A Study of Man / Machine Annotation
for
Word Sense Disambiguation

Project Report

Submitted in partial fulfillment of the requirements for the degree of

Bachelor of Technology

by

DIPTESH KANOJIA

Roll No: 0944910004

Under the guidance of

Prof. (Dr.) Pushpak Bhattacharyya

Department of Computer Science and Engineering
Indian Institute of Technology, Bombay

March 2012

Acknowledgement

I would like to express my profound gratitude to **Prof. (Dr.) Pushpak Bhattacharyya** for his *invaluable guidance* and support for rigorous study of literature. His essential and unparallel inputs and suggestions during difficult phases were much required. I am highly indebted to him for his kind help. I thank my senior research fellows, **Salil Joshi** and **Arindam Chatterjee** for the minutely detailed discussions we had on Word Sense Disambiguation, Eye Tracking and Human Annotation Techniques. I thank **Ms. Sudha B.** for getting in so much related work at such a crucial point of time. I thank **Mitesh Khapra** and other fellow researchers for their support and suggestions.

I would like to thank **Prof. G.G. Ray (HOD), Industrial Design Centre, IIT Bombay** for his kind support by providing the much needed Eye Tracking device for our experiments and making our research possible.

I would also like to thank **Hindi WordNet members** Mrs. Laxmi Kashyap, Mrs. Jaya Jha, Mrs. Rajita Shukla, Mr. Prabhakar Pandey and Akhlesh Meena for their immensely motivating attitude towards, and skilled annotation for this experiment. This report could not have been complete without all of them.

I would also like to thank **Prof. Y.P. Singh (Dean)** and **Mr. Jai Veer, (HOD)**, Computer Science and Engineering Department, K.S. Jain Institute of Engineering and Technology, for their constant moral support, help, and guidance.

Finally, I would like to thank my dear friend **Nitin Kumar Singh** for supporting me throughout all my tough times, and encouraging me to work harder every time I transgressed.

Diptesh Kanojia
(0944910004)

Abstract

Word Sense Disambiguation (WSD) is defined as the task of computationally finding the senses of words from a context. Our long time research on Word Sense Disambiguation shows that in almost all disambiguation algorithms, the sense distribution parameter *P(S|W)*, where P is the probability of the sense of a word W being S, plays the deciding role. The widely reported accuracy figure of around 60% for all-words-domain independent WSD is contributed to mainly by *P(S|W)*, as one ablation test after another reveals.

The story with human annotation is different though. Our experience of working with human annotators who mark with WordNet sense ids, general and domain specific corpora brings to light the interesting fact that producing sense ids without looking at the context is a heavy cognitive load. Sense annotators do form hypothesis in their minds about the possible sense of a word ('most frequent sense' bias), but then look at the context for clues to accept or reject the hypothesis. Such clues are minimal, just one or two words, but are critical nonetheless. Without these clues the annotator is left in an indecisive state as to whether or not to put down the first sense coming to his mind. The task becomes all the more cognitively challenging, if the senses are fine grained and seem equally probable.

The current work aims at understanding, at a deeper level, the meaning and significance of "contextual evidence", with respect to the human sense disambiguation task. The experiments were designed to probe into the disambiguation time across different POS categories and ontological categories. This would give us the level of difficulties associated with sense disambiguation across the same.

Another set of experiments were conducted to investigate into the set of clues in the neighboring words of a pre-defined target word, when it is annotated with a particular sense. The target words were selected based on frequency in the corpus and degree of polysemy, to ensure multiple instances of the word and to avoid monosemous cases, in which case the experiment becomes futile. Once substantial clue word sets are obtained, hypotheses can be constructed to design context-sensitive grammar or, rules to disambiguate words based on senses. Hence a rule based WSD engine can be developed. The experiments were conducted on a Remote Eye tracking Device.

Contents

Chapter 1

Introduction

Major Natural Language Processing applications rely on Word Sense Disambiguation, which depends on sense tagged corpus. Sense tagging involves assignment of Part of Speech tag and Wordnet Sense ID to the words in the document provided. Sense marker tool helps lexicographers in sense tagging or better known as human annotation. It assumes one sense per discourse for faster tagging, and also assumes stems of unique words in corpora are available in absence of stemmer.

1.1 Word Sense Disambiguation (WSD)

Word Sense Disambiguation (WSD) is one of the toughest problems today, not only in Natural Language Processing (NLP) but also in Artificial Intelligence (AI). The problem of WSD dates back to 1950s. It basically refers to the automatic disambiguation of word senses and it has been an interest and concern since the earliest days of treating languages computationally.

According to experts, sense disambiguation is an “intermediate task”, which is just a stepping stone to most Natural Language Processing tasks. Before moving into further detail, let us find out what exactly WSD is, and the various ways of looking at the problem, along with the various forms of the problem and *why the problem is favorite among NLP researchers*.

1.1.1 WSD Definitions

Initially it is important to know the basic definitions of the WSD problem and its different types. The following section deals with the evolution of WSD as a problem. In light of the motivation behind WSD, let us have a look at the colloquial and formal definitions of WSD.

1.1.1.1 General Definition

WSD is the ability to identify the meaning of words in context in a computational manner. Given a set of words (*e.g.*,, a sentence or a bag of words), a technique is applied which makes use of one or more sources of knowledge to associate the most appropriate senses with words in context. In light of the motivation behind WSD, let us have a look at the colloquial and formal definitions of WSD.

1.1.1.2 Formal definition

WSD is the task of assigning the appropriate sense(s) to all or some of the words in T, that is, to identify a mapping A from words to senses, such that $A(i) \in Senses_D(w_i)$, where $Senses_D(w_i)$ is the set of senses encoded in a dictionary D for word w_i, and A(i) is that subset of the senses of w_i which are appropriate in the context T. The mapping A can assign more than one sense to each word $w_i \in T$, although typically only the most appropriate sense is selected, that is, $|A(i)| = 1$. Where, T is a sequence of words (w_1, $w2$, . . . ,w_n).

1.1.1.3 As a Classification Problem

Word senses are the classes, and an automatic classification method is used to assign each occurrence of a word to one or more classes based on the evidence from the context and from external knowledge sources. Other classification tasks are part-of-speech tagging, named entity resolution, text categorization, etc. But important difference between these tasks and WSD is that the former use a single predefined set of classes (parts of speech, categories, etc.), whereas in the latter the set of classes typically changes depending on the word to be classified.

1.1.2 Variants of WSD

The problem of Word Sense Disambiguation can be broadly classified into two categories.

1.1.2.1 Lexical Sample (Targeted WSD)

Here, the system is required to disambiguate a restricted set of target words usually occurring one per sentence. It employs supervised techniques using hand-labeled instances as training set and then an unlabeled test set.

1.1.2.2 All-words WSD

Here, the system is expected to disambiguate all open-class words in a text (i.e., nouns, verbs, adjectives, and adverbs) therefore, they are termed as wide coverage systems to disambiguate all open-class words. Typically, such a system suffers from Data sparseness problem, as large knowledge sources are not available.

