ISSN 2007-9737

SNEIT: Salient Named Entity Identification in Tweets

Priya Radhakrishnan, Ganesh Jawahar, Manish Gupta, Vasudeva Varma

IIIT Hyderabad,
India

priya.r@research.iiit.ac.in, ganesh.jawahar@research.iiit.ac.in, manish.gupta@iiit.ac.in, vv@iiit.ac.in

Abstract. Social media is a rich source of information and opinion, with exponential data growth rate. However social media posts are difficult to analyze since they are brief, unstructured and noisy. Interestingly, many social media posts are about an entity or entities. Understanding which entity is central (Salient Entity) to a post, helps better analyze the post. In this paper we propose a model that aids in such analysis by identifying the Salient Entity in a social media post, tweets in particular. We present a supervised machine-learning model, to identify Salient Entity in a tweet and propose that the tweet is most likely about that particular entity. We have used the premise that, when an image accompanies a text, the text most likely is about the entity in that image, to build a dataset of tweets and salient entities. We trained our model using this dataset. Note that this does not restrict the applicability of our model in any way. We use tweets with images only to obtain objective ground truth data, while features for the model are derived from tweet text. Our experiments show that the model identifies Salient Named Entity with an F-measure of 0.63. We show the effectiveness of the proposed model for tweet-filtering and salience identification tasks. We have made the human annotated dataset and the source code of this model publicly available.

Keywords. Entity salience, named entity recognition, semantic search, named entity extraction.

1 Introduction

Social media content is growing rapidly. On an average, around 7600 messages are posted on Twitter every second [1]. As the content represents valuable public opinion, analyzing the same can provide important insights. However analyzing social media content is difficult. This is mostly due to the high volume and noisy and unstructured nature of the content. Analyzing social media content and identifying the salient information in it, is thus a challenging problem which we address in this paper. Here we present a method to identify salient information in tweets. Applications like online reputation management [14], social media summarization [24], identifying newsmakers [28] and tweet filtering [30], can benefit from identifying salient information in social media content.

When a text, be it a news paper article or a social media post, is accompanied by an image, it is intuitive to think that the text talks about the image. We use this intuition to identify Salient Named Entity (**SNE**) from among the set of named entities mentioned in the text. This intuition has been discussed in detail by Deschacht *et al.* [12]. Based on this intuition, we define salient entity of a tweet as "the entity present in the image accompanying the tweet." The definition is further discussed in detail in Section 3.2.

In tweets containing entity mentions, there are on an average 2.6 entity mentions per tweet, says Liu *et al.* [22] in an extensive study of evolution of Twitter ecosystem. This necessitates method to identify the salient entity mention from other entity mentions. We first extract all the named entities mentioned in the tweet and then determine which Named Entity (NE) is salient based on the features of the tweet. This approach is in line with the salient entity identification approaches [17] and [19].

Towards creating a salient entity identifier, we create a labeled dataset of tweets and their salient entities. A tweet with an associated image is called

[1] According to `http://www.internetlivestats.com/one-second/` as retrieved on Apr. 9, 2017.

imaged tweet [5]. We use imaged tweets to create the labeled dataset. Tweets are annotated with entity in the image as its SNE and Wikipedia page (if present) that describes the SNE. We train a salient entity identifier using features derived from processing the tweet text. We do not use any feature of the image. So tweets for which SNE should be identified using our model, need not be imaged tweets.

Our experiments show that the proposed method *can identify SNEs in tweets with an F-measure of* 0.63. We evaluate our system (SNEIT) using two methods. First is an intrinsic evaluation using the RepLab 2013 filtering task, where we observe that use of SNEs improves filtering task performance. The proposed method provides results better than the median of the submissions for the RepLab 2013 tweet filtering task. Second we compare the SNEIT performance with that of three state-of-the-art salient entity detection systems [17, 19, 25]. SNEIT system outperforms the first two systems achieving 1% and 3% overall F-measure improvement and achieves F-measure comparable with that of the third system.

The main contributions of this paper are as follows.

1. We model the task of salient named entity identification as a supervised machine learning task, and achieve an F-measure of 0.63.
2. We evaluate our system with a standard tweet filtering task dataset as well as with other salient entity detection systems. Our method performs better than median accuracy of the tweet filtering task submissions, and outperforms two of the three salient entity detection methods.
3. We publicly release the human-annotated dataset consisting of 3646 tweets, their SNEs and the Wikipedia articles they map to[2].

This paper begins with a survey of related research in Section 2 and presents the approach in Section 3. The generated dataset, experimental results and comparison with baseline systems are presented in Section 4. Section 5 presents analysis and design decisions followed by conclusion and future work in Section 6.

