Retrieval approach to extract opinions about people from resource scarce language News articles

Aditya Mogadala
Search and Information Extraction Lab, IIIT-H
Hyderabad
India
aditya.m@research.iiit.ac.in

Vasudeva Varma
Search and Information Extraction Lab, IIIT-H
Hyderabad
India
vv@iiit.ac.in

ABSTRACT

We wish to address the challenging task of opinion mining about organizations, people and places from different languages. It is known that resources and tools for mining opinions are scarce. In our study, we leverage comparable news articles collection to retrieve opinions about people (opinion targets) in resource scarce language like Hindi. Opinions expressed about opinion targets (Named Entities)given by adjectives and verbs known as opinion words are extracted from English collection of comparable corpora to get transliterated and translated to resource scare languages. Transformed opinion words are then used to create subjective language model (SLM) and structured opinion queries (OQs) using inference network (IN) for retrieval to confirm the opinion about opinion targets in documents. Experiments have shown that OQs and SLM with IN framework are effective for opinion mining tasks in minimal resource languages when compared to other retrieval approaches.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Languages, Experimentation

Keywords

Opinion Mining, Subjectivity Analysis, Resource scare languages

1. INTRODUCTION

News articles are written in different languages containing different entities like places, people or organizations. Factual information provided about these entities (mainly people) is sometimes added with support or expressions of the writer. This induces opinion about people in the article indirectly. Opinions can be present in various granularities such as a word, a sentence, a text etc. Although each granularity is important, we focus our attention on word-level opinion detection. For Example, Figure 1 highlights people (Opinion Targets) in green and opinions expressed about them in red at word-level from English and Hindi news articles.

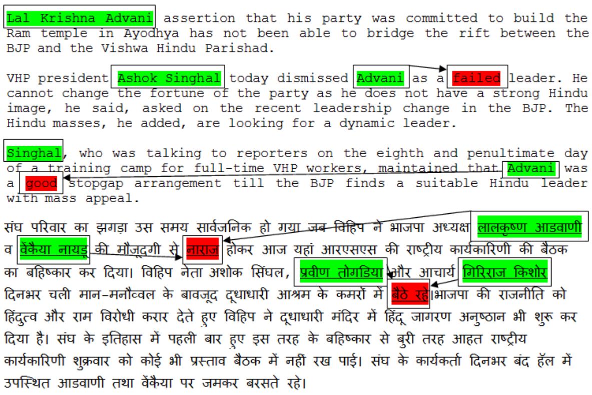
Mining opinions about people in news articles [1] will help us to distinguish the factual information and writers view on them. Also, it will help us understand the bias of news agencies or writers towards them. Comparison of opinions about people mentioned in different news agencies articles can be done to see the authenticity of information. It also helps us tracking opinions about people over different timelines.

Among the different problems observed in news articles, one that caught our attention is the fact that they contain unintelligent/unclear/misinterpreted opinions. We know that opinions are generally expressed in subjective text. There is some evidence from previous works [5] that 44% of sentences in a news collection are subjective. Hence, subjective text categorization from objective text at sentence level will help us in mining opinions. However, most of the techniques [18] which identify subjectivity at sentence level are supervised and require lot of labeled data. This creates a problem for languages other than English due to availability of less linguistic resources. Also, identification of people (opinion target) and opinions about them requires state of art word identification tools. But, resource scarce languages lack high precision state of art natural language processing tools like named entity recognizers, part of speech (POS) taggers, dependency parsers, semantic role labeling tools etc.

So, major problems that this paper tries to address are (1) can opinion extraction about people in Hindi news articles is achieved using minimal language specific resources and tools. (2) can we leverage on resource rich languages English using a comparable corpora. (3) can we surpass word identification tools in resource scarce languages to identify named entities, adjectives and verbs [14] used for subjective text. (4) can we design an approach which is less dependent on language specific tools and is easily scalable to different languages. (5) Build a reusable dataset, to facilitate follow-up research.

The remainder of the paper is organized as follows. In the next section 2, we describe related work. In Section 3 approach is described and in Section 4 query formulation is explained. We present our experimental setup and results in Section 5 and Section 6 respectively. We present result discussion in Section 7 and conclude with a discussion of

WISDOM'12, August 12 2012, Beijing, China Copyright 2012 ACM 978-1-4503-1543-2/12/08...$15.00



Lal Krishna Advani assertion that his party was committed to build the Ram temple in Ayodhya has not been able to bridge the rift between the BJP and the Vishwa Hindu Parishad.

VHP president Ashok Singhal today dismissed Advani as a failed leader. He cannot change the fortune of the party as he does not have a strong Hindu image, he said, asked on the recent leadership change in the BJP. The Hindu masses, he added, are looking for a dynamic leader.

Singhal, who was talking to reporters on the eighth and penultimate day of a training camp for full-time VHP workers, maintained that Advani was a good stopgap arrangement till the BJP finds a suitable Hindu leader with mass appeal.