1.1.3 WSD: Heart of the NLP Problems

As mentioned earlier, due its applicability and hardness, WSD has become an area of keen interest and challenge for NLP researchers. Word Sense Disambiguation is truly the heart of all NLP problems. This very fact is illustrated in the figure below:

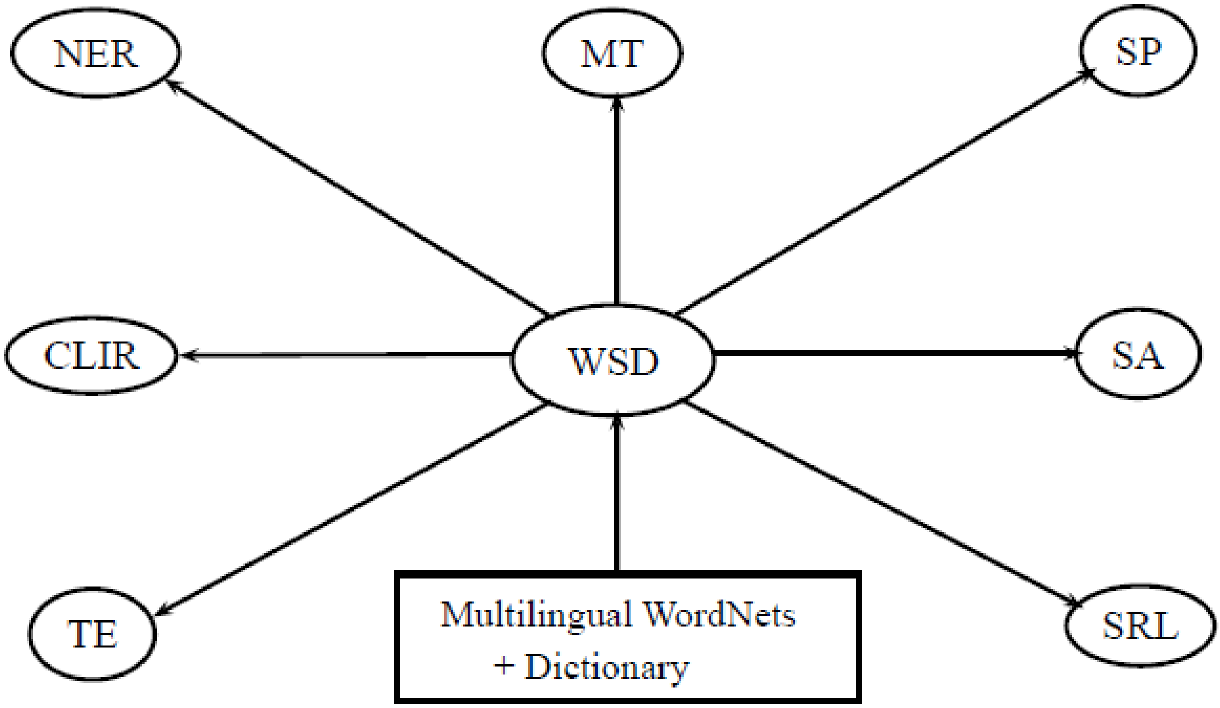


Figure 1.1: WSD - Heart of the NLP problems

- **SRL:** Semantic Role Labeling
- **TE:** Text Entailment
- **CLIR:** Cross Lingual Information Retrieval
- **NER:** Named Entity Recognition
- **MT:** Machine Translation
- **SP:** Shallow Parsing
- **SA:** Sentiment Analysis
- **WSD:** Word Sense Disambiguation

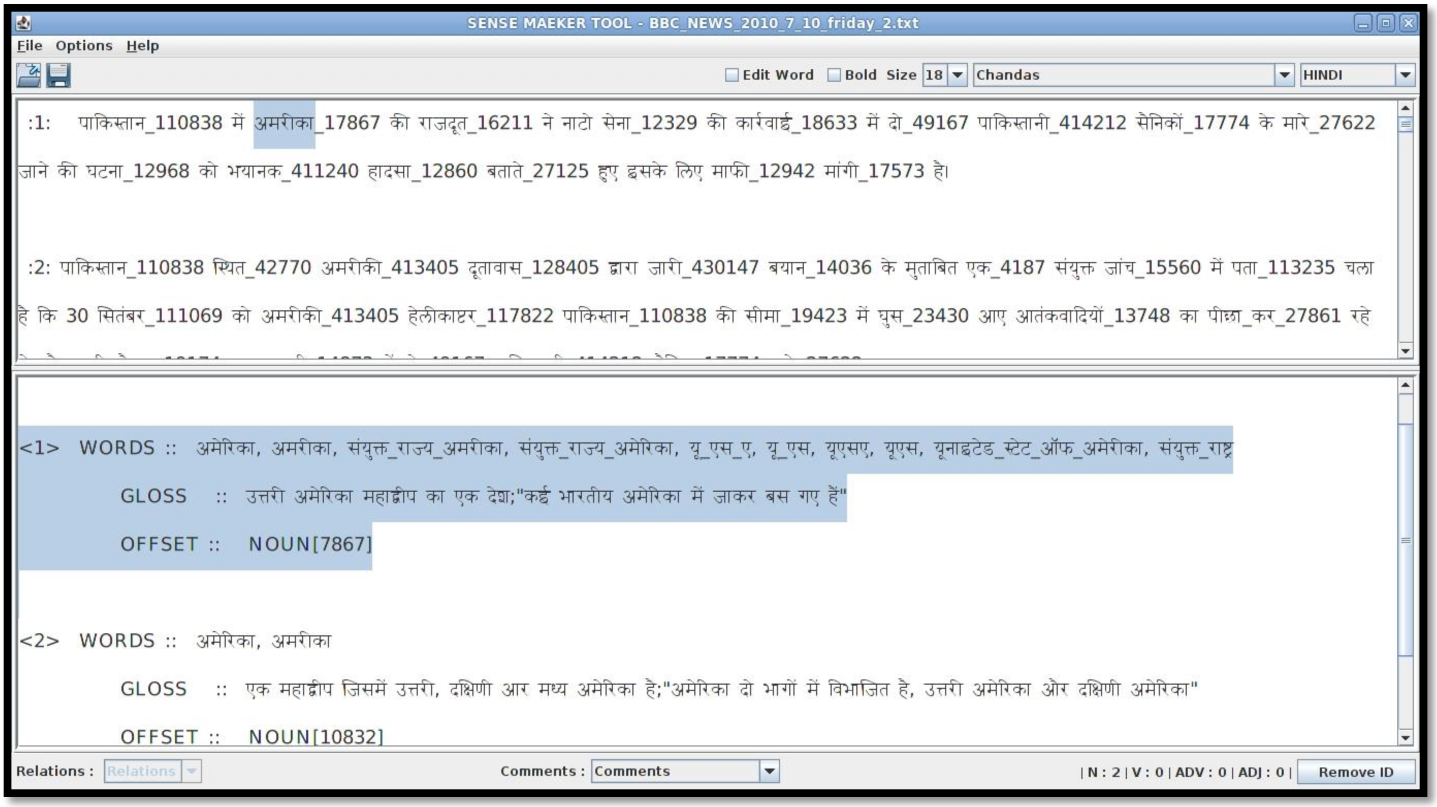
1.2 The Sense Annotation Process

In this section, we have demonstrated the backbone of our human annotation process, the sense marker tool. It serves as one single weapon to which helps our annotators mark the words with sense ids in 18 different languages. The machine needs to be trained by humans to understand the written language. Huge amount of accurately sense-marked data is supplied to the algorithm for its training.

1.2.1 The Sense Marker Tool

A word may have a number of senses and to identify and mark which particular sense has been used in the given context is known as sense marking.

