Building Tempo-HindiWordNet(THWn) : A Temporal Resource For Hindi

M. Tech Project Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Technology
in
Computer Science & Engineering

by

Dipawesh Rajendra Pawar
(1411CS04)

under the guidance of

Dr. Asif Ekbal

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY PATNA
BIHTA - 801103, PATNA, BIHAR, INDIA
MAY 2016

<u>CERTIFICATE</u>

*This is to certify that the work contained in this thesis entitled "**Building Tempo-HindiWordNet(THWn) : A Temporal Resource For Hindi**" is a bonafide work of **Dipawesh Rajendra Pawar (Roll No. 1401CS04)**, carried out in the Department of Computer Science and Engineering, Indian Institute of Technology Patna under my supervision and that it has not been submitted elsewhere for a degree.*

Supervisor: **Dr. Asif Ekbal**

Assistant Professor,

Department of Computer Science & Engineering,

Indian Institute of Technology Patna, Bihar.

18^{th} May, 2018

Patna.

DECLARATION

Date: 18^{th} May, 2016

I certify that

a. The work contained in this thesis is original and has been done by myself under the general supervision of my supervisor.

b. The work has not been submitted to any other Institute for degree or diploma.

c. I have followed the Institute norms and guidelines and abide by the regulation as given in the Ethical Code of Conduct of the Institute.

d. Whenever I have used materials (data, theory and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the reference section.

e. The thesis document has been thoroughly checked to exclude plagiarism.

Signature of the Student

Dipawesh Rajendra Pawar

Roll No: 1411CS04

Acknowledgments

I would like express heartiest gratitude towards my mentor Dr. Asif Ekbal for his constant encouragement, valuable guidance and support throughout the thesis work. Pattern Recognition and Artificial Intelligence courses taught by Dr. Sriparna Saha and Dr. Asif Ekbal helped me a lot in completing this thesis work. I extend my sincere gratitude towards them. Besides, courses taught by Dr. Arijit Mondal, Dr. Ashok Sairam, Dr. Somnath Tripathy and Dr. Abyayananda Maiti helped me to grasp various concepts in Computer Science. I can never forget their sincere efforts in making the things convincing. I would like to thank all Computer Science PhD Scholars and M.Tech. colleagues in IIT Patna for helping in direct or an indirect way in the thesis work. Special thanks to classmate Navneet Kishore Kunal, PhD scholars Deepak Gupta and Md. Shad Akhtar for discussing and giving various insights in the thesis work. I would like to acknowledge the assistance of PhD Scholar Sabyasachi Kamila and M.Tech. student Vikram Singh of our department for assisting me in manually annotating training set, gold standard set and Sentence classification corpora. Without them this thesis would have been a dream. At last I also thank our institute, IIT Patna, for providing best lab facilities to work.

Bibliographic Notes

Portion of this thesis is based on the following paper:

Dipawesh Pawar, Mohammed Hasanuzzaman and Asif Ekbal , “ Building Tempo-HindiWordnet – A resource for temporal information access in Hindi ”, *In 10th edition of the Language Resources and Evaluation Conference (LREC) 2016*

Contents

List of Figures

List of Tables

Abstract

In this thesis we build temporal ontology- an extension of Hindi WordNet for Hindi Language where we exploit inherent temporal sense(past, present, future, neutral and atemporal) associated with each synset of Hindi WordNet. Due to absense of annotated corpora, we here employ semi-supervised learning to build aforementioned ontology. To deal with the task, we develop two approaches namely One Step classification and Two Step classification. Ontology is built in single step in prior one whereas it is built in two steps in later. To achieve quality expansion of training synsets during semi-supervised learning, we developed two expansion strategies namely confidence based expansion(CBE) and Lexical and semantic relation encoding based expansion(LSEBE). With confidence based expansion we have performed two experiments. First experiment utilize gloss of synsets as feature whereas we employ Lexical and semantic relation encoding(LSE) of synset created over content word of gloss, synset, hypernymy gloss, hyponymy gloss as feature in second experiment.

*LSE oriented CBE as a feature has shown the improvement of **28.74%** in Two Step classification and **2.74%** in One Step classification approach over manually annotated data whereas LSEBE has shown the improvement of **24.52%** and **4.20%** in Two Step and One Step classification approach respectively. We also show sentence level temporality detection as one of the application of developed resource over manually tagged 940 sentences. WEBE has produced the highest accuracy of **66**% over this task.*

The resource is unique in nature as to the best of our knowledge, till no such resource is available for Hindi-a resource scarce language.

Chapter 1

Introduction

In recent years, Temporality has become a new crucial dimension in ranking the search results. Also looking at it's importance in several applications like Temporal question answering[SADCK06], temporal summaries[AGK01], temporal similarity of documents [JYT13], temporal clustering[AGBY09], sequencing of events, etc., it is becoming hot research topic in Natural Language processing(NLP) . Information retrieval where temporality plays pivotal role in deciding document relevance during crawling, indexing and ranking the search results is called as temporal information retrieval.

According to [MJPZ09], out of all queries searched over internet, 7% of queries are temporal. Considering the huge number of searches per day, this amounts to a significantly large value. Some of these queries are explicitly temporal, that is, time expressions like date, year are directly present in query and some are implicitly temporal. Queries like "World Cup 2011", "Indian Prime Minister 2000", "1990's Bollywood songs", etc. fall in former category, whereas quries like "Newton's childhood","World-war II", "Recent Bollywood songs", etc. fall in later one. [NRD08] has showed that 1.5% of the search queries are explicitly temporal which was revised to 1.21% after removal of false positive queries like "Windows 2000" by subsequent studies[CDJ11]. The rate for implicit temporal queries has not been measured so far. All this highlight significance of time in refining and ranking search engine results.

[JJR13] in conducted a survey over 110 people to understand or highlight temporal aspects in user information needs and presented interesting results. In the results they claimed that most of the time user queries needs to be addressed with recent information though sometimes it also needs to be addressed with past or future related information. For example, query “Did Mumbai Indians win today” needs recent information whereas queries like “release date of FAN movie” , “History of pyramids” needs future and past related information respectively. Results also showed that good amount of information needs are also due to seasonal or continuous interests. So, looking at these temporal aspects in user search queries and to enhance quality of results returned by search engines for such temporal queries, NTCIR-11 workshop conducted ’temporalia’ task which in turn was consisting of two sub-task namely Temporal query intent classification(TQIC) and Temporal information retrieval(TIR). In first sub-task participants needs to classify several web queries among four classes namely recency-related, past-related, future-related and atemporal.

Thus TQIC sub-task of NTCIR-11 workshop highlight the need to predict whether the user query is needing past related, recency related or future related information. However, most of the earlier studies in computational linguistics and data mining has concentrated on identifying temporal expressions and events associated with them, for example, TempEval task of SemEval 2007 workshop [VGS+07a]. Also most of the studies which tries to address these temporal aspects of information makes use of different linguistic constructs such as presence of temporal expressions like before, now, after, today, etc.,document creation time(DCT), explicit timexes (time expressions). But most of the time context does not contain such linguistic constructs which then makes it hard to predict temporal needs of the queries. E.g. the query “पानिपत कि लड़ाई हुई थी क्या (paanipat ki laDdaaEE huEE thee kyaa; Did the fight of Panipat happened) ? ” requires answer to be given from time view point but no time specific construct is present in query from which implicit presence of time could be identified. DCT and explicit timexes also have some problems [JYT13]. Prior need not depict the actual time to which content of document refers to. Secondly, it need not always available e.g. web documents may not contain time-stamps and if at all they contain, it can not be trusted as it might refer to time at which it is recently modified.

whereas, later if present in document need not actually represent time to which document's content belongs to as it may be weakly connected to topic of the document So, there is lack of work which exploits hidden temporal sense associated with word senses.

To exploit such hidden,inherent temporal sense associated with word senses, Dias et al. [DHFM14] developed TempoWordNet(TWn) an extended version of WordNet(Miller 1995) where they supplement temporal sense(past, present, future or atemporal) to current information like gloss, synset, sematic relations associated with each sense of all words in WordNet. They exploited semi-supervised learning wherein gloss associated with synsets are quantitatively analyzed for automatic creation of TempoWordNet. The Authors used this resource for solving TQIC sub-task of temporalia task in NTCIR-11 workshop and shown that significant improvement can be achieved if such resource is used. Their system ranked first in predicting recency related queries and shown average performance in past and future related queries in TQIC sub-task.

Being inspired from [DHFM14], we developed Tempo-HindiWordNet an extended version of Hindi WordNet. In Tempo-HindiWordNet, we augment each synset of HindiWordNet with hidden, inherent temporal sense(past, present, future, neutral, atemporal) it denotes or connote. One might argue that the word लंबा_समय(laNbaa samaya; Long time).01.n - एक लंबा समय(Ek laNbaa samaya;A long time) is having some time sense in it then with which temporal category it will be annotated. Such synsets can not be specifically categorized among past, present or future. So, in addition to three temporal categories in TWn, We here introduce new temporal category neutral which denotes overlapping time sense among past, present and future. Classes, their meanings and examples are depicted in Table 1.1.

Class	**Meaning**	**Example**
Past	already happened	कल(kala; yesterday)
Present	currently going on	आज(AAja: Today)
Future	about to happen	भावी(bhaavee; Prospective)
Neutral	Overlapping time sense	संध्या(saNdhyaa; Evening)
Atemporal	absence of time sense	उष्ण(UShNNa; Hot)

Table 1.1: Example classes with which HindiWordNet synsets are annotated

1.1 Task Definition

In our work we build a resource that can be effectively used for temporal information access. Given the Hindi WordNet, the task is to identify temporal information associated with each of the synsets present in it. We use four temporal senses (i.e past, present, future, and neutral) and one atemporal (denoting non-temporal instances) sense. Following frame shows some examples of the task.

