Sentence Level Temporality Detection using an Implicit Time-sensed Resource

Sabyasachi Kamila, Asif Ekbal, Pushpak Bhattacharyya
Department of Computer Science and Engineering, IIT Patna
Patna, India
{sabysachi.pcs16,asif,pb}@iitp.ac.in

Abstract

Temporal sense detection of any word is an important aspect for detecting temporality at the sentence level. In this paper, at first, we build a temporal resource based on a semi-supervised learning approach where each Hindi-WordNet synset is classified into one of the five classes, namely *past*, *present*, *future*, *neutral* and *atemporal*. This resource is then utilized for tagging the sentences with *past*, *present* and *future* temporal senses. For the sentence-level tagging, we use a rule-based as well as a machine learning-based approach. We provide detailed analysis along with necessary resources.

Keywords: Temporal Sense Detection, Semi-supervised Machine Learning, Sentence Level Temporality Detection

1. Introduction

Over the last few years, 'temporality' has drawn a significant attention to the community of Natural Language Processing (NLP) and Information Retrieval (IR). Time is an intrinsic property that aids in ordering events in a sequential order from the past to present to future. This ordering of events is very crucial in analyzing a document. Some of the applications where temporality plays an important role include automatic summaries (Allan et al., 2001), question-answering (Schockaert et al., 2006), clustering (Alonso et al., 2009), similarity of documents (Jatowt et al., 2013) etc. Queries can specify temporality both explicitly and implicitly. Queries like *"World Cup 2011"*, *"Indian Prime Minister 2000"* etc. denote the temporality explicitly, whereas the queries like *"Chomsky's childhood"*, *"Recent Bollywood songs"* etc. correspond to implicit temporality. All these highlight the significance of time in refining and ranking the results retrieved from a search engine.

In a survey, Joho et al. (2013) claimed that most of the time user queries need to be addressed with *recent* information. However, many situations demand the *past* or *future* related information. For example, the query *"ग्लोबल वार्मिंग की वर्तमान स्थिति (globala vArmiMga kI vartamAna sthiti - The current status of global warming.)"*[1] requires *present* related information whereas the queries like "डिजिटल अर्थव्यवस्था में भारत के लिए अवसर क्या है? *(DijiTala arthavyavasthA meM bhArata ke lie avasara kyA hai?- What are the opportunities for India in the Digital Economy?)"*, "अशोक का इतिहास *(ashoka kA itihAsa- History of Ashoka)"* need *future* and *past* related information, respectively. Here, tense related information does not help but the implicit temporal keywords *'current'*, *'opportunities'* and *'history'* help in finding the temporal information of the respective queries.

1.1. Motivation and Problem Definition

Most of the earlier studies, for example, TempEval tasks (Verhagen et al., 2009; Verhagen et al., 2010; UzZaman et al., 2013) in the computational linguistics, have concentrated on identifying the temporal expressions, event expressions and various relations among these. These studies tried to address the temporal aspects of information with the help of linguistic constructs such as the presence of temporal expressions like before, now, after etc., document creation time (DCT), or explicit time expressions.

Let us consider the following two example sentences: **Sentence-I**: *You should live in the present.* **Sentence-II**: *She gave him a nice present.* When these two sentences are subjected as input to the SUTime tagger,[2] we observe that, for both the sentences, the word 'present'[3] is tagged as a temporal expression. However, it should be temporal only for the first sentence. When these two sentences are subjected as input to the HeidelTime tagger,[4] no temporal mention is found in either of the sentences.

In Hindi-WordNet (Bhattacharyya, 2010), the word *"कल (kala)"* has 8 senses in total, both *temporal* and *atemporal*. Let us consider the following three example sentences: **Sentence-I**: *"यह लेख कल के अखबार में है (yaha lekha kala ke akhabAra meM hai-This article is in yesterday's newspaper.)"*; **Sentence-II**: *"कल कौनसी परीक्षा है? (kala kaunasI parIkShA hai?- Which examination is scheduled tomorrow?)"*; **Sentence-III**: *"नये-नये कल का निर्माण हो रहा है (naye-naye kala kA nirmANa ho rahA hai-New machines are being built)"*. Here, the same word *"कल (kala)"* corresponds to *'yesterday'*, *'tomorrow'* and *'machines'* in the first, second and third sentence, respectively. These denote the *past* and *future* senses in the first and second sentence, respectively, and *atemporal* in the third sentence. Unless the contextual information is taken into account these cannot be disambiguated appropriately.