संघ परिवार का झगड़ा उस समय सार्वजनिक हो गया जब विहिप ने भाजपा अध्यक्ष लालकृष्ण आडवाणी व वेंकैया नायडू की मौजूदगी से नाराज होकर आज यहां आरएसएस की राष्ट्रीय कार्यकारिणी की बैठक का बहिष्कार कर दिया। विहिप नेता अशोक सिंघल, प्रवीण तोगडिया और आचार्य गिरिराज किशोर दिनभर चली मान-मनौव्वल के बावजूद दूधाधारी आश्रम के कमरों में बैठे रहे।भाजपा की राजनीति को हिंदुत्व और राम विरोधी करार देते हुए विहिप ने दूधाधारी मंदिर में हिंदू जागरण अनुष्ठान भी शुरू कर दिया है। संघ के इतिहास में पहली बार हुए इस तरह के बहिष्कार से बुरी तरह आहत राष्ट्रीय कार्यकारिणी शुक्रवार को कोई भी प्रस्ताव बैठक में नहीं रख पाई। संघ के कार्यकर्ता दिनभर बंद हॅल में उपस्थित आडवाणी तथा वेंकैया पर जमकर बरसते रहे।

Figure 1: Example of sentences with Opinions and Opinion Targets in Comparable news Corpora

future work.

2. RELATED WORK

Opinion mining can be broadly categorized into classification and retrieval methods.

2.1 Classification Methods

In our study we found an approach to extract opinion holders and relevant opinions, it uses parsing and Maximum Entropy ranker methods based on parse features. [12] However the problem we face with this approach is that it is not easily portable to resource scarce languages due to low quality parsers. The second approach [13] collected opinion words labels manually and used them to find opinion-related frames (FrameNet) which are then semantic role labeled to identify fillers for the frames. However we cannot apply this approach as well due to non-availability of semantic role labeling tools in resource scarce languages.

2.2 Retrieval Methods

Traditional opinion retrieval systems in TREC [21, 7] tried on large blog collections extract opinions about a query. These systems are designed for retrieving opinions about a single entity i.e. a query. Using this systems on news articles might fail because of the presence of multiple entities with different opinions expressed on them. Some other approaches [11] used opinion weights and proximity of terms to the query directly. They considered proximity-based opinion density functions to capture the proximity information. But these approaches may not be effective in retrieving information from resource scarce language documents due to dependency on language specific features. Some systems developed for languages Romanian [2], Chinese [19, 22] , Japanese [20, 4] only concentrate on identification of subjective texts. Other multilingual systems participated in NTCIR-6 [8] opinion extraction track were dataset specific and their approaches cannot be easily ported to other languages.

This shows a need for better and new approaches for opinion mining in resource scarce languages like Hindi to overcome the above issues.

3. APPROACH

In order to achieve opinion extraction about opinion targets, retrieval approach is chosen to overcome the problems involved in availability of state of art NLP tools and resources. We leveraged resource rich language like English to extract opinions about people from resource scarce language like Hindi. Initially, Named entities, adjectives, verbs from English collection are extracted using named entity recognition tools and POS taggers. It is then transliterated and translated to other languages using conditional random field approach [15] and bilingual dictionaries respectively. Next, cross language ported words are used to identify subjective sentences to create a subjective language model (SLM) and opinion queries in Hindi. Inference Network(IN) framework which support proximity and belief between query words is then used to create structured opinion queries (OQs) with SLM similar to Language Model (LM) with IN [3] for retrieval of documents to confirm the presence of opinions about opinion targets in documents.

In the following sections detailed description for achieving subjective sentence extraction, subjective language model (SLM) and extension of SLM with IN for opinion retrieval is described.

3.1 Subjective sentence extraction

Subjective sentences were selected in the document using two approaches motivated from [10]. But, we consider named entities also to find opinion targets. Documents without subjective sentences are eliminated. Two different approaches are used to extract subjective sentences.

Method 1
If two or more strong subjective expressions occur in the same sentence mainly Named entities, Adjectives or Verbs, the sentence is labeled strongly subjective.

Method 2
If at least one of Named Entities or Adjectives or Verbs exist in a sentence then it is labeled as weakly subjective sentence.

Difference between performance of *Method 1* and *Method 2* can be observed in the experiments when OQs are used for retrieval.

3.2 Subjective Language Model for Opinion retrieval

Subjective language model (SLM) is created for resource scare language documents similar to language model (LM) [6] by selecting **subjective sentences in the documents** using two different methods mentioned earlier.

SLM approach to opinion retrieval is finding probability of an opinion query oq being generated by a probabilistic model based on a document D denoted as $p(oq|D)$. It is done by estimating the posterior probability of document D_i and opinion query oq using Bayes formula given by Equation 1.

$$p(d_i|oq) \propto p(oq|D_i)p(D_i) \qquad (1)$$

where $p(D_i)$ is prior belief that is relevant to any opinion query and $p(oq|D_i)$ is the query likelihood given the document D_i, which captures the particular opinion query oq information. $p(D_i)$ is considered to be multinomial distribution. This assumption helps in choosing better smoothing techniques which is mentioned later.

For each document D_i in the collection c, its subjective language model defines the probability $p(ow_1, ow_2, ..., ow_n|D_i)$ containing opinions and opinion targets given by $ow_1, , ow_n$ as sequence of n query terms. Documents are ranked according to this probability.