- **Word :** कल (kala:Yesterday)

  Synset: अतीत (Ateeta; Past), अतीतकाल (Ateetakaala; Past time),भूतकाल (bhootakaala; Past time), अतीत काल(Ateet kaala; Past time),कल (kala; Yesterday), भूत काल (bhoot kaala; Past time),गत काल (gat kaal; Past time),पिछला ज़माना; Previous genre (pichhalaa Zamaanaa),पूर्वकाल (poorvakaala; Anterior)

  Gloss : बीता हुआ समय या काल (beetaa huAA samaya yaa kaala; Passed time)

  Example sentence : "यह उपन्यास अतीत की घटनाओं पर आधारित है । (yah Upanyaas Ateet kee ghaTanaaON para AAdhaarit hai ;This novel is based on past events)/कल की बातों को याद करके दुखी होना अच्छा नहीं । (kal kee baatoN ko yaad karake dukhee honaa Achchhaa naheeN ;It is not good to get upset by remembering past events)"

  Temporal Annotation : Past

- **Word :**कल (kala; Yesterday)

  Synset :कल (kala; Yesterday)

  Gloss : आज के बाद आनेवाले पहले दिन को (AAj ke baad AAnevaale pahale din ko; Day that comes just after today)

  Example sentence : "मैं कल घर जाऊँगा । (maiN kal ghara jaaOONngaa; I will go to home tomorrow)"

Temporal Annotation : Future

1.2 Contribution

We develop two approaches to automatically time tag all synsets in HindiWordNet namely 'One step classification' and 'Two step classification'. In one step classification all synsets are classified among aforementioned classes in single step whereas in two step classification they are classified among two classes namely, temporal(presence of time sense) and atemporal(absense of time sense) first and synsets which are predicted to be temporal are further fine grained among past, present, future and neutral categories.

Due to absence of time tagged corpus, semi-supervised learning is employed in both the techniques. For achieving sound expansion during semi-supervised learning we develop three strategies. In first strategy, we expand training set with the synsets which are annotated with desired class with highest confidence. The technique is referred to as confidence based expansion(CBE). In the second strategy, we support confidence based expansion technique with Lexical and semantic relation encoding(LSE) of synsets which are created over gloss, synonyms, hypernymy gloss and hyponymy gloss of synsets. These word Embeddings developed in second strategy are further used in the third strategy to expand training set with synsets which are sematically very close to existing synsets. This strategy is referred to as 'Lexical and semantic relation encoding based expansion(LSEBE)' technique.

To evaluate developed techniques we created two gold standard sets. First one consists of 180 synsets manually classified among desired classes by 3 annotators with moderate agreement of 0.586. We used [Gee12] for measuring multi-rater kappa agreement. Second gold standard consists of 940 sentences retrieved from 300 news articles taken from IL-TIMEX 2012 corpus [RM15]. These are also manually classified among past, present and future by 3 annotators with a substantial agreement of 0.78. With the help of these gold standard sets further attempts to construct such resource can be easily evaluated.

We also show sentence level temporality detection as one of the application of our resource.

1.3 Structure of the resource

Each synset of our Tempo-HindiWordNet contains the following information: Offset, Word, PoS, Sense number, Gloss, Class and Prediction value. Offset is an unique identifier associated with each synset, PoS denotes the part-of- speech category of the target word, sense number denotes the number associated with the target synset in Hindi WordNet, gloss denotes the meaning associated with the synset, class denotes the predicted time sense of the synset by the temporal classifier and prediction value denotes the confidence with which annotated time sense is predicted. Number of word senses tagged as temporal and atemporal by different strategies are depicted as temporality statistics in methodology chapter. We show the structure of the resource with few examples in the following frame:

Synset_Offset, Word, PoS, Sense number, Gloss, Class, Prediction_Value

- 8741, 11, ADJECTIVE, 1, दस से एक अधिक (das se Ek Adhik; One more than ten), atemporal, 0.99
- 9234, अतरसों(AtarasoN; Atrason), NOUN, 2, गत परसों से पहले का दिन या आज से पहले का तीसरा दिन(gat parasoN se pahale kaa din yaa AAj se pahale kaa teesaraa din; The day before the day before yesterday or before the last day of the third day), past, 1
- 2688,अद्यतन(Adyatan; up to date), ADJECTIVE, 1, जिस पर इस समय की बातों या विशेषताओं की पूरी छाप हो(jis para Is samaya kee baatoN yaa visheShataaON kee pooree chhaap ho; The one on whome there is full impression of today's time), present, 1

- 23772, आइन्दा(AAIndaa; Hereafter), ADVERB, 1, इस समय के बाद से(Is samaya ke baad se; Since this time), future, 0.295
- 4216, स्वप्न (svapn; Dream), NOUN, 1, सोते समय दिखाई देने वाला मानसिक दृश्य या घटना (sote samaya dikhaaEE dene vaalaa maanasik dRishya yaa ghaTanaa; Psychic scene or event that appears at bedtime), neutral, 0.942

1.4 Hindi WordNet

In any natural language, word can have multiple meanings associated with it. Such words are referred to as polysemous words. Hindi WordNet is a lexical resource which stores words in Hindi Language not only with their meanings as that of the dictionary but also with their semantic and lexical relations. These content associated with each word in Hindi WordNet is as follows :

1. Gloss : It explains concept associated with that particular sense of the word. It consists of two parts-
   - Text definition : It explains the meaning associated with the synset. For example, "एक फल जो खाया या चूसा जाता है"(Ek phal jo khaayaa yaa choosaa jaataa hai; A fruit that is eaten or sucked) is the text definition of the synset आम(AAma; Mango).n.01 which explains the meaning of आम(AAma; Mango) as a fruit.
   - Example sentence : It shows example usage of the particular sense of the word in a sentence. For Example, "तोता पेड़ पर बैठकर आम खा रहा है ।"(totaa peDd para baiThakara AAm khaa rahaa hai; parrot eating Mango sitting on the tree) is the example sentence associate with synset आम.n.01(AAma; Mango).
2. Synset : It is group words with same meaning. For example, [आम, आम्र, अंब, अम्ब, आँब, आंब, रसाल, च्यूत, प्रियांबु, प्रियाम्बु, केशवायुध, कामायुध, कामशर, कामांग(AAm, AAmra,

ANb, Amb, AANnb, AANb, rasaal, chyoot, priyaaNbu, priyaambu, keshavaayudh, kaamaayudh, kaamashara, kaamaaNga)] are synonym set of synset आम.n.01(AAma; Mango).

3. Position in ontology : Ontology is the hierarchical organization of concepts. It shows position of the synset in that hierarchical organization. For example, Ontology for the synset पाठशाला(paaThashaalaa; School) is as follows :

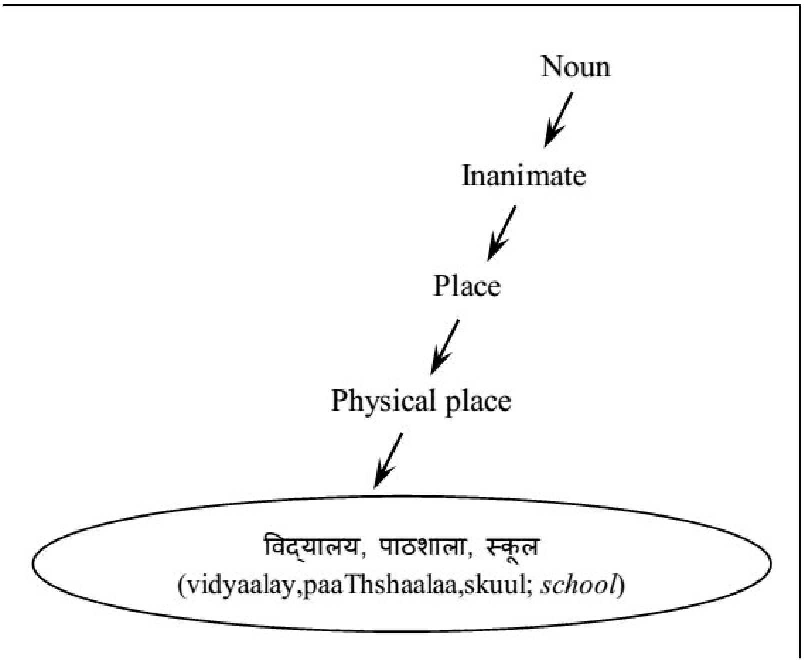

Fig. 1.1: Ontology for synset of पाठशाला(paaThashaalaa; School)

4. Relations in Hindi WordNet :

   - Hypernymy and Hyponymy(is a kind of): Hyernymy and hyponymy are relations among synsets i.e. semantic relation. A is hypernymy of B means that A is generalized entity of B. Hyponymy is inverse of hypernymy. Both of them are transitive and asymmetrical in nature. For Example शेर(shera;Tigar) is a kind of पशु(pashu;Animal). So, शेर(shera;Tigar) is hyponym and पशु(pashu;Animal) is hypernym.

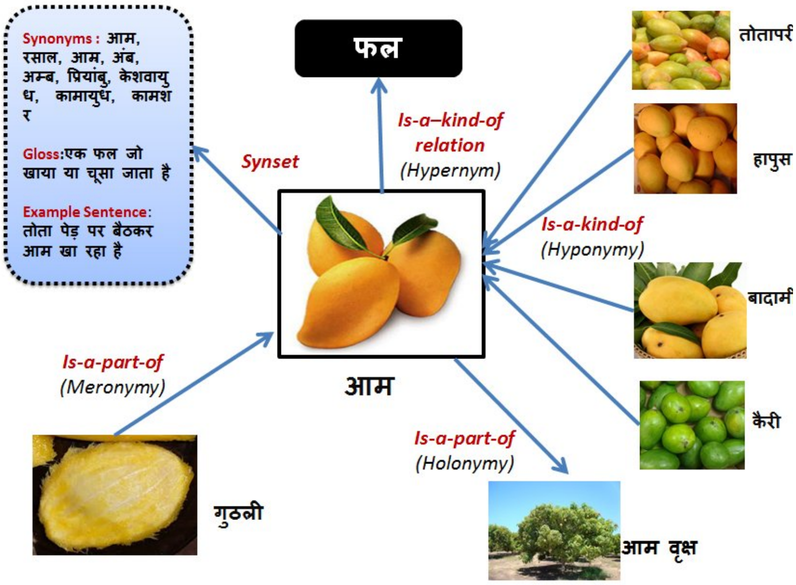

Fig. 1.2: Relations in Hindi WordNet for the synset आम.n.01(AAma; Mango)

- Meronymy and Holonymy (Part-whole relation): It is sematic relation among synsets. If C is part of D then C is said to be meronymy of D. Holonymy is inverse of meronymy. These relations are also transitive and asymmetrical in nature. For Example, जड़(jaDda;Root) is a part of पेड़(peDda;Tree) so, जड़(jaDda;Root) is meronym and पेड़(peDda;Tree) is holonym.
- Entailment : It is a relationship among two verbs. Verb C entails verb D means that if C is logically true then D must be logically true. It is unilateral relation i.e. it is one way relation. For Example, if someone is snoring then he/she must be sleeping. So, खर्राटे लेना(kharraaTe lenaa;Snoring) entails सोना(sonaa;sleep).
- Troponymy: It is a relationship among two verbs. If C ellaborates D in some specific manner then C is troponym of D. For Example, फुसफुसाना(phusaphusaanaa; Whisper) ellaborates the verb बोलना(bolanaa; to Speak) So, फुसफुसाना(phusaphusaanaa; Whisper) is troponym of बोलना(bolanaa; to Speak).
- Antonymy: It is a relation among two words i.e. it is a lexical relation. If word C express opposite meaning to the meaning expressed by the word D in a context then C is said to be antonymy of D or vice versa. So, antonymy is

a symmetrical relationship.