It is evident from the existing literature that there is a lack of attention in detecting the implicit temporal sense of words. In order to capture such implicit temporal senses, we propose an effective technique for determining the temporal sense of each synset of the

[1] Henceforth, all the Hindi examples are represented by Hindi glosses, ITRANS representations and using equivalent English translation.

[2] `http://nlp.stanford.edu/software/sutime.html`

[3] The word 'present' has noun Part-of-Speech (PoS) tag in both the sentences.

[4] `http://heideltime.ifi.uni-heidelberg.de/heideltime/`

Hindi-WordNet. We augment each synset of the Hindi-WordNet with one of the five temporal tags, namely *past*, *present*, *future*, *neutral*, and *atemporal*. For example, the synset *"प्राचीन (prAchIna-Ancient)"*, *"मौजूदा (maujUdA-Existing)"* and *"आगामी (AgAmI-Forthcoming)"* correspond to *past*, *present* and *future* time sense, respectively. The synset *"अयोग्य (ayogya-Unworthy)"* is characterized as *'atemporal'* as it does not depict any time sense. There are also some synsets, such as *"सुबह (subaha-Morning)"* that clearly represent a time sense, but cannot be specifically categorized to *past*, *present* or *future*. Such kind of instances are denoted as *neutral*.

At first, we propose a semi-supervised machine learning framework for detecting temporal word senses. The process initiates learning with a set of seed instances for each class, and then iteratively expands it following various expansion strategies. The temporal resource, Tempo-Hindi-WordNet that we build will definitely be an effective resource for the efficient temporal information access in the resource-poor languages like Hindi which is one of the widely spoken languages worldwide and one of the official languages in India. We show how this resource can be utilized for sentence-level temporal tagging.

Our present study is inspired from the prior works (Dias et al., 2014; Hasanuzzaman et al., 2014), where the authors attempted to annotate each synset of English WordNet with four temporal dimensions, namely *past*, *present*, *future* and *atemporal*. Our work differs from these existing works in terms of the following points: (i). present work attempts to build a temporal resource that can facilitate temporal information access in Hindi; (ii). new expansion strategies including word-embedding based techniques are proposed; and (iii). two approaches (i.e. rule-based and machine learning-based) for sentence-level temporality detection are developed. The present work also differs from an earlier work reported in (Pawar et al., 2016) in terms of expansion strategies, quality of temporal resource created, and application of the resource developed for sentence-level temporality detection in two different domains, *viz.* newswire and Twitter.

2. Word-level Temporal Sense Detection

Due to the unavailability of annotated dataset, we adapt a semi-supervised learning strategy for temporal word sense detection.

2.1. Seed Data Creation

We manually prepare a seed set based on the synsets of Hindi-WordNet. Three individuals (with post-graduate level knowledge) were asked to annotate the seed set based on the word knowledge and the information available in the gloss, and it was found to have a substantial multi-rater kappa agreement (Fleiss, 1971) of 0.73 among the annotators. The tag was finalized based on majority voting. The seed consists of 96 synsets, out of which 48 are *atemporal* and the rest are equally distributed among the *past*, *present*, *future* and *neutral*. While creating this, special care was taken to ensure that it is not biased towards any specific temporal class or Part-of-Speech (PoS) category. It is to be noted that in the Hindi-WordNet *samay*-time is biased towards the 'noun' PoS category.

2.2. Gold Standard Set for Evaluation

In order to evaluate the Tempo-Hindi-WordNet, we manually prepare a gold standard test set with the synsets taken from the Hindi-WordNet.[5] Same persons who created the seed set were employed for this annotation with the help of similar kind of information. Multi-rater kappa agreement was found to be 0.63 which gives an idea about the level of difficulty involved, as humans are also not in agreement for a number of decisions. One of the considerations was the fundamental fact that the core concepts of words do not exist in many cases; these are rather defined by the contextual information. For example, the synset "आवश्यकता (AvashyakatA-requirement) - आवश्यक होने की अवस्था या भाव (Avashyaka hone kI avasthA yA bhAva-State of necessity) has a connotative sense of *future*. However, from the inspection to the WordNet gloss, it was found not to have any time sense. For a second example, the synset "इमरजेंसी (imarajeMsI - Emergency)- संकट या विपत्ति का समय (saMkaTa yA vipatti kA samaya-Time of crisis or disaster)" describes a situation or condition of emergency where we can call something *"emergency"* by looking at its effects in the recent *past* or *present*. We cannot surely confirm a situation to be emergent that has not yet happened. Hence, it can have both the *past* and *present* time senses.