2 Related Work

Salience: Salience as a concept has been defined as aboutness, most noticeable, conspicuous and prominence in dictionaries. Gamon *et al.* [17] observe that salience is a function of the structure of text and the intention of the author.Boguraev *et al.* [4] mention that "Salience is a measure of the relative prominence of objects in discourse". Thus we can see that in our context of set of NEs in a tweet, a SNE is the NE that is central to the tweet.

Salience as a concept has received little attention in information retrieval and knowledge discovery research [26]. Two recent works on identifying SNE are that of Gamon *et al.* [17] and Gillick and Dunietz [19]. Gamon *et al.* assign salience scores to entities based on their centrality to the web-page. They assess the relevance and salience of an entity with respect to a web-page by mining a web search log and click log from a commercial search engine. Gillick and Dunietz [19] automatically generate salience labels for an existing corpus of document/abstract pairs using the assumption that the salient entities will be mentioned in the abstract. Both the works focus on structured documents, for which high-quality NLP tools are available, whereas our work focuses on tweets.

Our method of creating training data for salient entity is similar to that of Gillick and Dunietz [19]. While Gillick and Dunietz create training data by identifying and aligning SNE across document and abstract, we do so across tweet and image. However annotation is done manually in our case whereas it is done automatically in the Gillick and Dunietz method. Manual annotation was also done by Deschacht *et al.* [13] in their study of salience of entities to predict the probability of salient entity appearing in the image accompanying the text. To test their system, they annotated 900 image-text pairs of the Yahoo! News dataset. For every text-image pair one human annotator selected the entities that appeared both in the text and in

[2]Source code and dataset are available at `https://github.com/priyaradhakrishnan0/SNEIT`.

the image and sorted the entities based on their perceived importance in the image.

Imaged tweets: Chen *et al.* [5] studied why tweeters prefer imaged tweets over text-only tweets. They note that the preference of posting imaged tweets or text-only tweets correlates with the content. For example, advertisement tweets tend to include a product image to make it more informative; whereas tweets about the everyday routine or social babble are prone to be text-only. Thus imaged tweets are more likely to contain entities.

Recognizing NEs in tweets: The task of analyzing a tweet to recognize and extract NEs from it, is challenging in many ways. Tweets are short in length, resulting in reduced context and ambiguity. They are also dynamic, context-dependent and less grammatical than longer posts. The use of unorthodox capitalization leads to significant drop of recall in Named Entity Recognition (NER) on tweets compared to conventional text [11]. As tweet content is not curated (as done by editors in news-wire), tweets are largely superfluous, impacting information extraction performance. Ritter *et al.* [27] model the tweet NER problem as a problem of segmenting tweets and classifying the segments into entity types. They propose a distantly supervised approach based on LabeledLDA and show that it significantly outperforms generic NER [16]. Developed in parallel to this work, Gimpel *et al.* [20] built a POS tagger for tweets using 20 coarse-grained tags. In a survey of NERs for tweets, Derczynski *et al.* [11] observed that Twitter-specific NER is difficult due to lack of sufficient context and good human-annotated corpus covering distinct named entity types. They also reported that the highest achieved F1 score on NER in tweets is only 0.40 [3].

To disambiguate entities in tweets, Meij *et al.* [25] proposed 'whole tweet entity linking', where they identify concepts in tweets and link them to Wikipedia articles using a supervised learner. Their dataset is a concept disambiguated tweet dataset, as it contains tweets and their concepts which are Wikipedia articles. While Meij *et al.* aim at finding a ranked list of concepts of a tweet, we aim to obtain an SNE for a tweet. The concept has to be present in Wikipedia, while SNE need not necessarily have a Wikipedia page.

Research on salience of entities was rekindled in recent years by works of [32] and [31]; we extend it to tweets. Our approach is to first identify NEs in a tweet and then ascertain the salience of the NEs. We use the NERs described in this section for identifying NEs and use image in the imaged tweets to ascertain salience.

3 Approach

Fig. 1 shows the work flow of the proposed SNEIT system which is simple yet efficient. Given a tweet, the system identifies the NEs in it by using a set of NERs, as the first step. These NEs are candidate SNEs for the tweet. This step is explained in Section 3.1. To select the SNE from candidate SNEs, we train a sequence learner on a dataset of tweets annotated with their salient entities. Creation of this dataset is explained in Section 3.2. The learner associates a salience score for each candidate SNE, which is the probability of a candidate NE being a SNE. This task is modeled as a supervised sequence labeling task. The sequence labeler and features used in it are explained in Section 3.3.