- Gradation: It is a relation among words. So, its a lexical relation. It represents intermidiate concept among two opposite concepts. For Example,
  - सुबह(subaha; morning) दोपहर(dopahara; afternoon) शाम(shaama; evening)
  - गरम(garama; hot) गुनगुना(gunagunaa; Lukewarm ठंडा(ThaNDaa; Cool)
- Causative : In Hindi Language, some morphological changes in root form if a word denote causation. For Example, खिलाना(khilaanaa; to Feed) is a causative verb of खाना(khaanaa; Food). This relation show interdependency among base form of a verb and its causative form.

1.5 Thesis Overview

The roadmap of the thesis is as follows. Next section summarizes various project works which directly relates to our work. Various strategies developed to tackle the problem are explained in detail in chapter 3. Chapter 4 report evaluation results on manually annotated data and their analysis. We show sentence temporality detection as one of the application of our resource along with accuracy results in chapter 5. We conclude with contribution of the thesis work and probable ways for future research in chapter 6.

Chapter 2

Literature Survey

There are two kinds of work that are closely associated with this task of detecting inherent temporal sense associated with the words.

2.1 Temporality in NLP and IR

In recent years lots of research activities in NLP and IR are focusing on time as one of the important dimension. [MPG05] has exhaustively summarized initial works in this domain. The introduction of TempEval Temporal Relation Identification task[VGS+07b] in SemEval 2007 workshop and subsequent challenges(TempEval-2[VSCP10] and -3) has clearly established the importance of time to tackle different problems in NLP and IR. TempEval-1 was consisting of following task :

- Task A : Detecting relationship between time expression and events present at sentence level
- Task B : Detecting relationship between events and document creation time.
- Task C : Detecting relationship between main events among consecutive sentences

Six teams had participated in TempEval-1. Approaches used by some of these teams are as follows :

- Team 1[BM07]: They had annotated one of the relationship of TempEval-1 to each pairs of events /time using pairwise classification. They had used semantic and syntactic features to train SVM classifiers. Their system for task B performed the best. So results of task B are used as feature for other tasks.

- Team 2[MSF07]: They had used machine learning and rule based approaches along with NLP tools and linguistic resources developed at LCC to identify temporal relations. They had used features directly obtained from TempEval data and some features derived from these features including tense and aspect shifts and whether a model auxiliary is present.

- Team 3[HSG07]: They had used very simple approach for TempEval task wherein they have used Weka implementation of different classifiers like SVM, lazy.kstar rules, DecisionTable, etc. They used some simple features which are directly computed from TempEval event/time annotation or the ones which can easily be computed from documents without any deep NLP analysis.

In TempEval-2 the system which have used conditional random field juxtaposed with parsing techniques had given the best results[UA10]. In TempEval-3 all participated systems had performed very well for the same set of tasks with more rigorous restrictions.

Motivated from these set of tasks [RM15] has developed and analyzed various approaches to detect and classify temporal expressions in Hindi. First of various such approaches is rule based which on the basis of hand crafted set of rules returns tagged temporal expressions in the input. In the next approach they have built CRF classifier on manually tagged dataset. This approach not only detects temporal expression but also classifies them to various temporal categories. Later on they tested SVM classifier replacing the CRF classifier with the same methodology as followed in second approach and has shown that SVM classifier has produced comparable results. In the fourth approach they provided information produced by rule based classifier as an additional feature to classifier developed in second and third approach and has shown that it exceeds the performance of both the classifiers. In the next technique they provided the data tagged with rule based approach as an additional training data to CRF and SVM classifier developed in second and third approach. Further they have merged these three approaches with voting based system which selects the best possible tag for each token from the output of each of those approaches. They have shown that this voting based system has produced the highest accuracy for detecting and classifying temporal expressions in Hindi.

In IR, Time has increasingly becoming important parameter to rank search results. [Met07] has pointed out that time is one of the key five aspects that determine the document's credibility besides relevance, accuracy, objectivity and coverage. This denotes that value and quality of information is inherently time dependent. This has give rise to new field known as temporal information retrieval(T-IR). It deals with the traditional IR tasks such as crawling[KTSD11], indexing[ABBS12], ranking[KBM11] giving equal importance to time relevance along with document relevance. [CDJJ14] has handled temporal query classification as an application of T-IR and tackled the problem by identifying top relevant dates in web snippets with respect to given temporal query disambiguating them with distributional metric called GTE. Temporalia task of NTCIR-11 workshop[JJB+14] has further denoted importance of temporal query classification by introducing the sub-task temporal query intent classification(TQIC).

In TQIC subtask, participating teams were expected to classify the query into one of the four classes past, recency, future and atemporal. Six teams had participated in TQIC subtask. It was observed by organizers that identification of recency related queries was most difficult whereas, future related queries were easiest to classify. Also, it was observed that no single participant's system was effective across all the classes. This denotes difficulty of the task. Methodologies used by participating teams is given below:

- HITSZ[HTX+14] team has used various language related features such as N-Gram words of query, POS N-grams, named entities, normalized date, date distance, special words from time sensitive word dictionary. They extracted these features either from query itself or from results produced by search engine once query is submitted to it.

They then used voting technique to combine the results of Rule based methods and results produced by multiple classifiers. In rule based method they have used distance between query issue time and date mentioned in query, combination of verb tense and date distance, time sensitive word dictionary, etc.

- HULTECH[HDF14] team has used ensemble learning to deal with the bias of different classifier developed. They have developed Tempo-WordNet(TWn) to detect temporal orientation of words in wordNet[Mil95]. Then they have used these temporal orientation of words as feature along with various features such as web snippet results, query as a sentence. HULTECH system has outperformed all other participants in predicting recency related queries.

- Team TUTA1[YKR14] had used semi-supervised and supervised classifiers to predict temporal intent of queries. Time gap features, verb tense features and lemmas, and named entities are used as features by them. They also had employed AOL 500K query session datased to increase training data.They had submitted 4 runs, two with logistic regression classifiers tuned with different parameter settings, one with linear SVM using AOL dataset and last one with logistic regression classifier tuned with lemma and named entities only as a feature.

- Andd7 team[SSM14] has fed bag of words, query length, difference of query issue time and temporal expression in query and verbs in query as features to SVM and Naive bayes classifier. Out of three runs submitted, first two were aforementioned classifiers and in the third run they submitted agreement decision among these classifiers.

- MPII's team[BB14] had submitted three runs each with different feature sets picked by simulated annealing. They had used various features extracted by applying POS tagger, DMOZ directory [1], publication dates and the content of document collections used for finding query-time associations information.

[1] http://www.dmoz.org/

- Team UniMan[FN14] has used nineteen features such as features extracted from TWn, some computed from query, submission date related, extracted from Wikipedia titles, etc. They had submitted three runs with different group of features and found the system with minimal features to be the best one.

2.2 Association of words/senses to various cognitive features

Lot of research activities in the domain of computational linguistics and cognitive science has tried to link words to different cognitive features like abstractness-correctness[Col81], sentiment[ES05], imageability[Col81], colors[ÖSMP11], time[HDFM14, DHFM14], etc. Colthearts computerized psycho-linguistic database[Col81] provides the way to deal with various properties of words such as number of syllables, regularity or irregularity of their spelling, their rated imageability/correctness and their frequency of occurrence in printed English which when increases in length known to hamper experimentation. Esuli and Sebastiani[ES05] has determined the sentimental orientation(positive, negative or neutral) of subjective terms through semi-supervised quantitative analysis of definitions of these terms available in online dictionaries. These sentiment orientation is then used by many other researchers as a feature for the task of sentiment analysis. [ÖSMP11] has developed three approaches based on image analysis, language models, and latent semantic analysis to automatically map colors to emotions that we perceive about them. For example, the color black usually is associated with the words black or death and these two words are often associated with fear so, black can be associated to fear. They investigated such relationship among colors and words to depict emotion the words connote. Hasanuzzaman and others[HDFM14] in their paper of propagation strategies to build temporal ontology has depicted different strategies to build temporal resource for English. Through this resource they tried to extract hidden temporal orientation of subjective terms through quantative analysis of their gloss i.e. definition available in WordNet[Mil95]. They used this developed resource for automatic classification of temporal queries among the classes recency, past, future in TQIC sub-task of Temporalia task in NTCIR-11 workshop.

All these highlight different research activities carried out in past to associate words to various cognitive features.

Chapter 3

Methodology

We build Tempo-HindiWordNet using a semi-supervised learning approach where each synset of Hindi WordNet is classified to one of the five classes as mentioned earlier. We develop two strategies:

1. In our first strategy we build a multi-class classifier that directly classifies WordNet synsets into four temporal and one atemporal categories in single step. We call this as one-step classification framework.

2. The second strategy works in two-steps. In the first step we build a binary classifier that classifies WordNet synsets into temporal vs. atemporal. In the second step, we build a multi-class classifier to further classify the temporal synsets into four categories, namely past, present, future and neutral. This approach is called as two-step classification framework.

In both of these strategies, due to unavailability of labeled corpora, we explore semi-supervised learning strategy that learns iteratively starting with a very small set of seed entities. We create a list of seed entities that represent aforementioned classes. These seed words are represented by their respective gloss in the training set. We use unigram model for representing the glosses. These unigrams are weighted with their term frequencies. We developed two expansion strategies to achieve quality expansion during semi-supervised learning. Initial set of seed words is then iteratively expanded using these expansion strategies. Here, initial seed list is augmented by adding the instances which are predicted with high confidence by the classifier or on the basis of their semantic distance with respect to existing seed entities. Hence, we believe that our incremental learning process is always supported with good quality instances for training in each iteration. Once the expansion process stops, classifier built is used to annotate each synset of Hindi WordNet with time sense. We sketch the algorithmic steps in Algorithm 1. It composes of two key steps, viz. creation of initial seed list and confidence based expansion strategy.