Many idiomatic synsets such as "धूप-छाँह *(dhUpa-ChA.Nha)*- बारी-बारी से आने वाला अच्छा और बुरा समय *(bArI-bArI se Ane vAlA achChA aura burA samaya - ups and downs of life where good and bad times come alternately)"* etc. are very difficult to annotate properly. From the meaning represented in the gloss we can conclude that the synset signifies a time period (denoting neutral) or a state of life (denoting atemporal). However, majority agree it to be of atemporal type. Finally, instances of gold standard are annotated based on the majority agreement. The gold standard set finally contains 180 instances: 16 past, 8 present, 13 future, 22 neutral and 121 atemporal.

2.3. Framework

We propose a hierarchical classification framework for solving the problem. In the first level, we distinguish temporal vs. atemporal. In the second level, we classify temporal instances into *past*, *present*, *future* and *neutral* categories. Initial set of seed set is iteratively expanded using various expansion strategies. Steps of the algorithm are shown in Algorithm 1.

2.4. Expansion Strategies

We propose two expansion strategies: *"Confidence based Expansion (CBE)"* and *"Semantic Distance based Expansion (SDE)"*.

1. In the first model (i.e. CBE), we select the most informative instances based on the prediction confidence of the classifier.

[5] While creating models, we excluded these instances from the Hindi-WordNet for processing.

Algorithm 1 Basic steps of temporal resource creation

1: Select initial set of seed words.
2: **repeat**
3: Train the model on the training instances created from the seed set.
4: Evaluate the model developed on the rest of the synsets of Hindi-WordNet (in an incremental manner).
5: Expand the seed set according to the chosen expansion strategy.
6: **until** cross-validation accuracy drops.
7: Classify the Hindi-WordNet using the final trained model.

2. In our second model (i.e. SDE), we select the most informative instances based on the semantic distance rather than the classifier's confidence.

Prototype Vector Generation: For representing the instances, we create a 'prototype vector' from the glosses of synset, synonyms, hyponyms and hypernyms. We assume that the temporal senses of a synset are propagated through various semantic relations of the Hindi-WordNet, and hence the information represented in their glosses will provide an important evidence. Vectors of two semantically related synsets could assist in spreading or detecting connotative temporality. Such a vector representation should be able to quickly refine the classifier's decision boundaries.

For example, semantic relations such as hypernym and hyponym detect connotative temporality as hypernym is a generalization and hyponym is a specialization of the synset. For example, *"विराम_काल (virAma kAla-rest period)"* is the kind of *"काल (kAla-period)"*. Here, *"काल (kAla-period)"* is the hypernym and *"विराम_काल (virAma kAla- rest period)"* is a hyponym. Both of these indicate temporality. As we encode both hyponyms and hypernyms, one's presence ensures other's inclusion through expansion.

We use Word2vec tool (Mikolov et al., 2013) for generating word embedding vectors. The model is trained on Bojar's corpus (Bojar et al., 2014) of around 44 million Hindi sentences for the training of Word2Vec using Skip-gram model with the dimension set to 200 and window size set to 7. For each content word of the synset, hyponyms, hypernyms and their glosses, we extract the corresponding vector of 200 dimension. All these vectors are averaged over to create a 'prototype vector'. If there are m content words then the prototype vector is generated as shown in the Equation 1.

$$\frac{\sum_{i=1}^{m} WE(w_i)}{m} \tag{1}$$

where, m is the number of content words in the glosses of synset, synonyms, hypernyms and hyponyms; WE(w_i) is the word embedding vector of the i^{th} token.

2.4.1. Confidence based expansion (CBE)

This expansion strategy makes use of the classifier's confidence as a mean to expand the initial seed list. Higher the value of confidence, more is the chance of its belongingness in the expanded list. We use three classification algorithms, namely Support Vector Machine (SVM) (Joachims, 2002), Naive Bayes (NB) (John and Langley, 1995) and Decision Tree (DT) (Quinlan, 1993).