At IIT Bombay, this work is being done in 3 languages – English, Hindi and Marathi. The corpus used so far have been taken from Tourism, Health, Environment and Travel review domains and the Princeton WordNet is used as the sense inventory for English text while the Hindi and Marathi WordNets have been used for Hindi and Marathi texts respectively.
The sense-marker tool developed by IITB supports 18 languages (English, Hindi, Marathi, Assamese, Bengali, Bodo, Gujarati, Kannada, Kashmiri, Konkani, Malayalam, Manipuri, Nepali, Oriya, Punjabi, Sanskrit, Telugu and Urdu).



□ Edit Word □ Bold Size 18 Chandas HINDI

:1: पाकिस्तान_110838 में अमरीका_17867 की राजदूत_16211 ने नाटो सेना_12329 की कार्रवाई_18633 में दो_49167 पाकिस्तानी_414212 सैनिकों_17774 के मारे_27622 जाने की घटना_12968 को भयानक_411240 हादसा_12860 बताते_27125 हुए इसके लिए माफी_12942 मांगी_17573 है।

:2: पाकिस्तान_110838 स्थित_42770 अमरीकी_413405 दूतावास_128405 द्वारा जारी_430147 बयान_14036 के मुताबित एक_4187 संयुक्त जांच_15560 में पता_113235 चला है कि 30 सितंबर_111069 को अमरीकी_413405 हेलीकाप्टर_117822 पाकिस्तान_110838 की सीमा_19423 में घुस_23430 आए आतंकवादियों_13748 का पीछा_कर_27861 रहे

<1> WORDS :: अमेरिका, अमरीका, संयुक्त_राज्य_अमरीका, संयुक्त_राज्य_अमेरिका, यू_एस_ए, यू_एस, यूएसए, यूएस, यूनाइटेड_स्टेट_ऑफ_अमेरीका, संयुक्त_राष्ट्र

GLOSS :: उत्तरी अमेरिका महाद्वीप का एक देश;"कई भारतीय अमेरिका में जाकर बस गए हैं"

OFFSET :: NOUN[7867]

<2> WORDS :: अमेरिका, अमरीका

GLOSS :: एक महाद्वीप जिसमें उत्तरी, दक्षिणी आर मध्य अमेरिका है;"अमेरिका दो भागों में विभाजित है, उत्तरी अमेरिका और दक्षिणी अमेरिका"

OFFSET :: NOUN[10832]

Relations : Relations

Comments : Comments

| N : 2 | V : 0 | ADV : 0 | ADJ : 0 |

Picture 1.2: The Sense Marker tool screenshot.

The Sense Marker tool is a Graphical User Interface based tool developed using Java which facilitates the task of manual sense marking. This tool displays the senses of the word as available in the Marathi, Hindi and Princeton (English) WordNets and allows the user to select the correct sense of the word from the candidate senses.

The table shown alongside is the statistics of sense-marking done for Tourism and Health files (for two languages: Marathi & Hindi):

Domain	Total Documents	Total Sentences	Tagged Words
Tourism Hindi	152	15,200	1,80,525
Tourism Marathi	152	15,200	1,25,387
Health Hindi	89	8,900	94,209
Health Marathi	72	7,200	51,415
Tourism English	152	15,200	1,81,964
Health English	140	14,000	1,49,259
Total	757	75,700	7,82,759

Table 1.1: *Statistics of sense marking*

1.3 Report Outline

This report describes our work done on human sense annotation, in a context sensitive scenario. Chapter 1 provides with a brief introduction of WSD, and Sense Marker Tool. Chapter 2 outlines our motivation for going forward with such an experiment. Chapter 3 gives a detailed explanation of what an eye tracking device is, and how is it used. Chapter 4 is built around the literature survey of eye tracking and WSD. Chapter 5 details how we conducted the experiment and what exact settings did we use. Chapter 6 includes all the results of both the phases of our experiments. While Chapter 7 encompasses of all our discussions on our work here, Chapter 8 consists of conclusion and insight from current work and possibilities of much needed future work.

Chapter 2

Motivation

2.1 Human *vs.* Machine Annotation

The process of sense annotation of words with senses is more accurate for humans than machines. The deciding parameter in the human sense disambiguation process is contextual evidence. Considering the principle of *weak AI*, the annotation procedure employed by the machine should make use of contextual evidence for disambiguation purposes in some form, which also conforms to the classical definition of WSD.

2.1.1 Previous Work

We exhibited that **contextual evidence** is a *necessary attribute* for the human tagging process. Without contextual information, the human tagging process is crippled (Arindam Chatterjee, 2012). Machines, which use the *P(S|W)* statistic for WSD, take human context-sensitive information to learn the *P(S|W)* measure . This is an adaptation of the contextual evidence used by human beings. Hence the principle of *weak AI* holds for such WSD algorithms. Hence obtaining the *P(S|W)* values perfectly is of paramount concern for machines.

Supervised approaches to WSD deliver far better results, compared to knowledge-based or unsupervised methods (Navigli, Word Sense Disambiguation: A Survey, 2009). In a supervised framework, WSD is considered as a classification task, where senses of words are the classes. If we take a closer look at the state-of-the-art supervised algorithms for WSD, it will be evident that the parameters used by such algorithms are mostly statistical, *i.e.*, *corpus-based evidence*.

WSD researchers have tried to incorporate contextual support in the form of syntactical features, co-occurrence statistics and so on, but these algorithms do not perform significantly better over the Most Frequent Sense baseline. WSD researchers have tried to incorporate contextual support in the form of syntactical features, co-occurrence statistics and so on, but these algorithms do not perform significantly better over the Most Frequent Sense baseline.

Consequently, state-of-the-art WSD algorithms use the *P(S|W)* statistic for annotation. In our previous work, we tried to answer two basic questions regarding the annotation techniques of man and machine:

For Humans*: Can humans annotate data efficiently without contextual evidence?*

For Machines*: Do machines need context information during the annotation process?*

We successfully demonstrated:

- Contextual information is paramount for humans while disambiguating sense of a word.
- The annotation process of tagging without the context is cognitively strenuous and time consuming as compared to tagging with help of the context.
- In the case of machines, the *P(S|W)* measure can fetch high accuracies, provided that it has been correctly captured in the corpus by human beings, during annotation process. This in turn necessitates annotations with the help of context.

- In the case of machines, the *P(S|W)* measure can fetch high accuracies, provided that it has been correctly captured in the corpus by human beings, during annotation process. This in turn necessitates annotations with the help of context.

- WSD algorithms, if trained on corpus generated through Context Agnostic annotation process, would result in low accuracies, as the *P(S|W)* parameter is not efficiently captured in this case.
- Once the training process is over and *P(S|W)* statistic is captured, machines do not require further contextual information while annotating, unlike human annotation process. From this perspective, machines do not ape the human annotation technique but, through an adaptation of this technique provide high ac-curacies. Hence, machines conform to the principle of *weak AI* with respect to the an-notation process.

In case of machines, we have observed that the *P(S|W)* statistic is the machine's adaption of human context sensitive annotation process and the principle of weak AI is satisfied here. However, the accuracies for WSD algorithms are not yet at par with human annotation quality. For this, we would like to see if using better contextual parameters in the Iterative WSD scoring function and ranking the senses using a balanced formulation between statistical and contextual parameters.

WSD algorithms are mostly supervised and use the *P(S|W)* statistic for annotation. Besides, the *P(S|W)* statistic is obtained after training on a corpus in the context sensitive setting. Hence there is absorption of contextual information in the generation of the *P(S|W)* values from the context sensitive training data.