The probabilities of the opinion words ow_i in document D_i improves the weight of subjectivity in the document D_i. Equation 2 gives probability $p(ow_i|c)$ of finding opinion words ow_i for entire collection c while Equation 3 gives probability $p(ow_i|D)$ only for a document D.

$$p(ow_i|c) = \frac{cfreq(ow_i, c)}{\Sigma_{i=1}^{n} cfreq(ow_i, c)} \tag{2}$$

$$p(ow_i|D) = \frac{tfreq(ow_i, D)}{\Sigma_{i=1}^{m} tfreq(ow_i, D)} \tag{3}$$

Here, $cfreq(ow_i, c)$ represents collection frequency of ow_i in the collection c and $tfreq(ow_i, D)$ is term frequency of the ow_i in a document D. n is total opinion words in collection, while m is total opinion words in a document. The non smoothed model gives maximum likelihood estimate of relative counts. But if the word is unseen in the document then it results in the zero probability. So the smoothing is helpful to assign a non-zero probability to the unseen words and improve the accuracy of word probability estimation in general. In this paper, we used Dirichlet and Jelinek-Mercer smoothing to assign non-zero probabilities to unseen words in the documents and collection. Below are the two smoothing techniques that are used to remove the zero probability scores to unseen words as mentioned in [23].

Dirichlet Smoothing

As subjective language model prior is considered to be multinomial distribution, for which the conjugate prior for Bayesian analysis is the Dirichlet distribution. We choose the parameters of the Dirichlet to be μ and the values that needs to be added in the numerator of the Equation 3 for smoothing. Equation 4 gives values which is Dirichlet parameter multiplied with probability of each opinion word in collection. So after smoothing, the Equation 3 is converted into Equation 5.

$$\mu p(ow_1|c), \mu p(ow_2|c), ..., \mu p(ow_n|c) \tag{4}$$

$$p_\mu(ow_i|D) = \frac{tfreq(ow_i, D) + \mu p(ow_i|c)}{\Sigma_{i=1}^{m} tfreq(ow_i, D) + \mu} \tag{5}$$

Jelinek-Mercer Smoothing

In Jelinek-Mercer smoothing approach, we consider the mixture of document model $p(ow_i|D)$ and collection model $p(ow_i|c)$ as used in standard retrieval model. This approach takes parameter λ which needs to be estimated. Equation 6 gives the mixture model equation.

$$p_\lambda(ow_i|D) = (1-\lambda)p(ow_i|c) + (\lambda)p(ow_i|D) \tag{6}$$

Subjective language model with *Method 1* and *Method 2* forms our extended baseline from basic language model. We will further extend this model with **inference network which allows different types of structured query formulations** as explained below.

3.3 Subjective Language model with Inference network for Opinion retrieval

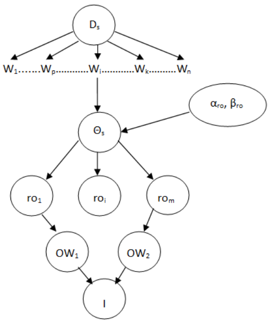


Figure 2: Inference Network Representation for SLM

This retrieval model combines SLM with IN [17] to confirm the presence of opinion about opinion targets. This model allows opinion queries containing possible opinions on opinion targets to use proximity and belief information between query terms similar to Indri [16]. We observe in IN framework that documents are ranked according to probability $p(I|D, \alpha, \beta)$ using the belief information need I calculated using a document D and hyper parameters α and β as evidence.

Information need I in our scenario is simply a belief node that combines all of the belief nodes ow_i's containing **opinion evidence** within the inference network into a single belief. In our scenario, belief nodes ow_i's are opinion words in the document. In order to obtain evidence of ow_i's, representation concept nodes ro_i's are used. ro_i's are binary random variables representing only opinion word unigrams from the total features extracted in the document representation. Features here are all word unigrams $w_1....w_n$ present in a document. In order to find the individual belief nodes ow_i's we need to calculate $p(ro_i|D)$ and then combine ow_i's to get information need I. Figure 2 shows the representation. Documents are then ranked accordingly using $p(I|D, \alpha, \beta)$.

To achieve it, we assume each subjective document D_s to be in multiple-Bernoulli subjective model θ_s and not multinomial distribution as assumed in previous section. As this model assumption imposes concept nodes ro_i's to be independent. First, we compute $p(\theta_s|D_s)$ which is the model posterior distribution given by Equation 7.

$$p(\theta_s|D_s) = \frac{p(D_s|\theta_s)p(\theta_s)}{\int_{\theta_s} p(D_s|\theta_s)p(\theta_s)d\theta_s} \tag{7}$$

Where $p(D_s|\theta_s)$ is the likelihood of generating D_s from model θ_s and $p(\theta_s)$ is the model prior. We see this posterior probability $p(\theta_s|D_s)$ is distributed according to multiple-Beta $(\alpha_{ro}, \beta_{ro})$ because beta distribution is the Bernoulli's conjugate prior.