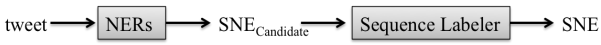

Fig. 1. SNEIT System work flow

We start with definitions of entity and salient entity. While there are many definitions of an entity in literature, for the purpose of this paper we use the working definition by Gamon *et al.* [17]. We consider something an entity if it has or reasonably could have a Wikipedia page associated with it. This would include people, places, companies, events, concepts and famous dates. There are varying definitions for salient entity too (as discussed in Section 2). For our purpose, we define as SNE as that entity, whose image the Twitter user attaches to the tweet.

[3] NER performance over the golden part of the UMBC dataset.

3.1 Identifying Candidates

First step in analyzing tweets as in most IE pipelines, is identification of NEs. Identifying NEs from tweets [16, 20, 27] is a well researched problem. NE identification task poses many challenges, especially for tweets [11]. A NER may identify none, one, or more than one NE in a tweet. In rare cases the NEs identified are completely incorrect. To overcome these and ensure better recall of NEs, Bansal *et al.* [3] combine the outputs from multiple NERs. Hence, for identifying NEs, we use three NERs, which are reported to give higher F1 scores by Derczynski *et al.* [11]. They are Ritter *et al.* [27], Gimpel *et al.* [20] and Finkel *et al.* [16]. All of them are open source NERs and are used 'as is' in SNEIT system. Their results are merged to obtain the candidate SNEs ($\text{SNE}_{candidate}$).

3.2 Creating Dataset

To create a dataset of tweets annotated with their SNEs, first we identify the SNE candidates and then choose the SNE from these candidates.

Identifying SNE Candidates: Twitter introduced in-stream images feature in late 2013. Ever since, imaged tweets are reported to have higher than average engagement, with 35% higher chance of retweet [4], than a text-only tweet. Tweeters include an image in the tweet capturing the central idea of the tweet, to increase engagement (and retweets) for the tweet. Hence our assumption is that the salient entities are represented in the image accompanying the tweet. Deschacht *et al.* [12] have proved that the chances of an entity appearing in the image and the accompanying text is very high when the entity is salient. So we identify the salient entities in tweet as the entity appearing in the image associated with the tweet.

Labeling SNE: Annotating a NE to indicate if it is salient to the tweet is a difficult task as annotators have different perspectives of salience. This is due to personal relevance or bias. For example, consider the tweet text (Fig 2 without the accompanying image)

[4] https://blog.twitter.com/2014/what-fuels-a-tweets-engagement.

Fig. 2. SNE of a Tweet

"Google Executive Dan Friedenburg Dies in Everest Avalanche Nepal Earthquake #google #techtalent sorry for your loss"

Annotator A1, who is a technology enthusiast, may mark "Dan Fredinburg" and "Google" as SNEs. Annotator A2, interested in mountaineering, may mark "Everest" as SNE. Annotator A3, interested in current affairs, may mark "Dan Fredinburg" and "Nepal" as SNEs. Thus SNE identification may vary with perspectives and/or interests of the annotator. This variance is captured as personal relevance. SNE annotation could also be affected by entity importance. Entity importance refers to influence or substantiveness of an entity outside of the scope of the paper. For example, although Barack Obama is a very important entity, he can be peripheral to some news stories.

These annotating difficulties can be overcome by using image as the evidence. For example, we look at the accompanying image of the tweet example (shown in Fig 2). We find that among the four NEs identified in the tweet, the only NE recognizable in the image is "Dan Fredinburg". We annotate this NE as the SNE of the tweet.

In their analysis of cross-media entity recognition in nearly parallel visual and textual documents, Deschacht *et al.* [13] prove that the ratio of entities in the text to that present in the image is 22.96%. In our dataset, we use only imaged tweets. Thus our annotators choose only those NEs that are present in the image as SNE of the tweet. (Please

refer to Section 4.1 for more information about the annotators, and the dataset creation.)

Table 1. Labeler Features

Feature Type	Value
POS tag	NNS, VBP, PRP
Chunk POS tag	B-NP, I-NP, B-VP, I-VP
Entity tag	B-ENTITY, I-ENTITY

3.3 Training the SNE Identifier

The task of identifying salience of $\text{SNE}_{candidate}$ considers context (neighboring words) of the $\text{SNE}_{candidate}$. Hence it is modeled as a sequence labeling problem. This is done using Conditional Random Fields (CRF)[5], a popular machine learning algorithm that assigns tags to token sequences [21]. CRF considers the features of current and neighboring tokens for tag assignment. In our implementation, we use 15 features of a token. The features are based on word characteristics and labeler features (POS tag, Chunk POS tag and Entity tag) of the token, as shown in Table 1. Twitter NLP toolkit[6] is used to tokenize the tweets and extract POS tag and Chunk POS tag features of tokens, whereas Entity tag is obtained from NERs (discussed in Section 3.1). Based on features of the token and features of preceding and succeeding tokens, CRF identifies the SNE. We refer to this as SNE Identifier.