2.2 Eye Tracking - Motivation

In Section 2.1 we have seen that human beings use contextual evidence as a prime parameter for sense disambiguation. This motivated us to look in depth at how human annotation is done. In this work we attempt to answer the following questions:

- *How does one find contextual evidence in the textual information?*
- *What are the deciding factors for sense disambiguation?*
- *What are the critical clue words responsible for the sense disambiguation of a particular word?*

Human annotators *form a hypothesis* as soon as they start reading the text, reaching the target word while reading might be sufficient to gain enough evidence to disambiguate it, in some cases even reading the whole text might not give sufficient clues to disambiguate a word. *Machines have no such facility*. **The paragraph** that the annotator is reading **always gives him a vague idea of the word sense**. In fact, **the domain** of the text being annotated gives away the most appropriate sense's idea (Mitesh Khapra, 2009). Also, being familiar with the text beforehand stimulates the idea of a winner sense in the mind. Hence, to assure genuineness of our experiment we separated line from different documents of the corpus and jumbled them up, such that, each sentence of text is taken from a different sort of contextual background.

The cognitive load over a human while annotating is way more than one imagines. As our expert lexicographers narrate, the hypothesis formation and rejection, work hand in hand as the senses are first narrowed down to a few most probable senses and then the winner sense is selected on the basis of matching the word with the gloss provided along with the sense.

One of the more important factors is the replaceability of synonyms provided along with in the sense window, if somehow narrowing down to a few senses and gloss matching tests are not enough, replaceability of the synonyms give the annotator a better understanding of the sense, which also works as verification in many cases.

The above mentioned factors along with the rich knowledge background form firm sense identification basis in one's mind and decides on an appropriate winner sense. Humans have a more powerful very imaginative visual sense of thinking, hence reading text stimulates visual background in a mind and this is again a very helpful factor in disambiguating a word written within a piece of text.

Hence the process of human annotation differs from machine completely.

. Questions like:

- *How much time do humans need to disambiguate a word?*
- *Which POS category takes longer to be disambiguated?*
- *What ontological category is tough to be understood?*
- *What are the context words which form a clue for sense disambiguation?*
- *What are the determining factors for sense selection of a particular word?*

are answered in our work here, along with the results of our experiments, and a strong possibility of further research in this area.

For sure, it will differ for every person, depending on factors like ones linguistic skill, native speaking, background of familiarity with the language. So we started exploring our options on how to study the concept of context deeply when we came across eye tracking device which could capture the eye gaze of a person while he reads.

We decided to go through with the eye tracking of an annotator as he marks the sense ids for given words in a document. This helps us form a set of context words for a particular sense id.

Chapter 3

Eye Tracking or Gaze Tracking?

Eye tracking is the process of measuring either the point of gaze ("where we are looking") or the motion of an eye relative to the head. An **eye tracker** is a device for measuring eye positions and eye movement. Eye trackers are used in research on the visual system, in psychology, in cognitive linguistics and in product design. There are a number of methods for measuring eye movement.

Eye tracking is more accurately called **gaze tracking**, because **it's not the eye that's being tracked as much as the gaze of the eye.**

Eye movement is a combination of **fixations and saccades**.

- *A fixation* is when the eye gaze pauses in a certain spot.
- *A saccade* is when it moves to another position.
- The resulting series of fixations and saccades is called a *scan path*. In eye tracking, the fixation is the most important point of data; researchers want to understand where someone is focusing and *often for how long*.

A Fixation Duration varies from 120-1000 ms, typically 200-600 ms. Typical fixation frequency is < 3 Hz, and they are interspersed with saccades. Saccades on the other hand are very rapid; their duration is typically only 40-120 ms. They are very fast (up to 600 ^{O}/s) and therefore the vision system is suppressed during the movement. They are ballistic in nature *i.e*. the end point of saccade cannot be changed during the movement. Saccades are used to move the fixation point, and if a movement of larger than 30 degree is required, then head moves with the eyes.

Initially developed and thought about by Louis Emile Javal, eye tracking was meant to study the reading of people. He noticed that people do not read smoothly across a page, but rather pause on some words while moving quickly through others. It later advanced to usability studies, marketing, cognitive linguistics and even neurosciences. Jamal published a series of revolutionary papers in 1878-79 to initiate this line of thought among others.

It has since then been quite a startling discovery on how the movement of eye gaze can evaluate your thought process. It is also used for more accurate interpretations of implicit feedback for machine learning. The section below gives a brief history of the eye tracking applications and the research areas since its discovery.

3.1 History

Eye tracking is quite an old method and the first study based on gaze measuring took place in 1878[1]. While photographing techniques developed in the early 20th century and later, when the techniques of the moving image evolved, also gaze study had more alternatives to develop. The earliest eye tracking methods were quite precarious from the aspect of the tested persons. (Nicholas Wade, 2005)

During the 1930's eye tracking was used in reading study. Gaze paths were researched while reading different kinds of fonts. Eye tracking were connected in usability study for the first time in 1950's. Then Fitts *et al.* used movie cameras to study air force pilots' eye movements while landing the plane.

1970's was more or less the golden decade of the gaze study. Techniques developed fast and psychological theories linked gaze study to cognitive processes. In 1980's eye tracking faced a new challenge as a method when computers started to increase. The interests of the scientists focused in human computer interaction and, thus also in gaze based writing.

Nowadays eye tracking is a very useful method for different kinds of purposes. It is a valid method when researching for example reading process or traffic behavior while driving. It is also a very important method when developing user interfaces for disabled peoples. Furthermore eye tracking is a very applicable method when analyzing users' behavior in WWW pages and of course, in usability and accessibility research.

Scientists still have different opinions whether there are a connection between eye movements and cognitive processes (Merja Lehtinen, 2005).

How do users interact with the list of ranked results of WWW search engines? Do they read the abstracts sequentially from top to bottom, or do they skip links? How many of the results do users evaluate before clicking on a link or reformulating the search?

The answers to these questions will be beneficial in at least three ways. First, they provide the basis for improved interfaces. Second, they suggest more targeted metrics for evaluating the retrieval performance in WWW search. And third, they help interpreting implicit feedback like click through and reading times for machine learning of improved retrieval functions.

In particular, better understanding of user behavior will allow us to draw more accurate inferences about how implicit feedback relates to relative relevance judgments. (Laura A. Granka, 2004)

[1] Javal, 1878/1879

3.2 Techniques and Data

3.2.1 Various Techniques

There are several techniques to measure and analyze eye gaze based on the kind of technique you are using. Nowadays the techniques are much more comfortable than a hundred years ago, but they are still in a developmental phase. Eye tracking techniques can be divided in three groups depending on the amount of physical touch the user needs to load.

First of these techniques is *one based on **infra red light** projected from the eye. Second are the techniques based on **electric potential measured from the skin,** around the eye*. Third techniques are *based on the **special contact lenses***. The first of the techniques is probably the most useful way, when searching WWW pages, and reading text. It is also the most comfortable way for the person in the test, because he doesn't have to wear any heavy helmet or special lenses.

***Picture 3.1:** Most comfortable technique to measure gaze based on infra red light projected from the eye.*

***Picture 3.2:** A bit more complicated way to measure gaze using electric potential around the eye.*

***Picture 3.3:** The eye tracking glasses are used for broad range of mobile eye tracking studies.*

***Picture 3.4:** The ergonomic chin rest eye tracking device for high speed and accurate measurements with a large visual field.*

3.2.2 Various Data Formats

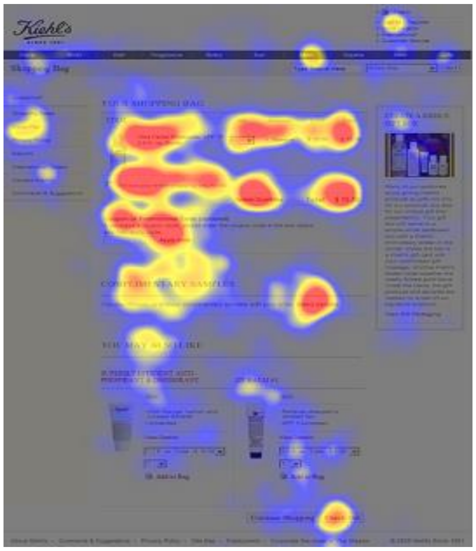
- *Stimulus* is the experiment data input provided to the participant in the experiment, who in our case, were skilled lexicographers.