In this IN framework only opinion terms are considered denoted by $optf_{ro}$ out of total terms in a document D_s for independent ro_i's. This is like finding x positive results for n trials. Thus $p(D_s|\theta_s)$ distribution is changed to multiple-Beta$(\alpha_{ro} + optf_{ro}, \beta_{ro} + |D_s| - optf_{ro})$ containing opinion terms, where $|D_s|$ is the total opinion word count of the document D_s. Expression 8 give estimate of $p(ro|D_s)$ representing individual beliefs ow_i's. It is nothing but a expectation over the posterior $p(\theta_s|D_s)$.

$$p(ro|D_s) = \int p(ro|\theta_s)p(\theta_s|D_s)d\theta_s \qquad (8)$$

But we know the expectation of a beta distribution given in terms of its parameters is $\frac{\alpha}{\alpha+\beta}$. Therefore, given that $p(D_s|\theta_s)$ is also distributed according to multiple-Beta(α_{ro}+$optf_{ro}$, β_{ro}+$|D_s|$−$optf_{ro}$) the Equation 8 is now represented by Equation 9.

$$p(ro|D_s) = \frac{optf_{ro,D_s} + \alpha_{ro}}{|D_s| + \alpha_{ro} + \beta ro} \qquad (9)$$

So for document D_s, $optf_{ro,D_s}$ is the opinion term frequency with ro_i's as features. Thus subjective language model θ_s is estimated based on hyper parameters α_{ro} and β_{ro} combined with the observed document D_s. From these models, concept nodes q_i's are used forming an opinion query. Overall belief I from this opinion query is used for ranking subjective documents.

But there can be chances of zero probability and data sparseness in the model. In order to eliminate this problem we employ smoothing.

Dirichlet Smoothing

Dirichlet smoothing is done in order to handle zero probability. α_{ro} and β_{ro} values were chosen given by Equation 10 and Equation 11 respectively to modify Equation 9 to Equation 12. $p(ro|c)$ gives beliefs of the representation concept nodes in entire document collection.

$$\alpha_{ro} = \mu p(ro|c) \qquad (10)$$

$$\beta ro = \mu(1 - p(ro|c)) \qquad (11)$$

$$p(ro|D_s) = \frac{optf_{ro,D_s} + \mu p(ro|c)}{|D_s| + \mu} \qquad (12)$$

where μ is free parameter and $optf_{ro,D_s}$ is the opinion terms frequency for feature ro in Document D_s.

Jelinek-Mercer Smoothing

This method involves a linear interpolation of the maximum likelihood model with the collection model, using a coefficient λ.

$$p(ro|D_s) = (1-\lambda)\frac{optf_{ro,D_s}}{|D_s|} + \lambda p(ro|c) \qquad (13)$$

Thus, this model combines SLM and IN with different smoothing techniques used for opinion retrieval. Queries formed for retrieval are mentioned in next section.

4. QUERY FORMULATION FOR RETRIEVAL

We understood from the previous section that belief nodes ow_i's combine evidence from the concept nodes ro_i's to estimate the belief that a opinion words are expressed in a document. But actual arrangement of the ow_i's and the way they combine evidence is dictated by the user through the query formulation. So we form structured opinion queries (OQs) to identify opinion targets and opinions in resource scare language news articles. These structured OQs contain named entities, adjectives and verbs as opinion words. OQs use proximity and beliefs between query words.

We will see difference in query formulation of opinion queries which use proximity and belief between query words and that don't use below.

Label	Opinion Queries
UQ	NE OExp
SQ 1	#10(NE OExp)
SQ 2	#filreq(NE #Combine(NE OExp))
SQ 3	#filreq(NE #uw10(NE OExp)
SQ 4	#filreq(NE #weight(2.0 #uw10(NE OExp)))
SQ 5	#uw10(NE OExp)
SQ 6	#uw5(NE OExp)

Table 1: Opinion Queries

Label	Opinion Queries Explanation
UQ	Unstructured query containing OT (named entity) and opinion expressed (OExp).
SQ 1	Match OT and OExp in ordered text within window of 10 words.
SQ 2	Evaluates combined beliefs of OT and OExp given OT in document.
SQ 3	Match OT and OExp in unordered text within window of 10 words given OT in document.
SQ 4	Match OT and OExp in unordered text within window of 10 words with extra weight of 2.0 and OT given in document.
SQ 5	Match OT and OExp in unordered text within window of 10 words.
SQ 6	Match OT and OExp in unordered text within window of 5 words.

Table 2: Explanation of Opinion Queries containing Opinion Target(OT) and Opinions Expressed (OExp)

4.1 Opinion queries without proximity and belief

Query terms in OQ are treated as bag of words. Each word in the opinion query is assumed independent without any conditional information. This query model may find relevant documents to query terms but may not extract opinions about opinion targets. In order to rank documents, opinion query likelihood is calculated using subjective language model θ_s given by Equation 14.

$$p(oq_1, oq_2|\theta_s) = \Pi_{i=1}^{2} p(oq_i|\theta_s) \qquad (14)$$

where oq_1 and oq_2 represent opinion targets and opinions on them respectively.

4.2 Opinion queries with proximity and belief

Indri Query language[1] is used to form OQs. In this approach, only those query operators are selected which use proximity and belief information in their representations. Proximity representations are used to map the opinion targets and opinions expressed on them appearing within ordered or unordered fixed length window of words. To use belief information of opinion words of OQ represented as belief nodes of subjective documents. Opinion words are combined using a belief operator.