We use standard BIO encoding [8] for tags. It subdivides the tags as begin-of-entity (B-), continuation-of-entity (I-) and Non entity (O-). Thus SNE Identifier encodes tokens with tags B-SNE, I-SNE and O-SNE.

The 15 features used by the SNE Identifier that gave best performance along with their respective weights are shown in Table 2. SNE Identifier was trained using tweets of CWC15 dataset (explained in Section 4.1). 1200 tweets (one-third of CWC15) was used for training the weights of features, as development set. Remaining 2400 (two-third of CWC15) was used in the 10-cross validation of the SNE Identifier, for training and validation.

[5]We used the implementation `https://github.com/tpeng/python-crfsuite`.

[6]`https://github.com/aritter/twitter_nlp`.

Table 2. SNE Identifier Features

Feature Type	Feature	Weight
Word	Lower : Change the case of word to lower case	3
Word	Upper : Change the case of word to upper case	1
Word	isTitle : Is word a in title format	1
Word	isDigit : Does word contain only numbers	1
Word	isUpper : Is the word in upper case	2
Word	isFirstCharHash : True if first character is #	3
Word	isFirstCharHashOrAt : True if first character is # or @	4
Word	isFirstCharCaps : True if first character is in uppercase	3
POS	Postag : POS tag of the token	4
POS	isStartsWithNN : True if POS tag starts with NN	2
POS	isStartsWithNNorPR : True if POS tag starts with NN or PR	1
Chunk	Chunk : Chunk POS tag	1
Chunk	isChunkNP : True if chunk POS tag is B-NP or I-NP	3
Entity	True if token is recognized as entity by NERs	4
Entity	isEntity : True if entity tag is B-ENTITY or I-ENTITY	1

4 Experiments

In this section we explain the dataset used in the experiments, the experiments conducted and how the results compare with those of baseline systems.

4.1 Dataset

To train the SNE identifier, we create a human-annotated dataset by annotating tweets with its SNE. As we consider the presence of a $\text{SNE}_{candidate}$ in the accompanying image as the evidence of salience of the $\text{SNE}_{candidate}$, our annotation guidelines (See source code) instruct the annotators to select the $\text{SNE}_{candidate}$ as the SNE only if it is present in the image. The human-annotated tweet dataset is referred to as CWC15 dataset henceforth. In this section, we explain the construction of CWC15 dataset detailing the motivation, tweet collection process, manual annotation, inter-annotator agreement scores on methods, domains and dataset statistics.

Motivation for new dataset: Existing Dataset [25] contain tweet and their concepts as Wikipedia articles. Though this originally contained 502 tweets, we could get only 363 tweets due to some twitter accounts getting banned and tweet deletions. As this size was not sufficient for creating a learner and as we are interested in annotating tweet with only SNEs (not all entities or NEs), we annotated tweets to create the CWC15 dataset. Besides the CWC15 dataset, we also use RepLab[7] task and dataset to evaluate our system. The task [2] tries to analyze tweets for potential mention of entities, filtering those tweets that refer to an entity.

CWC15 Dataset: As the annotator uses the image to validate the choice of SNE, we use only imaged tweets in constructing the CWC15 dataset. The textual content of the tweet is used for SNE Identification. The image is used only for helping the annotators select the SNE. No feature of the image is used in the SNEIT model, similar to the approach of Deschacht *et al.* [13].

Preprocessing: The tweets for CWC15 dataset creation were collected during a popular sporting event, the ICC Cricket World Cup 2015[8] (this choice is explained in '*Choice of domain*' paragraph in this section). Hashtags of the quarter final matches were used as queries. We collected tweets that contain at-least one image, is in English language and is not a re-tweet. A total of 10,938 tweets were collected.