- Most **common data forms** collected via eye tracking are probably **heat maps and hot spots, gaze paths and areas of interest**.

- From the length and duration of the gaze path it is possible to analyze the effectiveness of the visual search in the certain area.

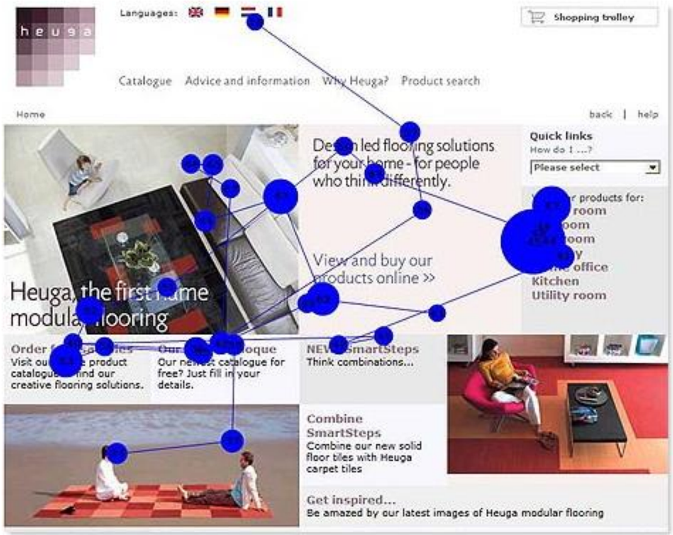
- *A gaze path* focused on the small area means usually *effective search* (Merja Lehtinen, 2005).

- *A heat map* shows the hot spots of the search. The warmer the color the more attention user has paid.



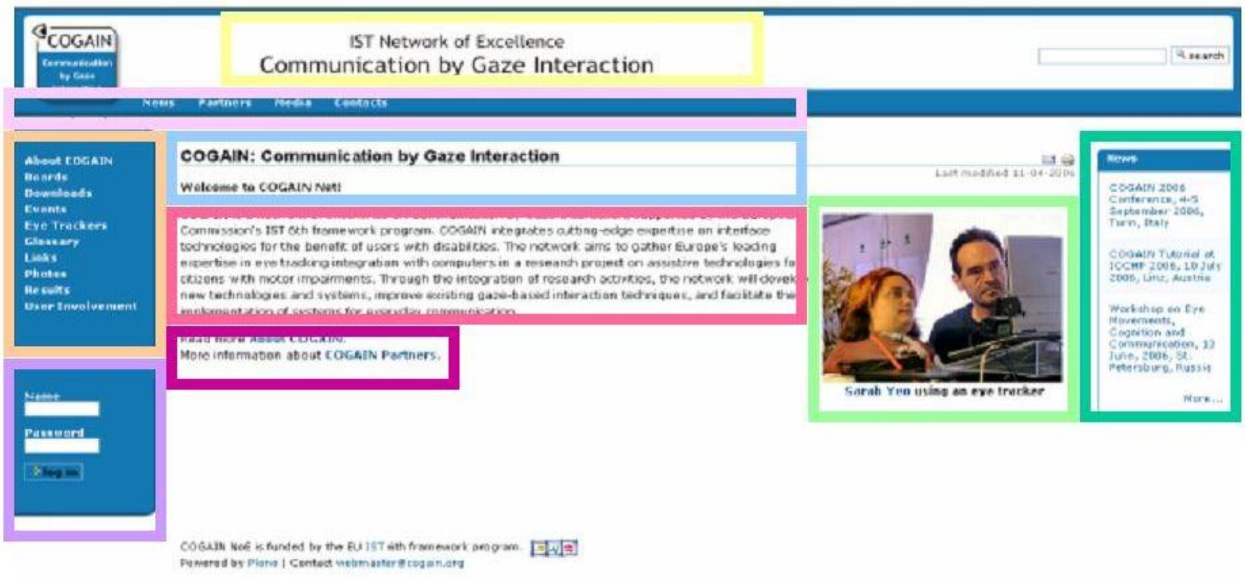
Picture 3.5: A Heat map from usability eye tracking study.

The areas where users looked the most are color red; the yellow areas indicate fewer views, followed by the least-viewed blue areas.


h e u g a
Languages:
Shopping trolley
Catalogue
Advice and information
Why Heuga?
Product search
Home
back | help
Design led flooring solutions for your home - for people who think differently.
Quick links
How do I ...?
Please select
View and buy our products online >>
Heuga, the first home modular flooring
Kitchen
Utility room
Order free samples
Our new catalogue
NEW SmartSteps
Think combinations...
Combine SmartSteps
Combine our new solid floor tiles with Heuga carpet tiles
Get inspired...
Be amazed by our latest images of Heuga modular flooring

Picture 3.6: A Gaze path with fixations and saccades.

The bigger the blue dots, the larger the time of fixation on that point. The Blue lines / paths connecting the blue dots are saccades.



COGAIN: Communication by Gaze Interaction

Welcome to COGAIN Net!

Last modified 11-04-2006

Commission's IST 6th framework program. COGAIN integrates cutting-edge expertise on interface technologies for the benefit of users with disabilities. The network aims to gather Europe's leading expertise in eye tracking integration with computers in a research project on assistive technologies fo citizens with motor impairments. Through the integration of research activities, the network will devel new technologies and systems, improve existing gaze-based interaction techniques, and facilitate the implementation of systems for everyday communication.

Read more About COGAIN.
More information about COGAIN Partners.

COGAIN 2006 Conference, 4-5 September 2006, Turin, Italy

COGAIN Tutorial at ICCHP 2006, 10 July 2006, Linz, Austria

Workshop on Eye Movements, Cognition and Communication, 12 June, 2006, St. Petersburg, Russia

COGAIN NoE is funded by the EU IST 6th framework program.
Powered by Plone | Contact webmaster@cogain.org

Picture 3.7: Areas of interests

- The searched area can be split in to smaller areas. After splitting it is possible to measure attention duration paid for the certain area. The data is called *areas of interests*.

- Eye-tracking is developing as a method all the time.

- In the future it might be a cheaper method, but right now *the techniques are relatively expensive*.

Nowadays eye tracking is a very useful method to collect data. It is easy to collect quantitative data via eye tracking and the data can be also evaluated qualitatively. Eye tracking is a valid method not only in usability and accessibility research, but also in psychological, and psycholinguistic research. Gaze study is used in developing new user interfaces. It is also a useful method when researching for example reading process or traffic behavior while driving.

Other applications of eye tracking studies are namely *Neurosciences*, *Reading*, *Brain Imaging*, *Scene Perception*, *Visual Search*, *and Auditory Language Processing*.

It is not only used for studies and scientific purposes but industrial usages of an eye tracking device are also very broad. Such as, Aviation industry uses it to track a pilot's gaze whilst in different circumstances. Driving hazards can be noted via the use of an eye tracking device, since most of them occur due to lack of visual attention. Thus a lot of other uses are possible and the scope of usage of these devices will only increase.