Size used for proximity window is intuitive. Each query represented in this framework uses SLM with IN for opinion retrieval. OQs used for retrieval are mentioned in Table 1 and their explanation in Table 2.

5. EXPERIMENTAL SETUP

[1] http://www.lemurproject.org/lemur/IndriQueryLanguage.php

Topics Used for Experiments					
Topic	T1	T2	T3	T4	T5
	78	80	85	90	91

Table 3: Topics used for Experiments

Inter-Annotator agreement		
Task	Agree	Observed kappa
Subjective sentences	86.5%	0.689
(Opinions,Opinion targets)	73.7%	0.493

Table 4: Inter Annotator agreement

Information about collection and evaluation metrics used for assessment is mentioned below.

Collection
Our experiments are based on 5 different topics selected from FIRE 2010[2] collection. It is a comparable corpora consisting of news articles in English and Hindi languages covering different areas like sports, politics, business, entertainment etc. The relevance judgments are provided for the topics. Relevance judgments of Hindi and English are used to select the relevant documents for a topic. Table 3 show the topics used for experiments. For each topic, Hindi documents are used to create a Gold standard dataset[3] of subjective sentences and tuples of opinion targets and opinions. While, relevant documents in English are used for extraction of NE's, adjectives and verbs. Gold standard dataset was prepared using two annotators who identified subjective sentences and opinion word with its target if they existed in the document. The inter-annotator agreement in identifying subjective sentences, opinions and opinion targets tuples averaged over 5 topics is given in Table 4.

Preprocessing on the Collection
English data from the FIRE 2010 collection is used to extract NE's, adjectives and verbs. Stanford NER[4] and POS Tagger[5] is used to extract NE's and adjective, verbs respectively. Manually prepared word-aligned bilingual corpus and statistics over the alignments is used to transliterate English to top 3 Hindi words. For this Hidden Markov Model (HMM) alignment and Conditional Random Fields (CRFs) are used. For HMM alignment GIZA++[6] is used while CRF++[7] is used for training the model. For translation of adjectives and verbs, English-Hindi dictionary shabdanjali[8] is used.

Word distribution of translated Opinions				
	Positive	Negative	Neutral	Total
Adjective	151	143	577	871

Table 5: Word Distribution of translated Opinions

English to Hindi Document Analysis	
Translation Coverage(ADJ)(After Exp)	63.9%
Translation Coverage(VB)(After Exp)	65.3%
Transliteration Error	13%

Table 6: English to Hindi Document Analysis

Before doing the translation adjectives and verbs are expanded with Wordnet[9]. Table 6 shows translation coverage, transliteration errors and Table 5 show the word distribution of translated opinions [9] averaged over 5 topics.

Evaluation
Relevant documents are retrieved for OQs created in each topic. But, documents retrieved may not represent opinion about opinion targets present in OQs. Evaluation is done using recall(R_{oq}), precision(P_{oq}) and F-measure(F_{oq}) for each OQ used for document retrieval to confirm whether OQ terms represent opinion about opinion targets in that document. Equation 15 and Equation 16 gives the metrics. Similar evaluation is done for subjective sentences using precision(P_s), Recall(R_s) and F-measure(F_s).

$$R_{oq} = \frac{Retrieved\_OQs\_in\_Document}{Total\_OQs\_present\_in\_Document} \quad (15)$$

$$P_{oq} = \frac{Relevant\_OQs\_in\_Document}{Retrieved\_OQs\_in\_Document} \quad (16)$$

$$F_{oq} = 2 * \frac{P_{oq} * R_{oq}}{P_{oq} + R_{oq}} \quad (17)$$

We also calculated mean average precision(MAP) scores to see whether the relevant documents are ranked first for the corresponding OQs. If the MAP scores are high for a model and its corresponding OQ, it can be derived that the model and query is efficient in retrieving more relevant documents first.

6. EXPERIMENTS

In this section we evaluate subjective sentence extraction, identification of opinion about opinion targets and ranking of relevant documents using the proposed approach on the gold standard dataset.

6.1 Detecting Subjective Sentences

Subjective sentences are identified using the two methods mentioned in Section 3. To analyze the accuracy of proposed methods P_s, R_s and F_s is calculated. Table 7 show the average scores for 5 topics. It can observed that Method 2 has 84.1% more recall but 7.3% low in precision compared to Method 1. For opinion mining we feel precision matters more in-order to get accurate and efficient results. This analysis was done in next section to analyze the accuracy of opinions retrieved using this two methods. We also did comparison of the following approach with classification methods by 10-cross validation on human annotated sentences using unigrams as features. Table 8 show the 10-cross validation results of learning methods.