Each classifier is trained with the feature vectors generated from the initial seed instances and tested with the rest of the Hindi-WordNet synsets. All the training and test instances are represented by *'prototype vectors'*. Instances which are predicted with higher confidence are considered to be the useful samples. Such instances are given more priority during expansion so as to preserve the connotational properties of the initial seed entities intact. In every iteration, we add the instances to the training set in such a way that the ratio of the instances of different classes are maintained at par with the initial class distribution. For expansion, we execute the following steps: (i). for each class, we select the top-10 instances from the test set based on the classifier's confidence; (ii). exclude such instances from the test set for evaluation in the next iteration; (iii). add the selected instances to the initial training set; and (iv). perform 10-fold cross validation experiment on the expanded training set. This process is repeated in an iterative fashion. If the cross-validation accuracy does not increase in the consecutive 3 iterations, then we terminate the process. Finally, we select the model that shows the best performance during all the iterations. This final selected model is used for classifying the entire Hindi-WordNet.

2.4.2. Semantic distance based expansion (SDE)

In this method, we expand the seed set in such a way that the newly added instances are always semantically closer to the existing seed instances. Unlike CBE, this method does not depend on the classifiers' decisions, rather it relies on the semantic distance between the two vectors. All the training and test instances are represented by *'prototype vectors'*. For each test instance, we measure its distance from all the training vectors by computing the cosine similarities. We choose the new candidate instances to be added to the initial training set in such a way that: (i). added instances are closer to the existing seed entities; and (ii). rejected instances seem to be dissimilar to the existing ones.

We select 10 most similar instances (based on cosine similarity) for each class from the test set, and add them to the training set in each iteration. We exclude such instances from the test set in the next iteration. We stop iterating when the cross-validation accuracy does not increase in the consecutive three iterations.Like the CBE based model, we finally fix the model that produces the highest cross-validation accuracy.

Through this process we ensure that in every run, a good quality of new instances are added to the existing training set. As the semantically closer instances are added during the expansion process, we believe that it preserves the soundness property. The process is more effective in detecting connotative temporal properties of the data as we expand our knowledge base by inducing word-embedding vectors and other WordNet semantic relations.

2.5. Results and Analysis

In this section we report the experimental results along necessary analysis.

2.5.1. Experiments: Cross-validation

Results of 10-fold cross-validation are reported in Table 1 for the hierarchical SDE based classification approach. It is to be noted that this model quickly converges but still attains better accuracy. It shows precision, recall and F-measure values of 85.53%, 89.66% and 87.55%, respectively for SVM. The SVM classifier performs better than DT and NB. Tempo-Hindi-WordNet that we obtain at the end contains 1,572 past, 3,650 present, 2,822 future, 5,429 neutral and 130,413 atemporal instances.

Iteration		1	2	...	9	10
DT	precision	61.80	70.56		78.89	78.85
	recall	65.89	72.45		78.91	78.81
	F-measure	63.78	71.49		78.90	78.83
NB	Iteration	1	2	....	9	10
	precision	62.23	67.28		73.78	74.20
	recall	64.88	68.03		76.92	74.10
	F-measure	63.53	67.65		75.32	74.15
SVM	Iteration	1	2	....	12	13
	precision	72.87	73.76		84.67	85.53
	recall	62.56	63.78		89.23	89.66
	F-measure	67.32	68.40		86.89	87.55

Table 1: Iteration-wise 10-fold cross-validation results. Here, we report the average performance of all the classes for the SDE based approach.

For CBE based expansion technique, experiments on 10-fold cross-validation yield precision, recall and F-measure values of 86.88%, 85.32% and 86.14%, respectively. In both the cases, SVM performs better-may be due to its robustness in efficiently handling high dimensional feature space. Finally, we observe the following statistics of the Tempo-Hindi-WordNet using CBE: 973 past, 431 present, 4,302 future, 1,977 neutral and 135,183 atemporal instances.