How we exploited it to our purpose is expressed in the oncoming chapters.

Chapter 4

Related work

Knowledge based approaches to WSD such as Lesk's algorithm (Lesk, 1986), Walker's algorithm (Walker & Amsler, 1986), conceptual (Agirre & Rigau, 1996) and random walk algorithm (Rada, 2005) essentially do Machine Readable Dictionary lookup. However, these are fundamentally *overlap based* algorithms which suffer from overlap sparsity, dictionary definitions being generally small in length. Further, these algorithms completely ignore the domain specific sense distributions of a word as they do not rely on any training data.

In order to obtain higher accuracies in knowledge base approaches the rule base of the algorithm should be robust and large enough to scale well in an all-words scenario. For this purpose, we used the eye tracking device to accumulate contextual evidence as has been done by numerous other NLP researchers. An eye movement experiment was conducted by Seppo Vainio, Jukka Hyönä and Anneli Pajunen to examine effects of local lexical predictability on fixation durations and fixation locations during sentence reading. (Seppo Vainio, 2009) The results showed that first fixation and gaze duration on the target noun were reliably shorter in the high-predictability than in the low predictability condition. As regards eye guidance in reading, their study indicates that local lexical predictability influences when decisions but not where the initial fixation lands in a word.

In another work based on word grouping hypothesis and eye movements during reading by Denis Drieghe, Alexander Pollatsek, Adrian Staub, and Keith Rayner the distribution of landing positions and durations of first fixations in a region containing a noun preceded by either an article or a high-frequency three-letter word were compared (Denis Drieghe, 2008). (Radach, 1996) inferred from a similar experiment that did not manipulate the type of short word that two words could be processed as a perceptual unit during reading when the first word is a short word. As this different pattern of fixations is restricted to article noun pairs, it indicates that word grouping does not occur purely on the basis of word length during reading; moreover, as we demonstrate, one can explain the observed patterns in both conditions more parsimoniously, without adopting a word grouping mechanism in eye movement control during reading.

Chapter 5

Experimental Setup

In our experiments, we have used an eye tracking device manufactured by SensoMotoric Instruments. Better known as RED (Remote Eye Tracking Device 60/120), this device provides us with the frequency of 60 Hz to 120 Hz. The device is presently under use by Ergonomics Lab, IDC, IIT Bombay who were kind enough to provide us access for our experiments.

RED measures gaze hotspots on the stimulus monitor. To accomplish this, it uses an infrared light source to illuminate the eyes, a CCD (Charge Coupled Device) sensor to capture a reflection of the user's eyes, as shown in the picture below, using eye gaze capture software (iViewX and Experiment Center) and eye-gaze analysis software (BeGaze) to process the data. By using a remote, digital eye-gaze tracker, we recorded saccades and fixations on the stimulus monitor, the length of each fixation, the distance to the eye, the pupils' diameter, events such as a click, position of eye gaze and retina on x and y coordinate respectively.

Eye-gaze analysis software produces various graphs that are useful for data interpretation, as follows:

- **Hotspots** - Generalize the behavior of a group of test subjects. They're very similar to heat maps.

- **Gaze plots** - Provide a comprehensive image of all the eye-gaze data from a single or both eye(s) tracking test.

- **Gaze replays** - Provide both *real-time* and *slow-motion* replay of the paths a user's eyes followed during an eye tracking test. This feature also records events such a click on the stimulus monitor, along with the gaze path and fixation scheme. Better known as the Screen Recording feature in Experiment Center Software.

- **Raw Data file** – BeGaze provided us with a raw data file which includes all the data such as the length of each fixation, the distance to the eye, the pupils' diameter, events such as a click, position of eye gaze and retina on x and y coordinate respectively.

We used the above mentioned Raw data file to classify the events separately and calculate the time difference between the different events.

5.1 Hurdles / Failures

Automated analysis of a screen recording which consisted of the gaze video to capture the words was not looking possible to us. We could not figure out a way to automatically capture Hindi text from the video.

Our SenseMarker Tool not being recognize some of the click and not selecting the words properly also became a huge problem, now since no automatic analysis was possible, we mapped the click events to no. of words in a file, words being white space separations.

We provided the annotators with a basic guideline to click twice for a word, so that we could map the time difference alternatively i.e. every alternative time difference is the time taken to tag a particular word.

So, now ignoring the punctuations and the function words or conjunct words we could map all the time difference manually into a spreadsheet. It had to be done manually taking into consideration the following:

- The annotators may make a mistake of clicking more than once for a word, such events had to looked up in the video and deleted from the filtered data file.
- The annotator may by mistake leave or ignore a word, for such words, blank events had to be inserted at that place.
- Many of the multi words could not be found and hence could not be tagged in the files.
- Since, clicks were made on such words, mapping had to be completed and hence we had to manually leave blank spaces for any such instances as well.

Now, all such mistakes could only be corrected when we would look at the video while simultaneously deleting the events from the spreadsheet. Hence, it took us quite a bit of time to analyze the experiments done by our annotators.

We also tried to map the x and y coordinate to the words on the SenseMarker Tool using modifications in our JAVA code, which again took a lot of time, but could not be possible due to the following limitations:

- Extra click on the words by our annotators were not being handled, there was no way to separate one event of extra click from multiple extra click events.
- If the annotator forgot to click a word twice and goes back to complete our mapping, we would not be able to predict that using the code, until and unless the video is analyzed manually.
- Hence, we finally set out to manually analyze the video and working on the spreadsheet.

5.2 Experiment Phases and Method

5.2.1 Phases

We had set out with an aim to find the approximate time being taken to tag a particular word, Further which can be categorized, POS and Ontology wise, and to find the context words.

Hence, we divided the work to be done in two major parts:

5.2.1.1 All word Annotation

We took help of *6 annotators* to tag the same set of files, so that our data set can be statistically verified. We divided the tagged corpus files into *11 separate files of 4 to 18 paragraphs each based on difficulty levels to be increased one by one*. We jumbled the corpus paragraphs so that the annotators don't have a good idea of the context beforehand, and domain remains as generic as possible. Making sure that the context of the next sentence or the paragraph is not easily given away, we then cleaned the pre-tagged files and started with the experiment. *Approximately 2000 words were tagged* by each of our annotators.

Since we knew that the earlier reading of related text gave away the context easily and that might have diluted our results, we made sure we jumbled up the paragraphs in each file and then they were provided with the files. Each and every word was to be clicked in the file twice, for the alternative time mapping.

5.2.2.2 Targeted Disambiguation

We took help of 4 annotators for this experiment keeping into mind that inter-annotator disagreement occurs for some word senses. Since, in this case a particular word was to be disambiguated. We searched for most frequently occurring words in the Hindi Wordnet and out of a list of top 100 we chose 17(15+2) most polysemous words for disambiguation. Searching for such word occurrences in the corpus we took 25 instances of each word in a separate file. The words were marked with '#' on both sides on every instance and the annotator was told to look for the words immediately and then for the context of helpful clues in the sentence provided.

So, we could monitor the gaze of the annotator in BeGaze and look out for longer fixations on some clues words / repeated look on clue words / set of clue words. Hence another spreadsheet was prepared with the particular word marked with its sense id and containing the set of crucial words against it.