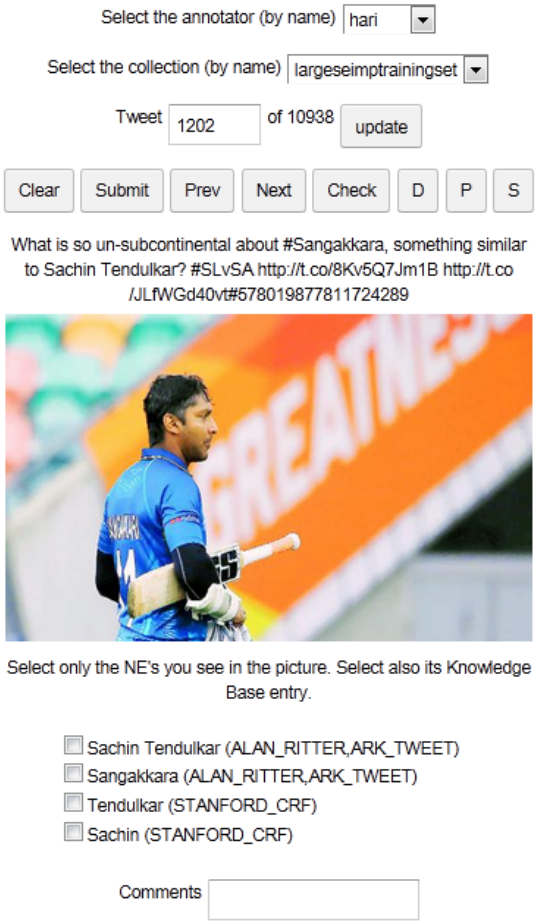

Fig. 3. Annotation Interface

Manual Annotation: We asked six volunteers to manually annotate upto 2000 tweets each, so that every tweet is annotated by one human annotator. The volunteers were graduate students in the age group of 20 to 30 with interest in the game of cricket. The group had three male and three female members. The annotators were presented with an annotation interface shown in Fig 3, which has the tweet, its image and the $\text{SNE}_{candidate}$. The annotation guidelines specified that the annotator should choose an $\text{SNE}_{candidate}$ as being salient to the tweet only if it is present in the image. On choosing

[7] Evaluation exercise for Online Reputation Management systems `http://nlp.uned.es/replab2013/`.

[8] `http://www.icc-cricket.com/cricket-world-cup`.

an $\text{SNE}_{candidate}$, the probable Wikipedia articles of that $\text{SNE}_{candidate}$ were presented and the annotator chose the Wikipedia article that best describes the $\text{SNE}_{candidate}$. Annotators could also choose more than one $\text{SNE}_{candidate}$ in cases where multiple SNEs exist. The annotator was asked to mark a tweet as **D** for Duplicate (Tweet text and/or image repeats in the dataset) or **P** for Pointless (Pointless conversation or Image not containing any NE or Image containing non English text or Advertisement) or **S** for Sarcastic (tweet text is sarcastic with respect to image) or **N** for Not Annotatable ($\text{SNE}_{candidate}$ is not presented or invalid) when the respective condition was satisfied. (Please refer section 5.1 for more details on annotation trivia.)

Inter Annotator Agreement - *Use of image:* The presence of image helps annotator to validate the choice of SNE. In order to measure this help, we created a smaller dataset by randomly sampling fifty tweets from CWC15 dataset. We asked four annotators to annotate this smaller dataset. The annotators were first presented with the 50 tweets without their images. After the first 50 tweets are annotated, they were presented with the same 50 tweets, with the images. The inter annotator agreement measured using Cohen kappa [7], is presented in Table 3. We report the annotator agreement on SNEs and their Wikipedia articles.

Table 3. Inter annotator agreement scores: Use of Image

Measure	With Image	Without Image
SNE	**0.67**	0.52
SNE and Wikipedia article	0.53	0.35

Table 4. Inter annotator agreement scores: Choice of domain

Measure	Overall	Sports	Movie	Product
SNE	0.73	**0.75**	0.75	0.65
SNE and Wikipedia article	0.60	0.62	0.61	0.51

The annotation round with images has better agreement score of 0.67 ('fair agreement' per [23]) compared to the round not using image. The higher inter annotator agreement has motivated us to annotate every tweet in CWC15 dataset by only one annotator, thereby quickening the annotation process. One human annotator annotated SNE across text-image pairs in the experiments of Deschacht*et al.* [13] too.

Inter Annotator Agreement - *Choice of domain*: The annotator's background knowledge and interest in the domain of the (tweet) text has a positive effect on the quality of the annotation. As the task was selecting the salient one of the $\text{SNE}_{candidate}$, background knowledge of the $\text{SNE}_{candidate}$ was needed. Considering our annotator's interest we chose a popular sporting event (cricket world cup), a popular entertainment event (annual film award) and a much awaited product release (apple watch release). Considering the age group of our annotators and their background, the popularity of these events ensured annotator's domain knowledge and hence a high quality of the annotation.