2.5.2. Experiments: Gold Standard

We evaluate our various models on the gold standard test set. We also show the evaluation on easy-to-classify instances.[6] The basis of showing this evaluation is to demonstrate how our classifier performs with respect to the humans. We report the results using all the expansion strategies in Table 2 for the first level classes in the hierarchy (i.e. temporal vs. atemporal). The results corresponding to the "Gold set" column denotes the overall performance (easy+hard instances). It shows that the SDE based approach performs better. The state-of-the-art system corresponds to the resource created in (Pawar et al., 2016) that was based on classifier's confidence and made use of only gloss-based features. The results reported for state-of-the-art system are for gold standard set.

Evaluation results of the second level (i.e. finer level) classification in the hierarchy are shown in Table 3. This again shows that the semantic distance-based instance selection strategy is more effective compared to the classifiers' confidence based selection strategy.

As evident from the experimental results, in both coarse (i.e. first level) as well fine-grained (i.e. second level) classification scenarios, word embedding and semantic relation based techniques improve the efficiency at a greater extent. Recall improves at a much faster rate, indicating the efficiency of word embedding features in correctly retrieving more and more instances. The gain in overall performance signifies the fact that the classifier is not only robust in handling easy-to-classify instances, but also generalizes well at predicting hard-to-classify instances. Although in CBE based method there is a drop in precision compared to the state-of-the-art system, our proposed model shows considerably higher F-measure due to the significant gain in recall. This phenomenon ensures that our current model is able to find a good trade-off between easy and hard cases.

We make a Multi-rater agreement (Fleiss, 1971) with the classification model and humans' annotations. While we look at the agreement, it was observed that for easy-to-classify instances, there is a considerably high agreement (with more than 85%) among machines (i.e. classifier) and humans. It was also observed that, for the instances where annotators had dis-agreement, classifier was also not able to properly classify-this was confirmed by an expert (non-annotator).

From our further analysis we come up with the following observations: (i). instances where both human and machine commit mistakes: for "नया *(nayA-new)* - जिसकी रचना अभी-अभी की गई हो *(jisakI rachanA abhI-abhI kI gaI ho- Which has just been created.)"*, machine assigns *'future'* whereas human assigns *'neutral'*. However, this should be *'present'* as confirmed by an expert. (ii). instances for which human makes mistakes but machine does not: e.g., 'तुरंत *(turaMta-Immediately)* - शीघ्रता से या बिना विलम्ब किए *(shIghratA se yA binA vilamba kie-Hastily or without delay.)'* is tagged as *'atemporal'* by human, but *'neutral'* by machine. (iii). instances where machine makes mistake but human correctly predicts: e.g., "ताजा *(tAjA-fresh)* - हाल ही का *(hAla hI kA-recent)"* is tagged as *'atemporal'* by the machine, but *'present'* by human.

Hierarchical First Level	CBE		SDE	
	Gold set	Easy cases	Gold set	Easy cases
Precision	80.40	80.80	88.90	91.80
Recall	82.80	88.40	89.90	95.50
F-measure	81.60	84.50	89.40	93.60
	State-of-the-art			
	Precision: 84.70, **Recall:** 72.40, **F-measure:** 78.10			

Table 2: Results of gold standard set with different expansion strategies for the first level in the hierarchy. Here, CBE: denotes candidate selection based on classifier's confidence score, SDE: denotes candidate selection based on semantic distance based measurement.

Hierarchical Second Level	CBE		SDE	
	Gold set	Easy cases	Gold set	Easy cases
Precision	72.50	72.30	73.83	74.02
Recall	69.60	76.90	70.82	77.35
F-measure	71.02	74.53	72.29	75.65
	State-of-the-art			
	Precision: 54.76, **Recall:** 55.23, **F-measure:** 54.99			

Table 3: Results with different expansion strategies for the second level classes in the hierarchy

[6] Correspond to the examples annotated by all humans with 100% agreement.

2.6. Error Analysis

We closely analyze the outputs of the classifiers to understand the behaviors of each expansion as well as classification strategy.

CBE: In this model, most of the miss-classified instances are also found to be difficult to the human annotators as temporal senses in these synsets are not directly denoted. As an example, "मरणासन्न (maraNAsanna- Moribund) - जो मरने के बहुत समीप हो *(jo marane ke bahuta samIpa ho- One who is very close to death)*" is classified as *'neutral'* even though it connotes 'futuristic' temporal sense.