5.2.2 Method

A minimum of 2 people are always required on a single experiment, a researcher or a computer scientist to setup the experiment and monitor the gaze and device, other one being the participant in the experiment, the lexicographer or the annotator to be precise. No other distraction should be allowed in the room / near the participant, which might cause the participant to move his eye gaze or even distract him in any such possible way.

Two display screens comprise of a workstation, one where the experiment is being performed by the participant, and the other where the researcher sits and watches his gaze being tracked. The output then is in the form of a screen recording video which has specific gaze point movement. This movement can further be analyzed by using the software provided to know the fixation scheme accurately, or to display the heat map as per required.

We needed to get the output in text format / words on which the gaze fixation was high. This is by far not known to us, and was the biggest hurdle. The automation of this output in text format would have made our task easier. We performed the above explained phases of the experiment and then set out to analyze them. The manual analysis was performed over recorded videos, and then the results were tabulated. The statistical significance of data was tested using a student's T-Test on the average of time taken for each POS category. Standard deviation and other statistical tests such as the correlation test were performed and are explained in detail in results section below.

Our work completed in a span of 8 weeks, both phases of the experiment were performed in approximately 5 weeks, and the analysis spanned the rest.

Chapter 6

Results

The experiments thus performed as explained above gave us the following results:

6.1 Phase 1:

6.1.1 POS Category wise:

6.1.1.1 RAW Data

	Lexicographer A	Lexicographer B	Lexicographer C
POS	**Average of Time Taken**	**Average of Time Taken**	**Average of Time Taken**
NOUN	1.478383567	2.253601387	3.579309502
VERB	2.21345865	5.127158281	5.77539597
ADVERB	1.346143964	2.046102697	2.919110521
ADJECTIVE	1.439151781	2.15240053	3.173684716
	1.619284491	2.894815724	3.861875177

	Lexicographer D	Lexicographer E	Lexicographer F	
POS	**Average of Time Taken**	**Average of Time Taken**	**Average of Time Taken**	**AGGR. AVG**
NOUN	2.674775041	4.971096848	2.008642602	2.827634824
VERB	3.983153642	7.900706749	3.156161565	4.692672476
ADVERB	3.403690013	5.626070987	2.04999925	2.898519572
ADJECTIVE	2.807961531	4.894643875	1.882537458	2.725063315
	3.217395056	5.848129615	2.274335219	

The tables shown above display the average amount of time taken by all our annotators categorized according to POS categories.

6.1.1.2 Normalized Data

	Lexicographer A	Lexicographer B	Lexicographer C
POS	Average of Time Taken	Average of Time Taken	Average of Time Taken
NOUN	0.667906566	0.439541996	0.619751359
VERB	1	1	1
ADVERB	0.608163141	0.39907149	0.505439027
ADJECTIVE	0.650182365	0.4198038	0.549518117
	0.731563018	0.564604322	0.668677126

	Lexicographer D	Lexicographer E	Lexicographer F	
POS	Average of Time Taken	Average of Time Taken	Average of Time Taken	AGGR. AVG
NOUN	0.671521935	0.629196477	0.636419448	0.610722964
VERB	1	1	1	1
ADVERB	0.854521397	0.712097179	0.649522912	0.621469191
ADJECTIVE	0.704959382	0.619519751	0.596464224	0.590074606
	0.807750678	0.740203352	0.720601646	

6.1.2 Standard Deviation based on Raw and Normalized data:

	RAW	NORMALIZED
NOUN	1.084403972	0.081899373
VERB	1.642706174	0
ADVERB	1.322128144	0.158876053
ADJECTIVE	1.065590054	0.093958687

6.1.3 Total no. of words annotated:

	A	B	C	D	E	F
POS	**Count**	**Count**	**Count**	**Count**	**Count**	**Count**
Noun	428	433	452	425	423	411
Verb	144	167	206	177	180	109
Adverb	84	66	96	81	78	60
Adjective	146	185	177	154	156	144
Grand Total	**802**	**851**	**931**	**837**	**837**	**724**

Each annotator tagged around 800 – 900 words, the rest of the words out of 2000 were function words.

6.1.4 Ontology wise:

Ontology is a much more abstract concept and is based on senses and there is a good probability of inter-annotator disagreement in senses, so we compiled the ontology results based on the annotator who had marked the maximum no. of senses i.e. Lexicographer 3

Ontological analysis here is distributed POS category wise:

Verbs:

Ontology	Average of Time Taken	No. of words
घटनासूचक (Event)	1.870816444	11
अनैच्छिक क्रिया (Verbs of Non-volition)	2.59201	1
अवस्थासूचक क्रिया(Verb of State)	4.403871355	77
शारीरिक कार्यसूचक bodily action	4.97281795	40
कर्मसूचक क्रिया (Verb of Action)	5.376058091	11
प्रेरणार्थक क्रिया (causative verb)	5.635743	5
संप्रेषणसूचक (Communication)	5.895843818	11
अधिकारसूचक (Possession)	6.00231725	9
परिवर्तनसूचक (Change)	6.517663706	17
विनाशसूचक (Destruction)	8.7992645	3
होना क्रिया (Verb of Occur)	12.06406657	7
भौतिक अवस्थासूचक (Physical State)	13.4773335	2
निरंतरतासूचक क्रिया (Verbs of Continuity)	17.896006	2
कार्यसूचक (Act)	20.2321495	2
मानसिक अवस्थासूचक (Mental State)	74.698983	1
Grand Total	**5.896812948**	**199**

Nouns:

Ontology	Average of Time Taken	No. of Words
भौतिक प्रक्रिया (Physical Process)	0.815685	1
Disease	0.823638667	3
पेशा (Occupation)	0.943731	4
शारीरिक वस्तु (Anatomical)	1.116099	2
खाद्य (Edible)	1.120084083	12
व्यक्तिवाचक संज्ञा (Proper Noun)	1.149471333	3
प्राकृतिक घटना (Natural Event)	1.195029167	6
माप (Measurement)	1.354816667	3
असामाजिक कार्य (Anti-social)	1.367934	1
शारीरिक कार्य (Physical)	1.406077	31
समय (Time)	1.506463	3
व्यक्ति (Person)	1.943074024	43
रोग (Disease)	2.064066	2
मानसिक अवस्था (Mental State)	2.481163667	3
अवस्था (State)	2.678653353	17
भौतिक स्थान (Physical Place)	2.703495675	83
आयोजित घटना (Planned Event)	2.856077	1
संकल्पना (concept)	3.276047	4
मनोवैज्ञानिक लक्षण (Psychological Feature)	3.3478462	5
वनस्पति (Flora)	3.424059	1
मानवकृति (Artifact)	3.435218269	26
प्राकृतिक वस्तु (Natural Object)	3.441991	4
समूह (Group)	3.580438389	54
ज्ञान (Cognition)	3.891006727	12
संज्ञापन (Communication)	4.277537333	3
अवधि (Period)	4.397335214	14
स्वामित्व (possession)	4.505523	1
अमूर्त (Abstract)	4.5785981	10

Nouns (Contd.):

भौतिक अवस्था (physical State)	4.664025	3
जानकारी (information)	4.79190175	4
संप्रेषण (Communication)	5.102635563	16
शारीरिक अवस्था (Physiological State)	5.323988	2
मात्रा (Quantity)	5.519856	1
वस्तु (Object)	5.647904125	8
गुणधर्म (property)	5.76	2
कार्य (Action)	6.074179641	40
बोध (Perception)	6.8069493	10
भाग (Part of)	6.919631	2
प्रक्रिया (Process)	7.687739	1
संज्ञा (Noun)	14.2589998	5
गुण (Quality)	16.381798	4
Grand Total	**3.581711002**	**450**

