td>0.586</td>
<td>0.243</td>
</tr>
<tr>
<td>LTR</td>
<td>151</td>
<td>0.204</td>
<td>0.771</td>
<td>0.322</td>
</tr>
<tr>
<td>FTR</td>
<td>151</td>
<td>0.309</td>
<td>0.689</td>
<td>0.427</td>
</tr>
<tr>
<td>STR</td>
<td>151</td>
<td>0.196</td>
<td>0.755</td>
<td>0.311</td>
</tr>
</tbody>
</table>

Table 5: Evaluation of Link (TT and Target) Prediction

Table 5 gives results for anchor text recognition and target prediction, this is a harder task so we expected lower scores, which the figures above confirm. A 77.1% recall in the case of LTR is low for a system that uses a supervised method, because what this method does is simply to assign links to terms that have been linked before. The method dose not take into consideration the frequency of the term or its contextual sense. Perhas this may be a reason for its low recall. The best system for both term recognition and automatic link generation is revealed as the FTR system that excels with F-scores of 45.8% and 42.7% for the two tasks respectively.

Choosing the most frequent sense (FTR) outperforms the (STR) method because the overlap ranks are very low for the term context and term definition paragraphs. We set a cut off as low as 5% for the overlap rank and still came out with these results. Sample outputs show that the words in a definition do not always overlap with the words in a context paragraph considering that we take the definition as the first paragraph of a target page. It is also possible that a whole paragraph is too long for generating an accurate word overlap. May be it is better to take a few words to the left and right of where the terms occur and overlap these instead.

4.4 Discussion of Results

The biggest bottle-neck seems to be adequate anchor term prediction. If we think of this as a machine learning (ML) problem then for each string in the text that has been used as an anchor, we could collect features like its frequency, its position, and words that preceded and follow it in an article. These features
can then be used in a ML method to improve the general performance of the system in term prediction thereby also improving the automatic link generation process of the system.

For the FTR method, only terms consisting of up to 7-grams were considered, so inorder to improve the precision, terms consisting of up to arbitrary n-grams could be considered. This method also predicts all anchor text that have a less than 5% frequency as a TT, and we take this as a strong point of the method because if an anchor text is minimally used then we presume that it is definitely a term within the running article. Evidently, this is a trick that improves the precision of the system, improving its F-score as well.

Another restriction that may be affecting the performance of the FTR system is that we set a cut-off of term anchors at 4% of the total number of words in the running article. If this maximum ratio is set to 3% or lower, perhaps it will gives precision and recall closer values, thus an optimum value of F-score shall be realised. Our primary goal is to assist readers maximally by providing background knowledge and lowering the term to total number of words ratio in a running article would be one way to do so.

We also predict that if the two methods BTR and the FTR are integrated, it would result in a system that has better recall than both methods have separately. But this is something we have to do as a future task and evaluate the results to ascertain this prediction.

5 Future Work

This section discusses both short and long term plans that should be carried out during the process of developing a competent document enrichment system. Future evaluation procedures shall also be briefly mentioned.

5.1 Evaluating the systems on External Data

At the moment we use plain text that was generated from the medical category of Wikipedia articles. We would like to explore the performance of the systems on external medical data. For this task, an appropriate corpus is still to be identified. We shall also involve humans in our evaluation procedure. That is to say, pick a few documents, have them annotated by humans and compare the annotation results with that of our system. It is also in our interest to compare the performance of our system and that of [Mihalcea and Csomai, 2007] on external data. Therefore both systems shall be run on external data other than that generated from Wikipedia and their performance evaluated.

5.2 Integrating the BTR and the FTR Methods

The two selection criteria for a TT shall be integrated into one system, so that we shall have a system that takes into consideration the longest n-grams of terms in combination with the most frequent terms. We hope that this system
will be better at predicting terms and should have a better precision than the two systems have separately at the moment.

5.3 Improving the Term Recognition Process
Besides using the list of technical terms extracted from Wikipedia, we would like to use open source applications for information extraction that can identify terms and compare the result with our systems. This is important because it will help us to find out if the Wikipedia anchor text is an adequate coverage of TTs. The open source for term extraction is still to be identified.

WordNet is another resourceful knowledge base which we plan to use in combination with Wikipedia to improve the term recognition and disambiguation features of our system. Also with the Google “define:” function, we hope to be able to connect text documents to information resources outside Wikipedia.

5.4 Improving the Systems Coverage
A possibility of linking TTs through Google search results to free online educational material will also be explored in order to facilitate research and further reading in the cases when term definitions do not give adequate information for complete comprehension of a technical term and the document at large.

For TTs which are depicted in a pictorial way, we plan to incorporate the available images in Wikipedia along with their definitions. This should help in understanding, when a reader sees an image then he knows precisely what the technical term is referring to in the real world scenario.