SDE: This is the most effective model, hence reduces the errors significantly. The model miss-classifies those instances which either do not have any denotative temporal evidence in their glosses or fall into the difficult-to-classify cases, i.e. the human annotators are not even in perfect agreement while classifying.

We observe that both the models have complimentary behaviors. There are instances which are correctly predicted by SDE, but CBE fails and the vice-versa. Significance t-test (De Winter, 2013) confirms that the performance improvement in SDE-based approach over the CBE-based approach is statistically significant.

3. Sentence Level Temporality Detection

As an application of Tempo-Hindi-WordNet that we develop, we evaluate its effectiveness for detecting temporality at the sentence level. Each sentence is classified with one of the three temporal classes, namely *past*, *present* and *future*. As there was no sentence level temporally tagged corpus, we manually create it for benchmarking. Three experts (with post-graduate level knowledge) were asked to manually annotate two kinds of datasets: (i). The first set contains 940 sentences of ILTIMEX corpus (Ramrakhiyani and Majumder, 2015) with 281, 533 and 126 instances of past, present and future, respectively; (ii). The second dataset contains 210 tweets chosen from SAIL dataset (Patra et al., 2015) with 18, 166 and 26 instances of past, present and future, respectively. We find inter-annotator multi-rater kappa agreement (Fleiss, 1971) of 0.80.

We develop two models based on rules and supervised machine learning.

3.1. Rule-based Approach

We define a set of generic rules which we apply for determining the temporal sense of any sentence for both Twitter and Newswire text. We apply the same set of rules for the following two cases: (i). Temporal sense of each word sense in the sentence is detected using our temporal resource. The most suitable sense of each word in the sentence is determined using an unsupervised Most Frequent Sense (MFS) algorithm (Bhingardive et al., 2015). (ii). We identify the tense of each word in a sentence using a Hindi Morphological Analyzer.[7] Verbs with the tense information (past, present or future) are used for developing the rule-based system.

We depict the rules in Algorithm 2.

[7] `http://www.cfilt.iitb.ac.in/~ankitb/ma/`

Algorithm 2 Basic Steps of Rule-based Approach.

1: If majority words in a sentence belong to a particular temporal/tense category *t* then label it as *t*.
2: If the words in the sentence are equally distributed among the three classes then
   2.1. Label the sentence as present if the classes are only past and present;
   2.2. Label the sentence as future if the classes are only present and future;
   2.3. Label the sentence as future if the classes are only past and future;
   2.4. Label the sentence as future if all the three classes occur.
3: Class label is assigned at random in case no temporal/tense word is detected in the sentence.

Experimental results of this rule-based approach are shown in Table 4. It shows that the classifier created based on our temporal resource performs better than the system based on the tense information. Significance t-test (De Winter, 2013) confirmed that the performance improvement in our resource-based approach over the tense-based approach is statistically significant.

	Tense based	Temporal Resource based
ILITIMEX Corpus	(63.67, 63.18, 63.43)	**(64.90, 67.90, 66.37)**
Twitter Corpus	(45.58, 53.10, 49.06)	**(61.78, 69.65, 65.48)**

Table 4: Results using rule-based approach. Here, (x, y, z): precision, recall, and F-score.

3.2. Supervised Machine Learning Approach

We develop a SVM-based model with the following set of features.

Unigrams(UN): Word unigrams of sentence are used as features of the classifier.

Tense Synset (TenseS): Synsets of words containing tense information are used as the features. Tense information is detected by the same Hindi morphological analyzer (c.f. Section 3.1.).

Temporal Synset(TempS): WodNet synsets of temporal words present in a sentence are used as features. We use Tempo-Hindi-WordNet to determine the temporal sense.

Results of machine learning-based approach are reported in Table 5. It shows that the best result is achieved when the unigrams and temporal synset features are used together. Significance *t-test* shows that the performance improvement with this feature combination is statistically significant over the others.