To measure inter-annotator agreement of domains, we created a smaller dataset with tweets from the three domains namely 'Product', 'Sport' and 'Movie' using keywords AppleWatch, SAvsNZ and NationalAwards respectively. We randomly sampled 20 tweets satisfying the tweet selection conditions (discussed in preprocessing paragraph), from the three domains and asked two annotators to annotate the 60 tweet corpus. The inter annotator agreement measured using Cohen Kappa is presented in Table 4. The Cohen Kappa scores show 'fair agreement' [23] and is comparatively higher for sports domain in SNE annotation. Hence we chose sports as domain.

CWC15 Dataset Statistics: A total of 3646 tweets have a $\text{SNE}_{candidate}$ marked as salient. These 3646 tweets form the CWC15 dataset[9] and is used to train the SNE identifier. Out of 3646 tweets, for 1812 tweets a Wikipedia title representing that SNE is chosen by annotator.

[9]We publicize two tweet corpora: (i) CWC15 dataset containing 3646 tweets, and (ii) the whole annotated corpus, containing 10,938 tweets.

This subset is used to train the Tweet Linker (see source code). On analyzing the annotated tweets

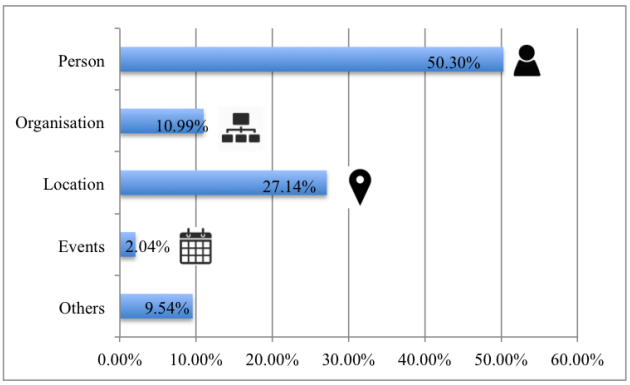

Fig. 4. NE Type distribution in CWC15

we find that a salient mention exists only for 33% of the total annotated tweets and a Wikipedia entity is identified only for 16.57% of them. Out of the 10,938 tweets annotated, 8425 have a $\text{SNE}_{candidate}$ and 3646 have a SNE, which put the probability of $\text{SNE}_{candidate}$ becoming SNE at 0.43. Fig 4 presents an analysis of the types of SNEs that are identified by the annotators. Half of SNEs is of the type persons, 27% is locations, 11% is organizations and 2% is events. Others including 'none' made up 9.5%. Thus SNEs of CWC15 dataset are spread across the entity types while the SNEs of comparable systems [19] are limited to type persons [10].

Table 5. NER Comparison

NER	Average NEs per tweet
Ritter *et al.* [27]	2.22
Gimpel *et al.* [20]	2.21
Finkel *et al.* [16]	0.89
Combined	**3.76**

4.2 Experimental Results

In this section we present the results and compare and discuss the design options considered.

[10] http://googleresearch.blogspot.in/2014/08/teaching-machines-to-read-between-lines.html.

Table 6. SNE Identification

SNE tag	P	R	F
B-SNE	0.74	0.55	**0.63**
I-SNE	0.59	0.38	0.46
O-SNE	0.93	0.97	0.95

Candidate SNE Generation: NERs are used to identify the NE mentions in the tweet. This has two possible outcomes. In the first case, the NER does not identify any NE mention in the tweet. We consider this tweet as a tweet for which SNE does not exist or cannot be determined (These tweets are marked P and N during annotation). In the second case, the NER finds one or more NE mentions in a tweet. These NE mentions are the $\text{SNE}_{candidate}$ of the tweet. We use three NERs in our experiments. The performance of the NERs on CWC15 dataset is shown in Table 5. Though we find that NERs of [27] and [20] give almost equal number of NE mentions for a tweet, the NE mentions of a tweet are not always the same. This is the case with [16] NER too. So, for our experiments, we combine the results of three NERs to get a super-set of NE mentions. The combined result is found to be higher than that of individual NERs and is presented in the last row of Table 5.

The results from the three NERs could be merged in two ways. First is by taking a union of all the results. For example, {"ind vs aus"} U {"ind vs aus quarter final","ind vs aus"} = {"ind vs aus", "ind vs aus quarter final"}. The other is to merge the NER results, i.e. if a NE mention is a sub-string of another NE mention as in {"ind vs aus"} U {"ind vs aus quarter final"} = {"ind vs aus quarter final"}. We choose the first as it ensures that a string having multiple NE mentions do not eliminate the string with a single NE mention. This method of choosing the mention is advocated by Gattani *et al.* [18].