5.5 Application of Machine Learning Techniques
In order to be able to recognize a term, we shall collect features like its frequency, its position, and words that precede and follow it in an article. These features will then be used in a machine learning (ML) technique to improve the general performance of our system in term prediction thereby also improving the automatic link generation process of the system.

5.6 Application of Statistical Modules
If time allows, we would also like to explore frequency counts for identifying likely technical terms without the use of external resources like term lists and encyclopedias.

5.7 User Evaluation
The main problem of this research is that text normally contains technical terms that may hinder the comprehension of a document even when the terms have been defined. The hypothesis of running a user evaluation for TermPedia is thus to show that it can provide relevant contextual information for technical terms. This in turn will help a reader to easily comprehend a document. With the same TermPedia, the time needed to acquire additional information is expected
to reduce. The learning experience is expected to become more interesting because unknown concepts shall be just “a click away!”

Since TermPedia is trained on medical data, we plan to carry out the user evaluation with first year undergraduate students at a medical school. This group of students are selected because they obviously read medical documents and also because they are at a lower reading level compared to medical students in their 2nd or 3rd year study. A low reading level here means that the students do not have a lot of knowledge in a particular knowledge domain. We suspect that students in their first year of study are at a low reading level and therefore more likely to encounter difficult technical terms. A students reading level is expected to improve as they advance through the years in medical school.

To make the user evaluation interesting, beneficial, and feasible a short piece of electronic text taken from a reading assignment shall be given to 50 students who are in their first year of medical school. The text shall contain a maximum of 400 words and the students shall be given ten (10) minutes to read it. During the period of text reading, the students shall be divided into two groups, A and B. Group A students will use TermPedia environment while reading the text and B students will be free to use any electronic aid while reading the text with the exception of TermPedia environment. After the reading session, both groups will then be required to answer a set of questions. The question will be geared to evaluate how well the students were able to comprehend the given text with and without the help of TermPedia. Examples of the evaluation questions may be given as follows:

i. Did you find any technical terms as you read the given text?
ii. If so, write down at least five of them.
iii. If you met technical terms during your reading, were they difficult for you to understand?
iv. If you met a technical term that was difficult to understand, were you able to find its meaning? If so how? If not why?
v. On average, how long did it take you to find the meaning of a term?
vi. If you were able to find the meaning of a difficult term, did it help you to understand the term?
vii. If the term meaning did not help improve your understanding, please give one reason why.
viii. For those using TermPedia, please write down its advantages and disadvantages.
ix. Did you get adequate help on the difficult terms by using TermPedia?
x. If not, what could be done to improve TermPedia?

Naturally, before the evaluation, the students shall be brief about the task that they are expected to perform so that the results from the evaluation will be useful. We hope that students who used TermPedia shall have understood the text better and faster than those who did not.
References


## Appendix A

### Time Schedule

<table>
<thead>
<tr>
<th>Year</th>
<th>Month</th>
<th>Activity</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>October - December</td>
<td>Evaluating the systems on External Data</td>
<td>3 months</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Identify medical corpus outside Wikipedia</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Run both Wikify! and our system on identified corpus</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Integration of BTR and FTR Methods</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Evaluate results</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Publication</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>January</td>
<td>Involve humans in TT recognition and evaluate</td>
<td>1 month</td>
</tr>
<tr>
<td></td>
<td>February - May</td>
<td>Improving the Term Recognition Process</td>
<td>4 months</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Employing WordNet for term disambiguation</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Integrating Google “define:” for term disambiguation</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Evaluation of system performance</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Publication</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>June - September</td>
<td>User Evaluation</td>
<td>4 months</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Designing of a user interface</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Selecting users for system evaluation</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>- User system evaluation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Publication</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>October - December</td>
<td>Expanding System Coverage</td>
<td>3 months</td>
</tr>
<tr>
<td></td>
<td>(September)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Integrating Google search results to system</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Integrating Wikipedia images to TTs</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Evaluation of system coverage expansion</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Second year progress report</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Publication</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td>January - March</td>
<td>Application of Machine Learning</td>
<td>3 months</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Collecting features for machine learning</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Developing a ML method of term recognition</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Developing a ML method for term disambiguation</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Evaluation of ML application</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Publication</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>April - May</td>
<td>Application of Statistical Modules</td>
<td>2 months</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Application of Naive Bayes assumption to system</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Evaluation of statistic module applications</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Publication</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>June - July</td>
<td>Third year progress report</td>
<td>2 months</td>
</tr>
<tr>
<td></td>
<td>August - December</td>
<td></td>
<td>5 months</td>
</tr>
<tr>
<td></td>
<td>Writing Thesis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>January - May</td>
<td>Writting thesis in progress</td>
<td>5 months</td>
</tr>
<tr>
<td></td>
<td>June</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>September</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>