	UN	UN+TenseS	UN+TempS
ILTIMEX corpus	86.92	87.42	**88.98**
Twitter corpus	84.29	84.61	**86.23**

Table 5: Results of 10-fold cross validation accuracy for machine learning based approach

3.3. Analysis of Results

In order to study the behaviors of sentence-level taggers, we analyse the outputs of the classifiers. It is found that a number of errors were contributed due to the incorrect sense marking by the MFS disambiguation algorithm (Bhingardive et al., 2015). Let us consider the following example, *"फिलहाल, सब इंस्पेक्टर तीन सप्ताह के बेड रेस्ट पर है (philahAla, saba iMspekTara tIna saptAha ke beDa resTa para hai-Currently, the sub inspector is in three week's bed rest)"*. Here, the MFS algorithm fails to identify the proper sense of the word *"फिलहाल (philahAla-Currently)"*, and thus can not directly detect temporality.

The temporal resource-based system correctly classifies many instances where the tense-based system fails. Let us consider the following example: *"हमारे पास मैच जीतने का मौका है (hamAre pAsa maicha jItane kA maukA hai-We have a chance to win the match)"* which refers to the *'future'* event. Here, the tense-based system classifies it as *'present'*, but our temporal resource-based classifier very correctly tags it as *'future'*. There are instances where the Hindi morphological analyzer fails to detect any tense information. For example, *"अब क्यों नहीं करती नारी अधिकार की बात (aba kyo.m nahii.m karatii naarii adhikaara kii baata-now, why do not you talk about women's right)"*. Here, our temporal resource-based tagger correctly classifies it as *'present'* with the help of temporal keyword *"अब (aba-now)"*. There are also some counter examples where the temporal resource-based classifier fails, but the tense-based classifier behaves properly. For example, *"मौका (maukA-opportunity)"* is a word having connotative *future* time sense. When this word appears in a sentence like *"हमने मैच जीतने का मौका गवा दिया था (hamane maicha jItane kA maukA gavA diyA thA-We missed an opportunity to win the match.)"*, it actually refers to a *past* time sense which is captured correctly by the tense-based model, whereas our resource-based system mis-classifies it as *'future'*.

Our close analysis reveals that the behaviors of rule-based and machine learning-based approaches are very often complimentary in nature, i.e. there are instances where rule-based model succeeds but the machine learning-based approach fails and the vice-versa. For example, consider the following sentence: *"उसके खिलाफ दर्ज केस वापस लेने के लिए सीबीआई पहले ही अर्जी दाखिल कर चुकी है (usake khilApha darja kesa vApasa lene ke lie sIbIAI pahale hI arjI dAkhila kara chukI hai-CBI has already filed an application for withdrawing the case against him)"*. Here, the rule-based approach classifies it as *'present'* but the machine learning-based approach correctly classifies it as *'past'*. In the following sentence: *"उन्हें चाजुन से कड़ी टक्कर मिलने की संभावना है (unheM chAjuna se kaDI Takkara milane kI saMbhAvanA hai-He is likely to get tough competition from Chajun)"*, the machine learning-based approach incorrectly predicts it as *present* but the rule-based approach correctly predicts it as *future*.

There are also some instances where both the rule-based and the machine learning-based approaches fail. For example, *"इस बार उनकी नजरें गोल्ड मेडल पर हैं (isa bAra unakI najareM golDa meDala para haiM-This time her eyes are on gold medal)"*. Here, both rule-based and machine learning-based methods incorrectly classify the sentence as *present*. However, this is actually an instance of *future*. In order to perform quantitative analysis we create confusion matrix that shows that the system is mostly confused in discriminating *present* from the *future* classes.

4. Conclusions

In this paper, we have presented a framework for sentence-level temporality detection in Hindi. In order to achieve this, we propose a semi-supervised learning framework for finding temporal sense of each word in the sentence. This classifies the entire Hindi-WordNet into five classes. We have used three learning algorithms and several expansion strategies. A gold standard test set is also created to perform detailed evaluation. Finally, we show how the temporal resource can be used for temporality detection at the sentence level. We develop two versions: rule-based and machine learning-based. These have been evaluated on two different domain corpora, namely Twitter (informal text) and newswire (formal text). Evaluation shows that such a temporal resource will facilitate research in temporal IR/NLP. Our proposed method is generic and can be adapted to other languages and domains with the availability of minimal resource such as the WordNet.

In future, we will like to investigate a hybrid expansion strategy for resource creation where probabilistic expansion and semantic distance based expansion will be joined together to exploit each other's merit. For sentence-level tagging, we will explore deep learning based methods.