Algorithm 1 Basic steps of the algorithm

```
Select initial set of seed words.
repeat
Train the model on initial seed entities
Test the model developed
Expand seed entries
until cross-validation accuracy drops
```

3.1 Creation of initial seed list

Just as words with positive or negative connotation or denotation are used in sentiWordNet [ES05], here we start with seed words representing the classes: past, present, future, neutral and atemporal. We show some of the examples in Table 3.1. Selection of such seed words is very important as semi-supervised learning greatly depends on its quality. Distribution of seed words among the various temporal classes is important in order to ensure that temporal classifier is not biased to any particular class. One of the ways to select such initial temporal seed words is to select the words which appear in sub-tree of word ”समय (samaya; Time)” in Hindi WordNet. But, almost all the words extracted from this sub-tree are found to belong to the noun POS categories. However, there are significant number of non-noun words which denote or connote temporal senses. In order to ensure that we do not miss these words and our process is not biased to some certain PoS categories, we follow a very rigorous process while creating the seed list manually, and thereby guarantee to cover all kinds of PoS categories. These seed words are annotated by three annotators and then we evaluated multi-rater agreement using online resource[1] which uses Fleiss multi-rater kappa statistics[Fle71]. The seed list is found to have substantial agreement of 0.73

We create a list of 85 seed words, out of which 37 are atemporal and the rest are temporal. Temporal entities are equally distributed among all the four temporal classes, namely past, present, future and neutral. In one-step classification scenario we consider the list of seed words that contain all these classes. In two-step classification framework we consider the seed words annotated with only two classes (temporal vs atemporal) in the first step, and then in the second step we consider the seed words annotated with the fine-grained classes (four classes, i.e. past, present, future and neutral).

past		**present**	
word(sense no)	PoS	word(sense no)	PoS
कल(kal; Yesterday)(1)	adv	मौजूदा(maojoodaa; Current)(3)	adj
गुज़रा(guZaraa; Gone)(2)	adj	वर्तमान(vartamaan; Present)(2)	n
विगत(vigat; Past)(2)	adj	आज कल(AAj kal; nowadays)(1)	adv
अतीत(Ateet; Past)(1)	n	आज(AAj; Today)(1)	adv
जन्म(janm; Birth) (1)	n	आज(AAj; Today)(2)	n
Future		**Neutral**	
word(sense no)	PoS	word(sense no)	PoS
पश्चात्(pashchaat; After)(1)	adv	इद्दत(Iddat;)(1)	n
आगामी(AAgaamee; Upcoming)(1)	adj	नित्यप्रति(nityaprati; Everyday)(1)	adv
कल(kal; Tomorrow)(2)	adv	धीमा(dheemaa; Slow)(1)	adv
कल(kal; Tomorrow)(2)	n	ठोहर(Thohara; Tohr)(1)	n
अपेक्षित(ApekShit; Expected)(2)	adj	अरसा(Arasaa; Ages)(1)	n
Atemporal			
word(Sense no)		PoS	
सर्वाधिक(sarvaadhik; Most)(1)		adj	
अनिवासी_भारतीय(Anivaasee_bhaarateeya;Non Resident Indian)(1)		n	
ताज्जुब_होना(taajjuba_honaa; To be surprised)(1)		v	
राष्ट्रीकरण(raaShTreekaraNN; Nationalization)(1)		n	
अस्मिता(Asmitaa;)(4)		n	

Table 3.1: Few examples of manually selected seed words

3.2 Propagation Strategies

We developed two propagation strategies to achieve quality expansion during iterative learning process. These are named as "Confidence based expansion(CBE)" and "Lexical and Semantic relation encoding based expansion(LSEBE)". We have used confidence based

[1]https://nlp-ml.io/jg/software/ira/

expansion in two experiments.

1. Only gloss of the synsets is used as feature to represent it in training set(Gloss oriented CBE).

2. Lexical and semantic relation encoding(LSE) over content words of gloss, synset, hypernym gloss, hyponym gloss and gloss is used as feature to represent synset in training set(LSE oriented CBE).

3.2.1 Confidence based expansion(CBE)

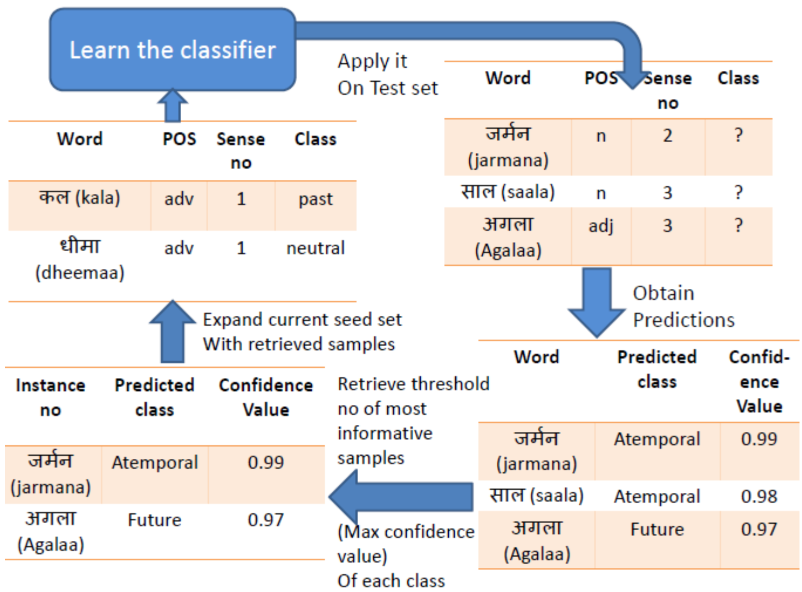

Fig. 3.1: Pictorial representation of CBE

Due to absence of annotated corpora, we here employ semi-supervised learning. First we train three different classifiers namely support vector machine[Joa02], Naive Bayes[JL95] and decision trees[Qui93] on initially selected seed list using 10-fold cross validation. Then we test these developed models on test set. Test set is composed of all synsets in Hindi WordNet. From the predictions generated over test set, we expand current training set with the synsets which are predicted to belong to certain class with highest confidence. The instances with highest confidence of prediction are given higher priority while expanding the initial seed data so as to keep denotational and connotational properties of initial seed entities intact. In every iteration we add the instances in such a way that the ratio of instances of different classes are maintained at par with the initial class distribution. This is to ensure that the classifier is not biased to any specific class. We define a threshold, and based on this value we select the instances to be added to the initial seed list. This threshold depends on the size of current seed word set, and we fix it through a series of experiments. Process is repeated iteratively till cross-validation accuracy drops.

1. **Gloss oriented CBE**

   In the first experiment of developing Tempo-HindiWordNet(THWn), we used gloss of synsets as the only feature to represent synsets in training set. These are encoded as unigrams in training set and weighted with number of their occurrences. Gloss defines the meaning of the synset hence we started the development of THWn with it as a feature. Some of the automatically extracted seed words during the expansion process are as shown in Table 3.2

past		**present**	
word(sense no)	PoS	word(sense no)	PoS
पूर्वाह्न(poorvaahna; Forenoon)(2)	n	शुरू(shuroo; Beginning)(2)	adj
उत्तरजीवक(Uttarajeevaka;)(2)	n	निर्माणाधीन(nirmaaNNaadheena; Under Construction)(2)	adj
अपमृत्यु(ApamRityu; Unnatural death)(2)	n	ग्रीक(greeka;)(2)	adj
सबेरे(sabere; Morning)(3)	adv	प्राकृतिक_वस्तु(praakRitik vastu; Natural object)(2)	n
उत्तरजीवी(Uttarajeevee; Survivor)(2)	adj	आज(AAj; Today)(2)	n
Future		**Neutral**	
word(sense no)	PoS	word(sense no)	PoS
प्रचालित(prachaalita; Operated)(2)	adj	श्लेषानक्षत्र(shleShaanakShatra;)(2)	n
अनचीत(Anacheeta;)(3)	adj	नछत्र(nachhatra;)(3)	n
परिपार्श्व(paripaarshva;)(4)	n	विषुव_वर्ष(viShuva varSha; Equinox) year(2)	n
उजबेकी(Ujabekee;)(2)	adj	मध्य-काल(madhya-kaala; Middle Ages)(2)	n
भौतिक(bhaotika; Physical)(7)	adj	तेज(teja; Fast)(3)	n
Atemporal			
word(Sense no)		PoS	
दिवालिया_बनाना(divaaliyaa banaanaa; To bankrupt)(2)		v	
नजरबन्दी(najarabandee; Internment)(2)		n	
खिलाफत(khilaaphata;)(2)		n	
मेज़बानी(meZabaanee; Host)(3)		n	
किनारा(kinaaraa; Coast)(6)		n	

Table 3.2: Few examples of automatically extracted seed words during CBE: Only gloss as feature

Temporality statistics: In Tempo-HindiWordNet created with this approach, out of all the words with all possible senses, 7215 are tagged as past, 172 are tagged as present, 501 are tagged as future, 2588 are tagged as neutral whereas 133390 are tagged as atemporal.

Cross-validation results: Cross-validation results of different classifiers developed in One Step classification approach are depicted in 4.1

Classifier		**Cycle 1**	**Cycle 2**	**Cycle 3**	**Cycle 4**	**Cycle 5**	**Cycle 6**
J48	precision	0.624	0.596				
	recall	0.667	0.609				
	F-measure	0.637	0.592				
SVM	precision	0.58	0.647	0.756	0.739		
	recall	0.594	0.655	0.738	0.738		
	F-measure	0.551	0.624	0.718	0.734		
Naive Bayes	precision	0.864	0.894	0.903	0.922	0.915	0.841
	recall	0.848	0.889	0.899	0.916	0.913	0.828
	F-measure	0.853	0.885	0.896	0.914	0.91	0.821

Table 3.3: Cross-validation Results of various classifiers developed in One Step classification framework: Only gloss as a feature

J48 classifier has iterated only for two iterations. This might be due to the reason that J48 could not handle too many classed and hence failed to capture their properties. On the other hand SVM and Naive Bayes has iterated very well and hence must have spreaded temporality over more number of synsets. Cross-validation Results of various classifiers developed in Two step classification framework are presented in Table 4.2. It has been observed that SVM classifier is very good at capturing and spreading temporality. Number of iterations, it has gone through is also the evidence of this. From cross-validation results it is seen that J48 and Naive Bayes captures properties of temporality too early and hence end up in less number of iterations. Due to this they are unable to spread temporality over more number of synsets.