SNE Identification: The SNE identifier performance in terms of Precision (**P**), Recall (**R**) and F-measure (**F**) in tagging a token with B-SNE, I-SNE and O-SNE tags is presented in Table 6. The results are macro-averaged scores from a 10-fold cross-validation on test set using all the features. The CRF labeler uses features of tokens preceding

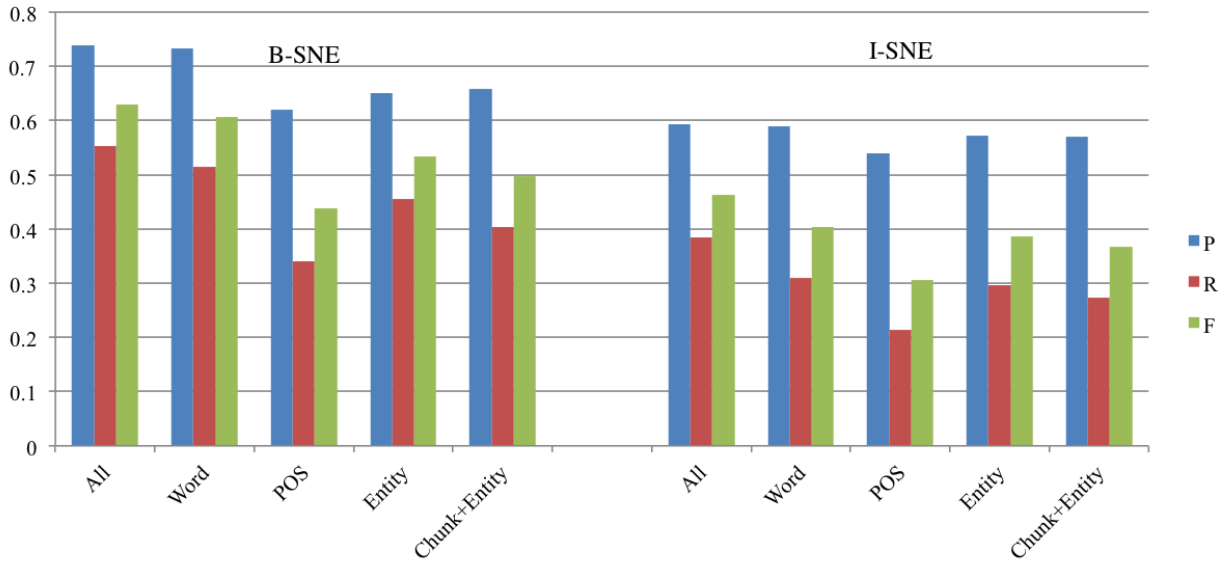

Fig. 5. Performance in identifying B-SNE and I-SNE

and succeeding the given token. The number of preceding and succeeding tokens used is referred as context. In our experiments, we found **context of 2** to give best results. Tag B-SNE is assigned with a F-measure of **0.63**. Tagging a token as B-SNE also accounts for identifying unigram SNEs and partial SNE matches [8].

SNE Identification - *Contributing features:* We analyze the performance of individual feature types used in the SNE identification. Fig 5 shows the graph plotted for the feature types discussed in Table 2. Here we see that word features give better P, R and F in assigning both B-SNE and I-SNE tags. The combination of all the features (denoted by 'All' in the Fig 5) gave best results.

4.3 Comparison with SNE Detection Baseline Approaches

We compare the performance of SNEIT system in two ways. First we evaluate it in a filtering application, using a standard dataset for filtering tweets containing relevant NEs. Second we compare it with three baseline systems [17], [19], [25] for identifying salient entity. Two of these evaluations namely filtering task and comparison with [25], require the NE to be a Wikipedia entity. So we first explain linking of SNE to its Wikipedia entity, followed by evaluating the performance.

Linking: The SNE identified is linked to a Wikipedia entity (we use Wikipedia page title and Wikipedia entity interchangeably). The SNE might represent multiple entities in Wikipedia. In order to disambiguate among the many entities, we use a supervised binary classifier, which we call Tweet Linker. Tweet Linker is an Entity Linking system [29, 6]. More details of Tweet Linker, like features and Wikipedia article sources used in it, are provided in shared source code. The Tweet Linker chooses a Wikipedia entity to link and classifies the choice as correct or wrong. The classifier is implemented using three algorithms. Their linking performance is presented in Table 7. Support Vector Machines (SVM) classifier performs well. Random Forest (RF) classifier gave the best precision, as also seen in the baseline work of Meij *et al.* [25] and in Yamada *et al.* [33, 34]. So a RF classifier is used as the linker. As Tweet Linker uses features specific to SNEs in tweet, we prefer using it over popular linkers like DBPedia Spotlight [9] or TagMe [15]. In the dataset used to train

Tweet Linker, data samples where SNE has a valid Wikipedia entity are positive samples, denoted as class '1'. Data samples where SNE does not have a valid Wikipedia entity are negative samples, denoted as class '0'. Class '1' samples are less (28% of the total samples) in the dataset. Performance values for classes '1' and '0' is presented in Table 9. The values in Table 7 are for '1' class alone.