Adverb:

Ontology	Average of Time Taken	No. of Words
अवधि (Period)	0.968144	1
निषेधात्मक Negative	1.389076083	12
गतिसूचक (Speed)	1.607493	1
मात्रासूचक Quantity	1.9137892	5
संभावनात्मक (Possibility)	2.13774275	4
रीतिसूचक (Manner)	2.51722069	29
क्रिया विशेषण (Adverb)	3.6254454	10
समयसूचक (Time)	3.674002179	28
स्थानसूचक (Place)	5.2460885	4
Grand Total	**2.877390809**	**94**

Adjective:

Ontology	Average of Time Taken	No. of Words
बाह्याकृतिसूचक (Appearance)	1.8357625	2
संख्यासूचक (Numeral)	2.602111133	46
संबंधसूचक (Relational)	2.881833089	56
अवस्थासूचक (Stative)	3.44904625	12
मात्रासूचक (Quantitative)	3.4977741	10
कार्यसूचक (action)	3.549164667	3
गुणसूचक (Qualitative)	3.937929295	44
समयसूचक (Time)	4.034163	4
Grand Total	**3.173684716**	**177**

A total no. of 450 Nouns, 177 Adjective, 94 Adverbs, and 199 Verbs were tagged.

6.2 Phase 2:

Detailed results of phase 2 are given in appendix and the analysis for this phase is still going on. Results are huge in number and comprise of a lot of details. A Small Sample though is provided here for a glimpse of our work.

Time Taken	Word Sense and ID	Context Words Set
13.200246	दुनिया_126760	हमारे आसपास की दुनिया कितनी तेजी से
6.216765	दुनिया_11427	आगे बढ़ने लगे तो दुनिया के लिए
1.184083	होने	मुस्लिम होने के नाते
0.737926	होने	विरोध मेखड़ा होने वाला
46.935237	बताया_24899	वर्णन करना,बताना,बयान करना
8.807983	बताया_27125	सीबीआई ने बताया कि नरसिंह राव

Chapter 7

Discussions

From the result tables in chapter 6, we made the following observations regarding the first phase of our experiments:

1. *Verbs* clearly take the highest amount of time among all the POS categories.
2. The average time taken by verbs is around 75% more than the time taken by other POS categories.
3. The average time taken for tagging a word varies greatly, depending on the skills, background and experience of a lexicographer *i.e.*, 1.619 to 5.848 sec.
4. Adjectives usually take smallest amount of time, followed by nouns and adverbs.

We also performed statistical significance tests (Student's T-Test) to verify if the results obtained during phase 1 are not stochastic. The table below illustrates this fact with an example test statistic between data obtained from lexicographer 1 and lexicographer 2.

	lexicographer *1*	lexicographer *2*
Mean	1.619284491	2.894815724
Variance	0.159983181	2.222001081
Observations	4	4
Hypothesized Mean Difference	0	
df	3	
t Stat	-1.652919294	
P(T<=t) one-tail	0.098458793	
t Critical one-tail	0.978472312	
P(T<=t) two-tail	0.196917586	
t Critical two-tail	1.637744352	

Table 7.1: t-Test: Two-Sample Assuming Unequal Variances

Chapter 8

Conclusion and Future work

In this report, we have presented the variance in difficulty levels of annotation across various POS and ontological categories, as recorded by the eye tracking device. These results are in most cases are not stochastic. They also conform to the views of the lexicographers. The second phase of the experiment which is more critical, is difficult to analyze from the data recorded by the eye tracking device.

But it does form a solid base for a future rule based framework, which we are hoping to call 'Discrimination Net'. It would be self sufficient or rather and independent framework which would be able to discriminate between polysemous word senses on its own, using the context word set. We have had limited no. of instances for each word as well, which if increased substantially, could lead to a healthy coverage of all word sense meanings.

During data recording in the second phase of the experiment we observed that we are constrained by the instruments and tools provided for data recording in the eye tracking device. This is severe hindrance in achieving the end goals. Most importantly inadequate data in such a critical experiment can lead to wrong hypotheses. For this reason, in the future, we would like to develop tools which would computationally and in an error free fashion provide the required data to us, without much manual analysis, as this can be erroneous and cumbersome. Manual analysis of eye gaze and recording of data word by word is not so encouraged but was a step of desperation. Since nothing else could be thought of at that moment, we had to move forward with such a step. Keeping in mind our deadlines, this was the right thing to do.

The tool that would track contextual clues during annotation of a word with a particular sense would record the clue words aiding in the disambiguation from the lexicographers themselves hence avoiding future errors during manual analysis.

References

Bibliography

Agirre, & Rigau. (1996). Word sense disambiguation using conceptual density. *16th International Conference on Computational Linguistics.* Copenhagen, Denmark.

Arindam Chatterjee *et.al*. (2012). A Study of the Sense Annotation Process: Man v/s Machine. *International Conference on Global Wordnets (GWC 2011)*, (p. 8). Matsue, Japan.

Denis Drieghe, A. P. (2008). The word grouping hypothesis and eye movements during reading., (p. 28).

Google. (n.d.). Retrieved from http://www.google.co.in

Laura A. Granka, H. A. (2004). The Determinants of Web Page Viewing Behavior: An Eye-Tracking Study. *Eye Tracking Research and Applications* (pp. 147-154). New York, NY, USA: ACM.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. *5th annual international conference on Systems.* Toronto, Ontario, Canada.

Merja Lehtinen, K. S. (2005). *Usability Research Methods.* University of Tampere, Department of Computer Sciences.

Mitesh Khapra, S. S. (2009). Projecting Parameters for Multilingual Word Sense Disambiguation. *Empirical Methods in Natural Language Prfocessing (EMNLP09).* Singapore.

Navigli, R. (2009). Word Sense Disambiguation: A Survey. *ACM Computing Surveys* , 69.

Navigli, R., & Velardi, P. (2005). Structural Semantic Interconnections: A Knowledge-Based Approach to Word Sense Disambiguation. *IEEE Transactions On Pattern Analysis and Machine Intelligence.*

Nicholas Wade, B. W. (2005). *The moving tablet of the eye: the origins of modern eye movement research.* New York: Oxford University Press.

Picture Source 1. (n.d.). Retrieved from http://www.smashingmagazine.com/2007/10/09/30-usability-issues-to-be-aware-of/

Rada, M. (2005). Large vocabulary unsupervised word sense disambiguation with graph-based algorithms. *Joint Human Language Technology and Empirical Methods in Natural Language Processing Conference (HLT/EMNLP)*, (pp. 411-418). Vancouver, Canada.

Radach, R. (1996). Blickbewegungen beim Lesen: Psychologische Aspekte der Determination von Fixationspositionen (Eye Movements in Reading). Münster/New York: Waxmann.

Seppo Vainio, J. H. (2009). Lexical Predictability Exerts Robust Effects on Fixation Duration, but not on Initial Landing Position During Reading. *Experimental Psychology* , 8.

SMI: Eye Tracking Solutions. (n.d.). Retrieved from http://www.smivision.com/

UXBooth: A Brief History of Eye-Tracking. (n.d.). Retrieved from http://www.uxbooth.com: http://www.uxbooth.com/blog/a-brief-history-of-eye-tracking/

Walker, & Amsler. (1986). The Use of Machine Readable Dictionaries in Sublanguage Analysis. *In Analyzing Language in Restricted Domains* , 69-83.

Wikipedia. (n.d.). Retrieved from http://en.wikipedia.org/wiki/Wiki