2. **Lexical and Semantic relation encoding(LSE) oriented CBE**

   In this experiment we support confidence based expansion strategy with vectorial representation(LSE) of the synset encoded over content words of gloss, lexical relation(synonyms) and semantic relation(hyponym and hypernym gloss) of the synset.

Classifier		**Cycle 1**	**Cycle 2**	**Cycle 3**	**Cycle 4**	
J48	precision	0.664	0.643	0.708	0.551	
	recall	0.643	0.613	0.676	0.587	
	F-measure	0.643	0.578	0.661	0.561	
SVM		**Cycle 1**	**Cycle 2**	**........**	**Cycle 23**	**Cycle 24**
	precision	0.209	0.361		0.958	0.939
	recall	0.375	0.407		0.958	0.939
	F-measure	0.227	0.255		0.957	0.938
Naive Bayes		**Cycle 1**	**Cycle 2**	**Cycle 3**	**Cycle 4**	
	precision	0.533	0.654	0.617	0.615	
	recall	0.554	0.597	0.603	0.6	
	F-measure	0.517	0.577	0.57	0.561	

Table 3.4: Cross-validation Results of various classifiers developed in Two Step classification framework: Only gloss as a feature

Hypernym and hyponym gloss assists to spread or detect connotative temporality as hypernym is generalization and hyponym is specialization of the synset. E.g. विराम_काल(viraam kaala; Rest period) is the kind of काल(kaala; period). So, काल(kaala; period) is hypernym and विराम_काल(viraam kaala; Rest period) is hyponym. Both of them are temporal words. As LSE is also encoded over hypernym and hyponym gloss, if विराम_काल(viraam kaala; Rest period) is present in training synsets then its hyernym काल(kaala; period) will be considered for expansion. Our hypothesis is such vectorial representation of semantically related synsets must be related in some manner to each other and hence such representation will help classifier to detect class of the synset.

Lexical and Semantic relation encoding(LSE)

LSE of the synset are obtained using Word Embeddings trained using word2vec[2] tool[MCCD13]. With the help of this tool, word Embeddings can be created using two techniques namely skip-gram model and continuous bag of word model(CBOW). Skip gram model transforms words to vector maximizing the probability of word based on words in its context in the same sentence whereas CBOW predicts the word on the basis of its context specified with the help of window parameter of word2vec

[2]https://code.google.com/archive/p/word2vec/

command. Word2vec tool is trained on Bojar's(2014)[BDR+14] corpus. It then creates vocabulary over Bojar corpus and tag each word in the vocabulary with its vectorial representation of specified length. Bojar corpus is consisting of 44 million Hindi sentences. We obtained word embedding of 200 length using skip gram model keeping the window size 7. We use skip gram model as it is known to produce better word embeddings over large training corpus.

To create vectorial representation of the synset, we take average of word embeddings of content words of gloss, Synset, hypernym gloss and hyponym gloss. Suppose, synset is consisting of m content word in its gloss, Synset, hypernym gloss and hyponym gloss then vectorial representation of the synset is given by the equation 3.1.

$$\frac{\sum_m WE(w_m)}{m} \tag{3.1}$$

Cross validation results: Cross-validation results for various classifier developed in this experiment using one step classification framework are reported in Table 4.3. Results shows that J48 stops at iteration 19 whereas SVM and Naive Bayes stops very early at an iteration 10 and 9 respectively. So, J48 spreads temporality over more number of synsets.

Classifier		**Cycle 1**	**Cycle 2**	**...**	**Cycle 18**	**Cycle 19**
J48	precision	0.601	0.755		0.996	0.996
	recall	0.646	0.771		0.996	0.996
	F-measure	0.611	0.757		0.996	0.996
SVM		**Cycle 1**	**Cycle 2**	**........**	**Cycle 9**	**Cycle 10**
	precision	0.719	0.72		0.908	0.906
	recall	0.615	0.619		0.888	0.887
	F-measure	0.635	0.638		0.889	0.888
Naive Bayes		**Cycle 1**	**Cycle 2**	**...**	**Cycle 8**	**Cycle 9**
	precision	0.646	0.69		0.895	0.891
	recall	0.667	0.695		0.889	0.885
	F-measure	0.649	0.68		0.89	0.886

Table 3.5: Cross-validation Results of various classifiers developed in One Step classification framework: LSE oriented CBE

Classifier		**Cycle 1**	**Cycle 2**	**...**	**Cycle 20**	**Cycle 21**
J48	precision	0.457	0.608		0.994	0.994
	recall	0.464	0.614		0.994	0.994
	F-measure	0.456	0.607		0.994	0.994
SVM		**Cycle 1**	**Cycle 2**	**........**	**Cycle 4**	**Cycle 5**
	precision	0.648	0.777		0.88	0.873
	recall	0.643	0.771		0.875	0.867
	F-measure	0.628	0.769		0.875	0.868
Naive Bayes		**Cycle 1**	**Cycle 2**	**...**	**Cycle 18**	**Cycle 19**
	precision	0.589	0.667		0.98	0.977
	recall	0.607	0.671		0.98	0.977
	F-measure	0.595	0.648		0.98	0.977

Table 3.6: Cross-validation Results of various classifiers developed in Two Step classification framework: LSE oriented CBE

Cross validation results for various classifiers developed in Two Step classification framework are depicted in Table 5.1 As SVM captures temporal properties in data very quickly, does not iterate too long. On the other hand both Naive Bayes and J48 iterate for 19 and 21 iterations respectively and hence end up in spreading temporality over more number of synsets. Some of the automatically expanded seed words in this experiment are depicted in Table 4.4.

Temporality statistics: In Tempo-HindiWordNet created with this approach, out of all the words with all possible senses, 1973 are tagged as past, 431 are tagged as present, 4302 are tagged as future, 1977 are tagged as neutral whereas 135183 are tagged as atemporal.

3.2.2 Lexical and Semantic relation encoding based expansion(LSEBE)

To support semi-supervised learning with quality expansion, here we expand training set with the synsets in the test set that are semantically closer to training synsets, instead of depending on classifiers decision as that in CBE. To measure semantic closeness, we use cosine similarity measure given by the equation 3.2. Cosine similarity of LSE of each synset in the test set is computed with LSE of each synset in the training set. Finally similarity of the test synset is its highest similarity found with any of the training synset. Then test

past		present	
word(sense no)	PoS	word(sense no)	PoS
गतवर्ष(gatavarSha; Last year)(1)	adv	अद्यतन(Adyatana;)(1)	adj
भूत_में(bhoot meN; before)(1)	adv	वर्तमान(vartamaana; Present)(1)	n
पहले(pahale; formerly)(1)	adv	इह-काल(Iha-kaala;)(1)	n
अधिरूढ़(AdhirooDdha;)(2)	adj	आधुनिक_काल(AAdhunik kaala; Modern Era)(1)	n
अतीतकाल(Ateetakaala;Past time)(1)	n	प्रौढ़(praoDdha; Mature)(2)	adj
Future		**Neutral**	
word(sense no)	PoS	word(sense no)	PoS
कल(kala; Tomorrow)(2)	adv	निदाघकाल(nidaaghakaala;)(1)	n
आगमनशील(AAgamanasheela;)(1)	adj	सेकेंड(sekeNDa; secs)(3)	n
आगन्तुक(AAgantuka;)(1)	adj	कार्यावकाश(kaaryaavakaasha;)(1)	n
अगवान(Agavaana;)(1)	n	विलंबन(vilaNbana; Postponement)(1)	n
प्रत्यावर्तन(pratyaavartana; Repatriation)(1)	n	घड़ी(ghaDdee; Watch)(1)	n
Atemporal			
word(Sense no)		PoS	
वास्तव(vaastava; Reality)(1)		adj	
प्रणाम_करना(praNNaam karanaa; To greet)(1)		v	
गौरी_शंकर(gaoree shaNkara;)(1)		n	
ठीक_ठाक(Theeka Thaaka; Smoothly)(1)		adv	
थरथर(tharathara;)(1)		adv	

Table 3.7: Few examples of automatically extracted seed words during CBE: LSE oriented CBE

synsets with highest semantic closeness are chosen to achieve expansion.

$$\frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{3.2}$$

LSE of the synset is obtained in the same way as explained in the previous section. As semantically closer instances are added during expansion process in this strategy, we believe that it increases soundness of expansion process and will effectively detect connotative temporal synsets as it incorporates LSE encoded over hypernym and hyponym gloss. In every iteration we add the instances in such a way that the ratio of instances of different classes are maintained at par with the initial class distribution. This is to ensure that the classifier is not biased to any specific class. We define a threshold, and based on this value we select the instances to be added to the initial seed list. This threshold depends on the size of current seed word set, and we fix it through a series of experiments.

Some of the automatically expanded seed words during this expansion strategy are reported in Table 4.5

Temporality Statistics: In Tempo-HindiWordNet created with this approach, out of

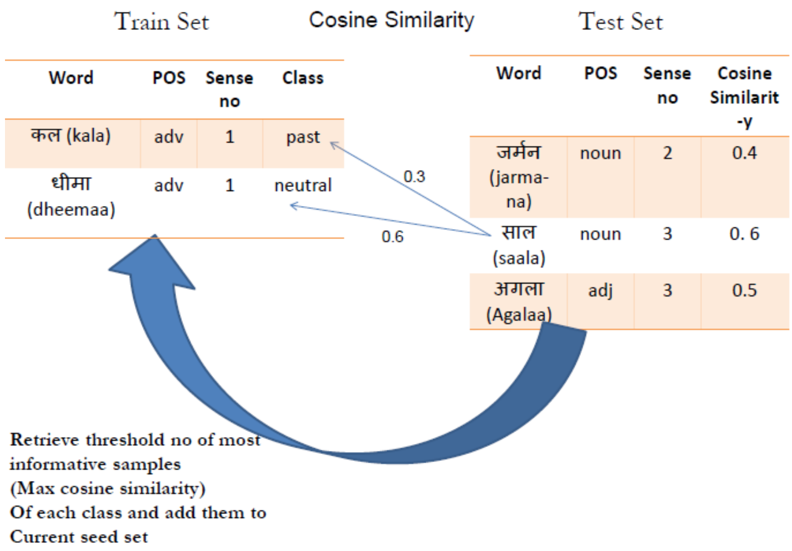

Fig. 3.2: Pictorial representation of LSEBE

all the words with all possible senses, 1088 are tagged as past, 665 are tagged as present, 6319 are tagged as future, 6474 are tagged as neutral whereas 129320 are tagged as atemporal.