5. Acknowledgements

Asif Ekbal acknowledges Young Faculty Research Fellowship (YFRF), supported by Visvesvaraya PhD scheme for Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia).

6. Bibliographical References

Allan, J., Gupta, R., and Khandelwal, V. (2001). Temporal Summaries of News Topics. In *SIGIR 2001: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 10–18, New Orleans, Louisiana, USA.

Alonso, O., Gertz, M., and Baeza-Yates, R. A. (2009). Clustering and Exploring Search Results using Timeline Constructions. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM*, pages 97–106, Hong Kong, China.

Bhingardive, S., Singh, D., V, R., Redkar, H. H., and Bhattacharyya, P. (2015). Unsupervised Most Frequent Sense Detection using Word Embeddings. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1238–1243, Denver, Colorado, USA.

De Winter, J. C. (2013). Using the Student's T-test with Extremely Small Sample Sizes. *Practical Assessment, Research & Evaluation*, 18(10).

Dias, G. H., Hasanuzzaman, M., Ferrari, S., and Mathet, Y. (2014). TempoWordNet for Sentence Time Tagging. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, pages 833–838, Geneva, Switzerland.

Fleiss, J. L. (1971). Measuring Nominal Scale Agreement among Many Raters. *Psychological Bulletin*, 76(5):378–382.

Hasanuzzaman, M., Dias, G., Ferrari, S., and Mathet, Y. (2014). Propagation Strategies for Building Temporal Ontologies. In *14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 6–11, Gothenburg, Sweden.

Jatowt, A., Yeung, C. A., and Tanaka, K. (2013). Estimating Document Focus Time. In *22nd ACM International Conference on Information and Knowledge Management, CIKM'13*, pages 2273–2278, San Francisco, CA, USA.

Joachims, T. (2002). *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms.* Kluwer Academic Publisher, Dordrecht, Netherlands.

John, G. H. and Langley, P. (1995). Estimating Continuous Distributions in Bayesian Classifiers. In *Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345, San Mateo, CA.

Joho, H., Jatowt, A., and Blanco, R. (2013). A Survey of Temporal Web Search Experience. In *22nd International World Wide Web Conference, WWW '13*, pages 1101–1108, Rio de Janeiro, Brazil.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *CoRR*, abs/1301.3781.

Patra, B. G., Das, D., Das, A., and Prasath, R. (2015). Shared Task on Sentiment Analysis in Indian Languages (SAIL) Tweets - An Overview. In *Mining Intelligence and Knowledge Exploration - Third International Conference, MIKE 2015*, pages 650–655, Hyderabad, India.

Quinlan, R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA.

Ramrakhiyani, N. and Majumder, P. (2015). Approaches to Temporal Expression Recognition in Hindi. *ACM Trans. Asian & Low-Resource Lang. Inf. Process.*, 14(1):2.

Schockaert, S., Ahn, D., De Cock, M., and Kerre, E. E. (2006). Question Answering with Imperfect Temporal Information. In *Flexible Query Answering Systems*, pages 647–658. Springer, Berlin, Germany.

UzZaman, N., Llorens, H., Derczynski, L., Allen, J. F., Verhagen, M., and Pustejovsky, J. (2013). SemEval-2013 Task 1: TempEval-3: Evaluating Time Expressions, Events, and Temporal Relations. In *Proceedings of the 7th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2013*, pages 1–9, Atlanta, Georgia, USA.

Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Moszkowicz, J., and Pustejovsky, J. (2009). The Tempeval Challenge: Identifying Temporal Relations in Text. *Language Resources and Evaluation (LRE)*, 43(2):161–179.

Verhagen, M., Saurí, R., Caselli, T., and Pustejovsky, J. (2010). Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval@ACL 2010*, pages 57–62, Uppsala, Sweden.

7. Language Resource References

Bhattacharyya, P. (2010). Indowordnet. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

Bojar, O., Diatka, V., Rychlý, P., Stranák, P., Suchomel, V., Tamchyna, A., and Zeman, D. (2014). HindEnCorp - Hindi-English and Hindi-only Corpus for Machine Translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, pages 3550–3555, Reykjavik, Iceland.

Pawar, D., Hasanuzzaman, M., and Ekbal, A. (2016). Building Tempo-HindiWordNet: A Resource for Effective Temporal Information access in Hindi. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016*, Portorož, Slovenia.