We can observe that *Method2* achieves 2.3% more F-measure compared to naive bayes, but 4.8% less compared to SVM

[2] http://www.isical.ac.in/ fire/2010/index.html
[3] Will be made available on request
[4] http://nlp.stanford.edu/ner/index.shtml
[5] http://nlp.stanford.edu/software/tagger.shtml
[6] http://www.fjoch.com/GIZA++.html
[7] http://crfpp.sourceforge.net/
[8] http://www.shabdkosh.com/archives/content/
[9] http://wordnet.princeton.edu/

Topic		Method 1(**M1**)	Method 2(**M2**)
(T1,T2,T3,T4,T5)	P_s	**0.573**	0.534
	R_s	0.543	**1**
	F_s	0.557	**0.696**

Table 7: Subjective Sentence Accuracy

Topic		Naive Bayes	SVM	Decision Tree
(T1,T2,T3,T4,T5)	P_s	0.71	**0.73**	0.69
	R_s	0.66	**0.73**	**0.73**
	F_s	0.68	**0.73**	0.71

Table 8: Subjective Sentence Classification

and 2.0% less compared to decision trees. Similarly, we observe that *Method2* does not fair well in getting good F-measure. But was able to achieve good precision scores compared to learning methods. Since our approaches are unsupervised and achieves decent accuracy in identifying subjective sentences compared to supervised learning approaches. We feel these approaches for resource scarce languages can show significant results in identifying subjective sentences.

6.2 Detecting Opinions about Opinion Bearers

In our retrieval approach, query terms are used to confirm their presence in the document. So, SLM is created from sentences obtained using *Method 1*(M1) and *Method 2*(M2). SLM is then extended with IN for forming OQs for retrieval. Our approach is compared with other standard retrieval approaches like LM based retrieval, LM with IN retrieval using lemur toolkit[10]. Since each topic can produce as many OQs given by Equation 18 from English collection. Only those queries which retrieved documents are considered for evaluation.

$$Total\_OQ's = Total\_NE\_Hindi * OW \tag{18}$$

$$OW = (Total\_Adj + Total\_VB's) * (NumofOQ's) \tag{19}$$

P_{oq}, R_{oq} and F_{oq} is calculated for 5 topics using baseline LM, LM with IN, SLM and SLM with IN based retrieval using Dirichlet and Jelinek-Mercer smoothing techniques given by Table 9 and Table 10 respectively. In each column best performing model and its corresponding query is highlighted.

6.3 Documents Ranking

We analyzed the efficiency of OQs and models in retrieving the relevant documents first using mean average precision(MAP) scores. For that we calculated the MAP@4 scores of all the queries which retrieved at-least 4 documents. Average MAP@4 scores are calculated for 5 topics using baseline LM, LM with IN, SLM and SLM with IN based retrieval using Dirichlet smoothing technique given by Table 11 and Jelinek-Mercer smoothing by Table 12. Figure 3 and Figure 4 shows the MAP@4 scores calculated for 5 different topics using SLM+IN+M1 model and different OQs using two different smoothing techniques.

7. RESULT DISCUSSION

Model		UQ	SQ1	SQ2	SQ3	SQ4	SQ5	SQ6
Baseline LM	P_{oq}	0.156	-	-	-	-	-	-
	R_{oq}	1.000	-	-	-	-	-	-
	F_{oq}	**0.270**	-	-	-	-	-	-
SLM+M1	P_{oq}	0.125	-	-	-	-	-	-
	R_{oq}	1.000	-	-	-	-	-	-
	F_{oq}	0.222	-	-	-	-	-	-
SLM+M2	P_{oq}	0.156	-	-	-	-	-	-
	R_{oq}	1.000	-	-	-	-	-	-
	F_{oq}	**0.270**	-	-	-	-	-	-
LM+IN	P_{oq}	0.156	0.076	0.093	0.214	0.214	0.214	0.250
	R_{oq}	1.000	0.406	1.000	0.397	0.397	0.397	0.125
	F_{oq}	**0.270**	0.128	0.171	0.278	0.278	0.278	0.167
SLM+IN+M1	P_{oq}	0.125	0.272	0.093	0.333	0.333	0.333	0.250
	R_{oq}	1.000	0.343	1.000	0.375	0.375	0.375	0.125
	F_{oq}	0.222	**0.304**	0.171	**0.352**	**0.352**	**0.352**	0.167
SLM+IN+M2	P_{oq}	0.156	0.076	0.167	0.214	0.214	0.214	0.250
	R_{oq}	1.000	0.406	1.000	0.437	0.437	0.437	0.125
	F_{oq}	**0.270**	0.128	**0.286**	0.288	0.288	0.288	0.167

Table 9: Results obtained using Dirichlet Smoothing

Model		UQ	SQ1	SQ2	SQ3	SQ4	SQ5	SQ6
Baseline LM	P_{oq}	0.116	-	-	-	-	-	-
	R_{oq}	1.000	-	-	-	-	-	-
	F_{oq}	0.207	-	-	-	-	-	-
SLM+M1	P_{oq}	0.125	-	-	-	-	-	-
	R_{oq}	1.000	-	-	-	-	-	-
	F_{oq}	0.222	-	-	-	-	-	-
SLM+M2	P_{oq}	0.156	-	-	-	-	-	-
	R_{oq}	1.000	-	-	-	-	-	-
	F_{oq}	**0.270**	-	-	-	-	-	-
LM+IN	P_{oq}	0.116	0.076	0.081	0.204	0.204	0.204	0.250
	R_{oq}	1.000	0.406	1.000	0.397	0.397	0.397	0.125
	F_{oq}	0.207	0.128	0.149	0.269	0.269	0.269	0.167
SLM+IN+M1	P_{oq}	0.125	0.272	0.093	0.333	0.333	0.333	0.250
	R_{oq}	1.000	0.343	1.000	0.375	0.375	0.375	0.125
	F_{oq}	0.222	**0.304**	0.171	**0.352**	**0.352**	**0.352**	0.167
SLM+IN+M2	P_{oq}	0.156	0.076	0.156	0.214	0.214	0.214	0.250
	R_{oq}	1.000	0.406	1.000	0.437	0.437	0.437	0.125
	F_{oq}	**0.270**	0.128	**0.270**	0.288	0.288	0.288	0.167