Cross-validation Results: Cross-validation results for various classifiers developed with LSEBE strategy in One Step classification framework are reported in Table 4.5. Results shows that J48 and SVM iterate more than that of Naive Bayes. Also, Naive bayes stops with less precision, recall and F-measure as compared to other two classifiers.

Similarly, Cross-validation results for various classifiers learned in Two Step classification framework which are utilizing LSEBE to achieve solid expansion during iterative learning are presented in Table 4.6.

past		present	
word(sense no)	PoS	word(sense no)	PoS
गया(gayaa; went)(1)	adj	सम्प्रति(samprati; Currently)(1)	adv
पिछला(pichhalaa; Last)(3)	adv	इन_दिनों(In dinoN; nowadays)(1)	adv
बीता(beetaa; Past)(1)	adj	फ़िलहाल(Filahaala;)(2)	adv
अवर्तमान(Avartamaana; Non-existent except)(2)	adj	अधुना(1)	adv
व्यतीत(vyateeta; Passed)(1)	adj	आजकल(AAjakala; nowadays)(1)	adv
Future		**Neutral**	
word(sense no)	PoS	word(sense no)	PoS
वाँछित(vaaNnchhita; Required)(1)	adj	जोबन(jobana;)(2)	n
आशंसित(AAshaNsita;)(1)	adj	दिन(dina; day)(8)	n
बाँछित(vaaNnchhita; Required)(1)	adj	पलभर(palabhara;moment)(1)	adv
कमनीय(kamaneeya; Bonny)(1)	adj	दिनावसान(dinaavasaana;)(1)	n
अभिवांछित(AbhivaaNchhita;)(1)	adj	पीछे(peechhe; behind)(7)	adv
Atemporal			
word(Sense no)		PoS	
प्रत्याशानुकूल(pratyaashaanukoola;)(1)		adj	
प्रणाम_करना(praNNaam karanaa; To greet)(1)		v	
पीड़ादायक(peeDdaadaayaka; Sore)(1)		adj	
राष्ट्रीयकरण(raaShTreeyakaraNNa; Nationalization)(1)		n	
निस्सन्देह(nissandeha; Undoubtedly) (1)		adv	

Table 3.8: Few examples of automatically extracted seed words: LSEBE

Classifier		Cycle 1	Cycle 2	...	Cycle 8	Cycle 9
J48	precision	0.601	0.681		0.861	0.856
	recall	0.646	0.703		0.861	0.861
	F-measure	0.611	0.686		0.86	0.858
SVM		Cycle 1	Cycle 2	....	Cycle 8	Cycle 9
	precision	0.719	0.739		0.872	0.858
	recall	0.615	0.678		0.864	0.855
	F-measure	0.635	0.691		0.868	0.856
Naive Bayes		Cycle 1	Cycle 2	Cycle 3	Cycle 4	
	precision	0.646	0.747	0.767	0.759	
	recall	0.667	0.746	0.774	0.746	
	F-measure	0.649	0.743	0.767	0.745	

Table 3.9: Cross-validation Results of various classifiers developed in One Step classification framework: LSEBE

Classifier		Cycle 1	Cycle 2	...	Cycle 8	Cycle 9
J48	precision	0.457	0.698		0.967	0.963
	recall	0.464	0.671		0.967	0.963
	F-measure	0.456	0.676		0.967	0.963
SVM		Cycle 1	Cycle 2	....	Cycle 11	Cycle 12
	precision	0.648	0.765		0.994	0.991
	recall	0.643	0.757		0.994	0.991
	F-measure	0.628	0.76		0.994	0.991
Naive Bayes		Cycle 1	Cycle 2	Cycle 3		
	precision	0.589	0.732	0.722		
	recall	0.607	0.729	0.718		
	F-measure	0.595	0.715	0.702		

Table 3.10: Cross-validation Results of various classifiers developed in Two Step classification framework: LSEBE

Chapter 4

Results Over Gold Standard Test Set & their Analysis

4.1 Results

In order to evaluate tempo-HindiWordNet developed with various techniques, we prepared manually annotated gold standard test set. It is prepared following a thorough discussion among three individual and is annotated by all of them. We then measured muti-rater agreement[Fle71] among these three annotators and found to have moderate agreement of 0.58. This also indicate the difficulty of the task as humans are also not in agreement for lot of decisions. This is mainly due to the fact that these temporal annotations depend largely on the gloss of the synset and not on their inherent concept. E.g.Synset शशिवाटिका(shashivaaTikaa; Shashiwatika) - एक औषधीय पौधा जो दो से तीन मीटर लंबा होता है और हर साल वर्षा ऋतु में निकलता है और गर्मी में सूख जाता है (Ek AOShadheeya paodhaa jo do se teen meeTara laNbaa hotaa hai AOra hara saal varShaa RItu meN nikalataa hai AOra garmee meN sookh jaataa hai;A medicinal plant which is two to three meters long and comes out every year in the rainy season and dries up in summer) allows both atemporal as well as neutral annotation. Finally instances in gold standard test set are annotated with the annotation on which majority of the annotators agree. So, final gold standard test set is consisting of 180 instances, out of which 16 are past, 8 are present, 13 are future, 22 are neutral and 121 are atemporal. We evaluated various techniques developed against this gold standard test set as well as against easy cases in it. Here, by easy cases we mean those synsets against which all annotators are in 100% agreement. Through this, we want to observe quality of our classifiers developed using different techniques against easy cases. We also check agreement of machine annotation with our human annotators. These results in terms of agreement are also reported in this section.

Results in terms of average precision, recall, F-measure for tempo-HindiWordNet developed using propagation strategies discussed in methodology chapter are presented in Tables 4.1 and 4.2. Table 4.1 reports results obtained using One Step classification framework while Table 4.2 reports results of Two Step classification framework. NAN in table 4.2 denotes not a number. In two step classification framework of Gloss oriented CBE classifier does not predict a single instance to be present. Due to this precision of present class is NAN and hence average precision and average F-measure of gloss oriented CBE are NAN.

One Step	**G**loss oriented CBE		**L**SE oriented CBE		**L**SEBE	
Framework	Gold set	Easy cases	Gold set	Easy cases	Gold set	Easy cases
Precision	0.494	0.514	0.644	0.739	0.792	0.812
recall	0.502	0.521	0.601	0.644	0.736	0.821
F-measure	0.498	0.518	0.622	0.700	0.763	0.816

Table 4.1: Results in terms of precision, recall, F-measure for THWn developed with propagation strategies discussed in methodology chapter: One Step classification framework

Two Step	**G**loss oriented CBE		**L**SE oriented CBE		**L**SEBE	
Framework	Gold set	Easy cases	Gold set	Easy cases	Gold set	Easy cases
Precision	NAN	NAN	0.665	0.793	0.65	0.738
recall	0.382	0.562	0.636	0.839	0.581	0.712
F-measure	NAN	NAN	0.650	0.816	0.613	0.725

Table 4.2: Results in terms of precision, recall, F-measure for THWn developed with propagation strategies discussed in methodology chapter: Two Step classification framework

In two step classification framework errors induced in first step(temporal vs. atemporal classification) are propagated to second step(past, present, future, neutral classification).

This is why One step classification has produced good results for gloss orineted CBE and LSEBE. However Two Step classification LSE oriented CBE has succeeded in taking the advantage of classification in two steps. This shows that when LSE is used as feature, classifiers predictions are more effective as comapred to semantic closeness of synsets for achieving expansion. Detailed classwise precision, recall, F-measure results shows that gloss oriented CBE fails to detect Future class in One Step framework and present class in Two Step. Whereas all the classes are detected with moderate precision, recall, F-measure with LSE oriented CBE and LSEBE. Results in Table 4.1 also shows that LSEBE has detected all the classes with highest precision and recall. We present multi-rater agreement results in Table 4.3. Results are reported only for the classifiers(Decision Tree, SVM, Naive Bayes) which achieved highest agreement in respective propagation strategies.

	Gloss oriented CBE		**L**SE oriented CBE		**L**SEBE	
	Gold set	Easy cases	Gold set	Easy cases	Gold set	Easy cases
One Step Classification	0.547	0.83	0.562	0.876	**0**.57	**0**.889
Two Step Classification	0.371	0.747	0.478	0.879	0.462	0.855

Table 4.3: Multi-rater agreement results over manually annotated data for propagation strategies explained in methodology chapter

Highest agreement is achieved using LSEBE in One Step classification framework. With each of the propagation strategy agreement is more on easy cases. This shows that classifiers built with the discussed approaches are performing very well on instances where humans are in 100% agreement. Hence classifiers built are sound. With the help of LSE oriented CBE 2.74% improvement is achieved on One Step classification approach whereas 28.84% of improvement is achieved in Two Step classification framework. LSEBE has achieved an improvement of 4.2% and 24.52% in One Step and Two Step classification frameworks respectively.

4.2 Error Analysis

In this section we report quantitative and qualitative error analysis for THWn developed with propagation strategies discussed in methodology chapter. This error analysis is done with respect to best classifier in corresponding strategy.

Gloss oriented CBE: Confusion matrix showing the possible errors for One Step classification framework is depicted in 4.4. It shows that Gloss oriented CBE could not detect a single instance of future class correctly. Most of these errors are in the instances where future sense is connotative e.g. वासना(vaasanaa; desire)- कुछ पाने की इच्छा या कामना(kuchh paane kee Ichchhaa yaa kaamanaa; Some desire or wish to get something). Instances where explicit future related keywords are present and still miss-classified are mainly due to erroneous instances added during the expansion process. Also lot of instances are falsely predicted to belong to past class. This is mainly due to presence of explicit past related keywords such asहुआ(huAA; Happened), पहले(pahale; formerly, etc. in the gloss of those instances. e.g. आतपशुष्क(AAtapashuShka; Atpsushk) - धूप में सुखाया हुआ(dhoop meN sukhaayaa huAA; dried in sun light is atemporal but due to keyword हुआ(huAA; Happened) is predicted to belong to past.

Classified as ——>	Past	Present	Future	Neutral	Atemporal
Past	9	0	1	5	1
Present	1	3	0	1	3
Future	1	1	0	6	5
Neutral	3	0	1	17	1
Atemporal	8	0	1	1	111

Table 4.4: Confusion matrix for one Step classification framework: Gloss oriented CBE

LSE oriented CBE: We provide quantitative analysis of possible errors in Table 4.5. With this approach false positives for the past class are reduced. Most of the atemporal and neutral which were classified as past with gloss oriented CBE are now correctly detected. Atemporal instances are correctly detected as in two step classification classifiers have to descriminate atemporal synsets only against temporal class in first step. Most of the instances which are miss-classified with this approach are found to be tough even to human annotators as temporal sense in these synsets is not directly denoted. e.g. मरणासन्न(maraNNaasanna; Moribund) - जो मरने के बहुत समीप हो(jo marane ke bahut sameep ho; One who is very close to dying is classified as neutral though it connote futuristic temporal sense.

Classified as ——>	Past	Present	Future	Neutral
Past	9	4	0	3
Present	1	5	1	1
Future	0	0	7	6
Neutral	1	1	2	18

Table 4.5: Confusion matrix for second step of Two Step classification: LSE oriented CBE

LSEBE: Table 4.6 reports confusion matrix for one step classification. With this expansion strategy most of the classes are more accurately predicted. Miss-classified instances are classified wrongly either due to absence of denotative temporal words of that temporal category or they fall in difficult to classify cases i.e. even human annotators are not in perfect agreement while classifying them.

Classified as ——>	Past	Present	Future	Neutral	Atemporal
Past	9	0	1	4	2
Present	0	4	1	0	3
Future	0	0	5	2	6
Neutral	0	0	4	15	3
Atemporal	1	0	2	5	113

Table 4.6: Confusion matrix for One Step classification: LSEBE

Chapter 5

Sentence Level Temporality Detection

We show sentence level temporality detection as one of the application of our resource. For this task we developed annotated corpora of 940 sentences manually tagged with past, present and future. Owing to confusion arising in annotating some of the sentences, we annotated these corpora with 3 human annotators and measured muti-rater agreement[Fle71]. It is found to have substantial agreement of 0.78. Finally sentences are tagged with the tag on which majority of human annotators agree. Out of 940 sentences in this corpora, 281 are belonging to past, 533 are belonging to present and 126 are tagged as future. Some of the sentences in this corpora and their taggs are as shown below.

1. दो साल पहले निर्माता जवाहर एल जयरथ ने इस प्रॉजेक्ट पर काम शुरू किया(do saal pahale nirmaataa javaahara El jayarath ne Is prawjekT para kaam shuroo kiyaa; Two years ago, began work on the project by producer Jawahar L. Jairath) - Past

2. इसी साल इन दोनों परियोजनाओं का काम देखने के लिए मूंदड़ा पोर्ट एंड स्पेशल इकनॉमिक जोन लिमिटेड का गठन हुआ(Isee saal In donoN pariyojanaaON kaa kaam dekhane ke liE mooNdaDdaa porT END speshal Ikanawmik jon limiTeD kaa gaThan huAA; This year Mundra Port and Special Economic Zone Ltd. are set up for keeping an eye on the work) - Present

3. यह कमिटी अगले 6 महीने में बदले हुए हालात को देखते हुए दिल्ली नगर निगम के नए ऐक्ट का ड्राफ्ट तैयार करेगी(yah kamiTee Agale 6 maheene meN badale huE haalaat ko dekhate huE dillee nagara nigam ke naE AIkT kaa DraaphT taiyaara karegee; This committee, in the next 6 months , given the changed circumstances of the MCD will draft the new Act) - Future

Rule Based Approach for sentence level temporality detection

Our manually annotated corpora consists of only past, present and future class sentences. Hence we don't consider words detected to be neutral or atemporal with the help of developed THWn for determining sentence level temporality. We have used rule based approach for sentence level temporality detection. Rules are on the basis of number of past, present, future related words detected with the help of THWn in the sentence. Rules are as follows: If maximum words in sentence are detected to belong to temporal category t then it is annotated with t. If sentence is detected with words of more than one temporal category and they are equal in number for each category then

- if words of past and present category are detected then annotate class 'present'.
- if words of present and future category are detected then annotate class 'future'.
- if words of past and future category are detected then annotate class 'future'.
- if words of all temporal categories are detected then annotate class 'future'.

If no temporal words are detected in the sentence then it is annotated randomly. Words in the corpora are disambiguated with their most frequent sense computed using unsupervised algorithm[RRB]. Results over this task are depicted in Table 5.1.

Results shows that THWn developed with LSEBE has produced highest accuracy of 66.4% over sentence temporality detection task. Accuracy results depicted in the table don't consider random annotation done when our resource fail to detect temporal words in the sentence. With LSE oriented CBE and LSEBE, Two Step classification has excelled as these approaches makes first step of Two Step classification more sound as compared to

	Gloss oriented CBE		**w**ord embedding oriented CBE		**w**ord embedding based expansion	
	One Step	Two Step	One Step	Two Step	One Step	Two Step
% Accuracy	61.9	54.5	50.5	59.5	42.9	**66.4**
Precision	0.540	0.368	0.473	0.561	0.589	**0.649**
recall	0.554	0.374	0.407	0.577	0.535	**0.679**
F-measure	0.55	0.37	0.44	0.57	0.56	**0.66**

Table 5.1: Results of Sentence level temporality detection with THWn developed using various propagation strategies

gloss oriented CBE. Hence less errors propagates to second step. In one step classification as number of iterations goes on increasing number of atemporal instances in training set increases. This causes training set to become skewed towards atemporality. Two Step classification do not have to deal with such biasing. This is also one of the reason behind success of Two Step approach over this task. However gloss oriented approach fails to take its advantage as error propagation from first step to second step nullifies this advantage in gloss oriented CBE.

Chapter 6

Conclusion and Future Work

In this thesis we have presented various techniques to automatically construct a temporal resource for Hindi language that could be effective for temporal information access. We have proposed different models based on semi-supervised learning framework that learns iteratively starting from a small set of seed entities. As a learning algorithm we experimented with SVM, Naive Bayes and Decision Tree trained over glosses and/or word embeddings of synsets. We have presented an exhaustive evaluation framework where we report cross-validation accuracies as well as accuracies on a manually created gold standard test set. We also show sentence level temporality detection as an application of our resource. We reported accuracies of sentence temporality detection achieved with different techniques.

We believe that our contribution towards building the temporal resource in Hindi will be an useful resource to the community, and will facilitate the research related to NLP and IR applications.

In future we would like to identify more features for the target task. In addition, we would like to explore different expansion strategies such as combination of confidence based expansion and Lexical and semantic relation encoding based expansion. We would also like to investigate whether deep learning techniques can improve accuracies of the task.

Appendix A

Hindi WordNet API

Hindi WordNet is construction is still on its way. Presently, it is consisting of 28687 synsets and 63800 unique words in its 1.0 version. Information in Hindi WordNet can be accessed using its API. Hindi WordNet API is written in JAVA and all in information is stored in Java Hindi WordNet Library(JHWNL) library. Important classes in Hindi WordNet and their methods are explained in detail in following sections.

- **Dictionary** Dictionary class is mainly used to fetch words, meanings, semantic and lexical relations in Hindi WordNet(HWn). It mainly comprises of following methods-

  - getInstance()

    It instantiate the object of Dictionary class

  - getSynsetAt(POS pos, long synset-Id)

    Returns Synset with the given synset-Id and pos category

  - getAdjWordList()

    Return all the words of adjective POS category

  - getAdvWordList()

    Return all the words of adverb POS category

  - getVerbWordList()

    Return all the words of Verb POS category

- getNounWordList()

  Return all the words of Noun POS category

- getIndexWord(POS pos, java.lang.String lemma)

  Returns an IndexWord with the specified POS containing the lemma as a word.

  E.g.

  * Input : गढ़(gaDdha;)
  * Output:1. गढ़ना(gaDdhanaa; Concoct)(POS = VERB) 2. गढ़(gaDdha; Stronghold(POS= NOUN)

- getOntoSynset(long onto-Id)

  Return the Ontology Hierarchy from the given Synset -Id

  E.g.

  * Input : गढ़ना(gaDdhanaa; Concoct) sense 1 i.e. 2585 - NOUN - [अड्डा, गढ़, केंद्र, केन्द्र](ADDaa, gaDdh, keNdra, kendra; Den , defenses , center , center)
  * Output : ONTO_NODES : भौतिक स्थान(bhaotik sthaana; Physical place) (Physical Place) (PHSCL उदाहरण:- पाठशाला,पहाड़,बैंक इत्यादि (e.g.:-paaThashaalaa,pahaaDd Ityaadi; School , mountains , bank, etc.)); स्थान (Place) (PLACE उदाहरण:- मैदान,घर,विद्यालय इत्यादि(UdaaharaNNa:-maidaan,ghara,vidyaalaya Ityaadi; Field , home , school, etc.)); निर्जीव (Inanimate) (INANI उदाहरण:- पुस्तक,घर,धूप इत्यादि(UdaaharaNNa:- pustak,ghara,dhoop Ityaadi; Example: - books, home , sun , etc.)); संज्ञा (Noun) (N उदाहरण :- गाय,दूध,मिठाई इत्यादि(UdaaharaNN :- gaaya,doodh,miThaaEE Ityaadi ; Example: - cows , milk , sweets , etc.)); TOP (Top Level Node)

- lookupAllIndexWords(java.lang.String lemma)

  Returns a set of IndexWords, with each element in the set corresponding to all POS of the lemma in which synsets are present.

  E.g.

    * Input : गढ़ना(gaDdha;)
    * Output : 1. गढ़ (gaDdha;) 2. गढ़ना (gaDdhanaa; Concoct)
  - lookupMorphedIndexWords(POS pos, java.lang.String lemma)

    Returns a set of IndexWord for all root forms of the lemma for the specified POS
  - lookupAllMorphedIndexWords(java.lang.String lemma)

    Returns a set of IndexWord for all root forms of the lemma for all POS in which synsets are present
  - lookupIndexWord(POS pos, java.lang.String lemma)

    Returns all Synsets with the specified POS containing the root form of lemma as a word. Morphed forms of words can be supplied to this method.

    E.g.
    * Input : गढ़ना(gaDdhanaa; Concoct)
    * output : गढ़ (POS= NOUN)(gaDdha;)

- **Pointer**

  It is mainly used to retrieve pointers to lexical and semantic relations.
  - getOntoPointer()

    Returns pointer to Ontology nodes.
  - GetOntoNodes()

    Get the Ontology nodes for this pointer.
  - getTargetSynset()

    Get the synset that is the target of this pointer.

- **Synset**

  It is mainly used to retrieve all information related to synset such as POS category, number of words in it, word at a specific word in it, etc. Main methods in it are depicted below

- containsWord(java.lang.String lemma)

  Returns wheather Synset contains lemma in it or not.