Table 10: Results obtained using Jelinek-Mercer Smoothing

MAP@4 (Dirichlet)							
Model	**UQ**	**SQ1**	**SQ2**	**SQ3**	**SQ4**	**SQ5**	**SQ6**
Baseline LM	0.277	-	-	-	-	-	-
SLM+M1	**0.295**	-	-	-	-	-	-
SLM+M2	0.285	-	-	-	-	-	-
LM+IN	0.277	0.294	0.275	0.310	0.310	0.320	**0.322**
SLM+IN+M1	**0.295**	**0.314**	**0.295**	**0.325**	**0.320**	**0.392**	0.275
SLM+IN+M2	0.285	0.312	0.283	0.314	0.310	0.361	0.275

Table 11: MAP@4 Values with Dirichlet Smoothing

MAP@4 (Jelinek-Mercer)							
Model	**UQ**	**SQ1**	**SQ2**	**SQ3**	**SQ4**	**SQ5**	**SQ6**
Baseline LM	0.284	-	-	-	-	-	-
SLM+M1	**0.310**	-	-	-	-	-	-
SLM+M2	0.300	-	-	-	-	-	-
LM+IN	0.284	0.298	0.282	0.302	0.310	0.320	**0.310**
SLM+IN+M1	**0.310**	**0.348**	**0.324**	**0.315**	**0.320**	**0.351**	0.294
SLM+IN+M2	0.300	0.334	0.310	**0.315**	0.315	0.344	0.294

Table 12: MAP@4 Values with Jelinek-Mercer Smoothing

[10] http://www.lemurproject.org/

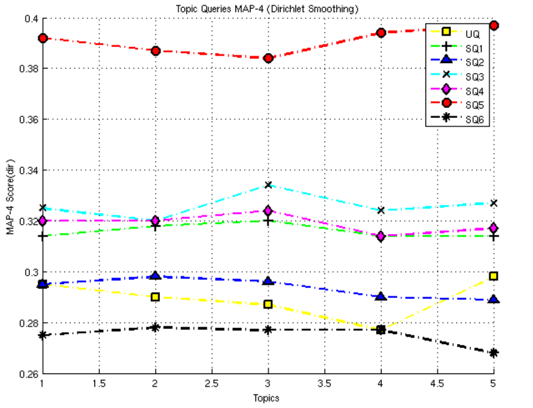


Figure 3: MAP@4(Dirichlet) for SLM+IN+M1 Model

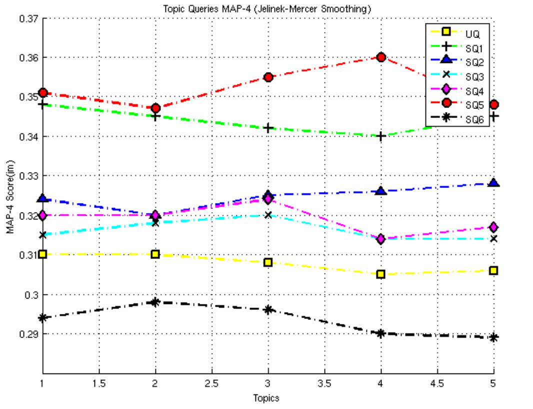


Figure 4: MAP@4(JM) for SLM+IN+M1 Model

It can observed that the best performing queries from Table 9 which used Dirichlet smoothing technique are SQ3, SQ4 and SQ5 of SLM+IN+M1 model in terms of F-measure for retrieving opinions about opinion targets. These queries in SLM+IN+M1 model outperformed LM+IN model by 26.6% in F-measure. Similar comparison was made between the performance of SQ3, SQ4, SQ5 of models SLM+IN+M1 and SLM+IN+M2. It showed that recall of SLM+IN+M1 model is 16.5% low, but performed 22.2% better in F-measure and 55.6% more in precision than SLM+IN+M2. This is contrasting to subjective sentence extraction results. Although the SLM+IN+M2 model had large coverage of sentences. The extracted sentences had weak subjective clues which just improved its recall. But, SLM+IN+M1 model had strong subjective sentences which improved its precision and F-measure.

We can also observe and derive from table 9 that unstructured queries (UQ) in all the models achieve 100% recall. But, precision levels vary between the models and are less compared to structured queries in the same and across models. This can be attributed to the retrieval of documents containing only query words but not documents which are opinions about opinion targets. This clearly shows the need for structured queries to achieve better performance.