- getGloss()

  Returns gloss of the synset

- getOffset()

  Returns unique id associated with the synset.

- GetPointers()

  Return an array of pointers for this Synset, which can be used to access all relations of the synset

- getPOS()

  Returns POS category of the synset

- getWord(int index)

  Returns Word at given index in the synonymous set(Synset)

- getWords()

  Returns all the words in synonymous set(Synset)

- GetWordsSize()

  Returns number of words in synonymous set

• **IndexWord**

All words in Hindi WordNet(HW) are indexed. All the information related to words in HW can be obtained with the help of this class. Main functionalities provided by this class are summarized below

- getLemma()

  Returns root form of the indexed word

- getPOS()

  Returns POS category of indexed word

- getSense(int index)

  Returns synset at a given index where index is sense number that you want to retrieve

- getSenseCount()

  Returns number of possible senses with which word can be used

- getSenses()

  Returns synset array

- **IndexWordSet**

  This class is mainly used to store array of index words. Main functionalities provided by it are summarized below.

  - getIndexWord(POS pos)

    Returns IndexWord of given POS category

  - GetIndexWordArray()

    Returns IndexWord Array

Appendix B

Precision, Recall and F-measure

- **Precision**

  Precision is defined as the ratio of total number of instances of a class correctly classified by the classifier(true positives) to total number of instances classified among that class. In general terms it is basically is a measure which tells how many instances, out of total no of instances classified among a perticular class, are correctly classified. It is given by equation B.1

  $$\frac{tp}{tp + fp} \tag{B.1}$$

  where,

  tp = Instances belonging to class i and classified among the same class(true positive)

  fp = Instances not belonging to class i but classified among class i(false positive)

- **Recall**

  It is defined as the ratio of total number of instances of a class correctly classified by the classifier to total number of insatnces belonging to that class. In general terms it is a measure which tells how many instances, out of total number of instances belonging to the certain class, are correctly classified. It is given by the equation B.2

  $$\frac{tp}{tp + fn} \tag{B.2}$$

where,

tp = Instances belonging to class i and classified among the same class(true positive)

fn = Instances belonging to class i but classified among some other class(false negative)

- **F-measure**

It is defined as harmonic mean of Precision and Recall. Sometimes it happens that precision is too low and recall is too high or vice varsa. In such situation it becomes tough to say whether classifier has perform good or bad only on the basis of precision and recall. Hence, these two measures are combined to produce only one figure on the basis of which classifier performance can be judged. Instead of taking average of precision and recall, harmonic mean is taken as harmonic mean of recall and precision is closer to whichever is smaller out of them. It is given by the equation B.3

$$\frac{2 * Recall * Fmeasure}{Recall + Fmeasure} \tag{B.3}$$

References

[ABBS12] A. Anand, S. Bedathur, K. Berberich, and R. Schenkel. Index maintenance for time-travel text search. In *Proceedings of the 35th International ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pages 235–244, 2012.

[AGBY09] O. Alonso, M. Gertz, and R. Baeza-Yates. Clustering and exploring search results using timeline constructions. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM)*, pages 97–106. ACM, 2009.

[AGK01] James Allan, Rahul Gupta, and Vikas Khandelwal. Temporal summaries of new topics. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 10–18. ACM, 2001.

[BB14] Robin Burghartz and Klaus Berberich. Mpi-inf at the ntcir-11 temporal query classification task. In *NTCIR*, 2014.

[BDR+14] Ondrej Bojar, Vojtech Diatka, Pavel Rychlỳ, Pavel Stranák, Vít Suchomel, Ales Tamchyna, and Daniel Zeman. Hindencorp-hindi-english and hindi-only corpus for machine translation. In *LREC*, pages 3550–3555, 2014.

[BM07] Steven Bethard and James H Martin. Cu-tmp: Temporal relation classification using syntactic and semantic features. In *Proceedings of the 4th International*

Workshop on Semantic Evaluations, pages 129–132. Association for Computational Linguistics, 2007.

[CDJ11] Ricardo Campos, Gaël Dias, and Alípio Jorge. An exploratory study on the impact of temporal features on the classification and clustering of future-related web documents. In *Progress in Artificial Intelligence*, pages 581–596. Springer, 2011.

[CDJJ14] Ricardo Campos, Gaël Dias, Alípio M. Jorge, and Adam Jatowt. Survey of temporal information retrieval and related applications. *ACM Computing Survey*, 47(2):15:1–15:41, 2014.

[Col81] Max Coltheart. The mrc psycholinguistic database. *The Quarterly Journal of Experimental Psychology*, 33(4):497–505, 1981.

[DHFM14] Gaël Dias, Mohammed Hasanuzzaman, Stéphane Ferrari, and Yann Mathet. Tempowordnet for sentence time tagging. In *Companion Publication of the 23rd International Conference on World Wide Web Companion (WWW)*, pages 833–838, 2014.

[ES05] Andrea Esuli and Fabrizio Sebastiani. Determining the semantic orientation of terms through gloss classification. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 617–624. ACM, 2005.

[Fle71] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[FN14] Michele Filannino and Goran Nenadic. Using machine learning to predict temporal orientation of search engines' queries in the temporalia challenge. In *NTCIR*, 2014.

[Gee12] J Geertzen. Inter-rater agreement with multiple raters and variables. *URL:, Retrieved May*, 8:2014, 2012.

[HDF14] Mohammed Hasanuzzaman, Gaël Dias, and Stéphane Ferrari. Hultech at the ntcir-11 temporalia task: Ensemble learning for temporal query intent classification. In *The 11th NTCIR Conference on Evaluation of Information Access Technologies*, pages p–478, 2014.

[HDFM14] Mohammed Hasanuzzaman, Gaël Dias, Stéphane Ferrari, and Yann Mathet. Propagation strategies for building temporal ontologies. In *14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 6–11, 2014.

[HSG07] Mark Hepple, Andrea Setzer, and Rob Gaizauskas. Usfd: preliminary exploration of features and classifiers for the tempeval-2007 tasks. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 438–441. Association for Computational Linguistics, 2007.

[HTX+14] Yongshuai Hou, Cong Tan, Jun Xu, Youcheng Pan, Qingcai Chen, and Xiaolong Wang. Hitsz-icrc at ntcir-11 temporalia task. In *NTCIR*, 2014.

[JJB+14] Hideo Joho, Adam Jatowt, Roi Blanco, Hajime Naka, and Shuhei Yamamoto. Overview of ntcir-11 temporal information access (temporalia) task. In *NTCIR-11 Conference (NTCIR)*, pages 429–437, 2014.

[JJR13] Hideo Joho, Adam Jatowt, and Blanco Roi. A survey of temporal web search experience. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 1101–1108. International World Wide Web Conferences Steering Committee, 2013.

[JL95] George H. John and Pat Langley. Estimating continuous distributions in bayesian classifiers. In *Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345, San Mateo, 1995. Morgan Kaufmann.

[Joa02] Thorsten Joachims. *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms.* Kluwer Academic Publisher, 2002.

[JYT13] A. Jatowt, C.-M. Au Yeung, and K. Tanaka. Estimating document focus time. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management (CIKM)*, pages 2273–2278, 2013.

[KBM11] N. Kanhabua, R. Blanco, and M. Matthews. Ranking related news predictions. In *Proceedings of the 34th International ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pages 755–764, 2011.

[KTSD11] A. Kulkarni, J. Teevan, K.M. Svore, and S. Dumais. Understanding temporal query dynamics. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM)*, pages 167–176, 2011.

[MCCD13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[Met07] M.J. Metzger. Making sense of credibility on the web: Models for evaluating online information and recommendations for future research. *Journal of the American Society for Information Science and Technology*, 58(13):2078–2091, 2007.

[Mil95] Georges A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[MJPZ09] Donald Metzler, Rosie Jones, Fuchun Peng, and Ruiqiang Zhang. Improving search relevance for implicitly temporal queries. In *Proceedings of the 32nd* *international ACM SIGIR conference on Research and development in information retrieval*, pages 700–701. ACM, 2009.

[MPG05] Inderjeet Mani, James Pustejovsky, and Robert Gaizauskas. *The language of time: a reader*, volume 126. Oxford University Press, 2005.

[MSF07] Congmin Min, Munirathnam Srikanth, and Abraham Fowler. Lcc-te: a hybrid approach to temporal relation identification in news text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 219–222. Association for Computational Linguistics, 2007.

[NRD08] Sérgio Nunes, Cristina Ribeiro, and Gabriel David. Use of temporal expressions in web search. In *Advances in Information Retrieval*, pages 580–584. Springer, 2008.

[ÖSMP11] Gözde Özbal, Carlo Strapparava, Rada Mihalcea, and Daniele Pighin. A comparison of unsupervised methods to associate colors with words. In *Affective Computing and Intelligent Interaction*, pages 42–51. Springer, 2011.

[Qui93] Ross Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[RM15] Nitin Ramrakhiyani and Prasenjit Majumder. Approaches to temporal expression recognition in hindi. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 14(1):2, 2015.

[RRB] SudhaBhingardive DhirendraSingh RudraMurthyV, Hanumant Redkar, and Pushpak Bhattacharyya. Unsupervised most frequent sense detection using word embeddings.

[SADCK06] Steven Schockaert, David Ahn, Martine De Cock, and Etienne E Kerre. Question answering with imperfect temporal information. In *Flexible Query Answering Systems*, pages 647–658. Springer, 2006.

[SSM14] Abhishek Shah, Dharak Shah, and Prasenjit Majumder. Andd7@ ntcir-11 temporal information access task. In *NTCIR*, 2014.

[UA10] N. UzZaman and J.F. Allen. Trips and trios system for tempeval-2: Extracting temporal information from text. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 276–283, 2010.

[VGS+07a] M. Verhagen, R. Gaizauskas, F. Schilder, M. Hepple, G. Katz, and J. Pustejovsky. Semeval-2007 task 15: Tempeval temporal relation identification. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 75–80, 2007.

[VGS+07b] Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. Semeval-2007 task 15: Tempeval temporal relation identification. In *4th International Workshop on Semantic Evaluations (SEMEVAL)*, 2007.

[VSCP10] Marc Verhagen, Roser Saurí, Tommaso Caselli, and James Pustejovsky. Semeval-2010 task 13: Tempeval-2. In *5th International Workshop on Semantic Evaluation (SEMEVAL)*, pages 57–62, 2010.

[YKR14] Haitao Yu, Xin Kang, and Fuji Ren. Tuta1 at the ntcir-11 temporalia task. 2014.