Similar analysis is done for Table 10 which uses Jelinek-Mercer smoothing. We can observe that SQ3, SQ4 and SQ5 of SLM+IN+M1 model outperforms other methods and queries in terms of F-measure. These queries in SLM+IN+M1 model perform 30.8% better compared to LM+IN model in F-measure though its recall is 5.8% low. This is observed as LM+IN does not have any constraints in sentence selection, while SLM+IN+M1 have only subjective sentences. There is same difference observed as in Dirichlet smoothing when compared between SLM+IN+M1 and SLM+IN+M2.

Different smoothing methods did not show much difference in best performing queries, although, they had minor differences in queries like SQ2.

From the Figure 3 and Figure 4 we can observe that opinion query SQ5 in SLM+IN+M1 model performs better than other queries in retrieving relevant document first for different smoothing techniques. This shows the correlation between the F-measure, as we saw that SQ5 outperforms other queries in different models. In conclusion, we can say that SQ5 of SLM+IN+M1 can be used to mine opinions to achieve decent accuracy if the resources in the language are less.

8. CONCLUSION AND FUTURE WORK

In this paper, we treated opinion mining in resource scarce languages as a retrieval problem by leveraging comparable corpora. Adjectives, verbs and NE's depicted as opinions on opinion targets are extracted from resource rich language like English to mine resource scarce languages. Structured opinion queries formed using opinions and opinion targets are retrieved using LM, LM with IN, SLM and SLM with IN having different smoothing techniques to confirm their presence in documents. We found that SLM with IN performs better for opinion mining. In Future, more complex queries are used which improves F-measure. Also, language independent techniques needs to be explored as current technique depends on dictionaries for translation and transliteration.

9. REFERENCES

[1] A Balahur, R. S. Rethinking sentiment analysis in news: from theory to practice and back. In *In Workshop on Opinion Mining and Sentiment Analysis* (2009).

[2] C. Banea, R. M. J. W., and Hassan., S. Multilingual subjectivity analysis using machine translation. In *EMNLP* (2008).

[3] Donald Metzler, W. B. C. Combining the language model and inference network approaches to retrieval.

[4] H. Takamura, T. I., and Okumura, M. Latent variable models for semantic orientations of phrases. In *11th Meeting of the European Chapter of the Association for Computational Linguistics.* (2006).

[5] J. Wiebe, T. W., and Bell, M. Identifying collocations for recognizing opinions. In *In ACL Workshop on Collocation: Computational Extraction, Analysis, and Exploitation.* (2001).

[6] John, L., and Zhai., C. Document language models, query models, and risk minimization for information retrieval. In *SIGIR* (2001).

[7] L Jia, C. T. Y., and Zhang, W. Uic at trec 2008 blog track. In *TREC* (2008.).

[8] N. Kando, T. M., and Sakai., T. Introduction to the ntcir-6 special issue. In *ACM TALIP, 7(2)* (2008).

[9] Piyush Arora, A. B., and Varma, V. Hindi subjective lexicon generation using wordnet graph traversal. In *In Proceedings of CICLing.* (2012).

[10] Riloff, E., and Wiebe., J. Learning extraction patterns for subjective expressions. In *EMNLP* (2003), pp. 105–112.

[11] S Gerani, M. J. C., and Crestani, F. Proximity-based opinion retrieval. In *SIGIR* (2010).

[12] Soo-Min Kim, E. H. Automatic detection of opinion bearing words and sentences. In *In IJCNLP* (2005).

[13] Soo-Min Kim, E. H. Extracting opinions expressed in online news media text with opinion holders and topics. In *In Proceedings of the Workshop on Sentiment and Subjectivity in Text at COLING-ACL.* (2006).

[14] Soo-Min Kim, E. H. Identifying and analyzing judgment opinions. In *In Proceedings of HLT/NAACL* (2006).

[15] Surya Ganesh, S. H. P. P., and Varma, V. Statistical transliteration for cross language information retrieval using hmm alignment model and crf. In *In Workshop on CLIA addressing the Information Need of Multilingual Societies.* (2008).

[16] T. Strohman, D. M. H. T., and Croft, W. B. Indri: A language model-based serach engine for complex queries. In *International Conference on Intelligence Analysis.* (2004).

[17] Turtle, H., and Croft, W. B. Evaluation of an inference network based retrieval model.

[18] Vincent Ng, S. D., and Arifin., S. M. N. Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews. In *COLING/ACL, 611–618.* (2006).

[19] Y. Hu, J. D. X. C. B. P., and Lu., R. A new method for sentiment classification in text retrieval. In *IJCNLP, 1-9.* (2005).

[20] Y. Suzuki, H. T., and Okumura., M. Application of semi-supervised learning to evaluative expression classification. In *7th International Conference on Intelligent Text Processing and Computational Linguistics.* (2006).

[21] Yang, K. Widit trec blog track:leveraging multiple sources of opinion evidence. In *TREC* (2008).

[22] Zagibalov, T., and Carroll., J. Automatic seed word selection for unsupervised sentiment classification of chinese text. In *Conference on Computational Linguistics.* (2008).

[23] Zhai, C., and Lafferty, J. A study of smoothing methods for language models applied to information retrieval. 179–214.