Table 7. Tweet Linker Performance with different classifiers

Classifier	P	R	F
SVM	0.62	0.09	0.16
Adaptive Boosting	0.5	0.07	0.12
RF	**0.75**	0.05	0.10
NaïveBayes	0.27	0.05	0.09

Filtering task - *Dataset*: We evaluate the SNEIT system using RepLab Filtering task. The RepLab tasks [2] is about monitoring the reputation of entities (companies, organizations, celebrities, etc.) on Twitter. Human analysts were asked to identify the potential mentions from the stream of tweets and map them to the corresponding entities. Thus the RepLab dataset contains manual annotations of tweets with entities, annotated with two possible values: related and unrelated. The filtering task is about determining which tweets are related to the entity and which are not. In this evaluation, rather than filtering all tweets containing the entity-of-interest, we filter tweets where entity-of-interest is the SNE, as related. We use the RepLab test collection for evaluation, while CWC15 dataset is used for training the SNEIT system.

Filtering task - *Performance*: For a test tweet, we determine the SNE using SNEIT system. The SNE is then linked to Wikipedia entities. If the linked Wikipedia entity is the related entity we count it as an accurate identification. Accordingly we report the Accuracy (Acc), Reliability (R), Sensitivity (S) and F-measure (F) in lines of RepLab filtering task specifications [2], in Table 8. In Replab filtering task, median of Acc, R, S and F are 0.85, 0.49, 0.32 and 0.27. The accuracy of SNEIT system is lower than median accuracy of the replab participant systems. This is because we count an identified entity as accurate only when it is SNE. This rigor leads to higher R and S values for SNEIT than the median R and S values of Replab competition, which also reflects in higher F values of SNEIT. Thus we see that the SNEIT system performs better than many RepLab participant systems. In Table 8 column 'Number of tweets' gives number of tweets mentioning entity-of-interest in the Replab test collection and column 'Tweets with SNE' gives number of tweets where the entity-of-interest is identified as SNE by SNEIT system. This proves that SNE is better candidate for filtering task.

Table 8. Evaluation with RepLab 2013 dataset

RepLab Query	Number of tweets	Tweets with SNE	Acc	R	S	F
Porsche	779	208	0.46	0.86	0.44	0.58
Lexus	809	141	0.58	0.55	0.54	0.55
Ferrari	800	505	0.13	0.92	0.12	0.22
Volvo	714	201	0.48	1.00	0.50	0.66
Kia	804	140	0.10	0.86	0.22	0.35
Ford	782	194	0.36	0.56	0.44	0.66
Fiat	752	153	0.43	0.92	0.45	0.60
Barclays	747	198	0.45	0.99	0.47	0.62
MIT	666	140	0.21	0.86	0.31	0.46
Shakira	944	209	0.72	0.97	0.72	0.82

Baseline System - *Identify salient entity*: In their work on identifying SNE from news articles, Gillick and Dunietz [19] identify SNE with F-measure of 0.62. SNEIT identifies SNE with F-measure of 0.63 on tweets. Similarly Gamon *et al.* [17] identify salient entities in popular web pages (denoted HEAD) with a F-measure of 0.75 and random web pages (denoted TAIL) with a F-measure of 0.64. SNEIT results are in this range as well. We note that these baseline systems identify SNE on news articles and web-pages which have a well structured text. Whereas SNEIT identifies SNE on tweets which is mostly unstructured text.

Baseline System - *Identify concept of tweet*: Meij *et al.* [25] propose 'whole tweet' entity linking in their work, where they link the tweet to the

topic or the entities describing the theme of the tweet. As this leads to identifying tweet's salient entity, we consider this work as third baseline system and evaluate the performance of the SNEIT system against this system.Towards this goal, we have re-implemented the Meij *et al.* system. The comparison is presented in Table 9. We have used the Random Forest (RF) classifier and COMMONNESS concept ranking, which is reported to have produced the best results by Meij *et al.*. The features of Meij *et al.* system used in our re-implementation include N-gram features (LEN, SLINKPROB), Concept features (INLINKS, WLEN, CLEN), N-gram + concept features (SPR, NCT, TCN, TEN, COMMONNESS) and Tweet features (TWCT, TCTW, TETW, TAGDEF, URL). The average precision and F-measure (on CWC15 dataset) of the re-implemented system are 0.85 and 0.87 whereas that of original Meij *et al.* system were 0.57 and 0.48 respectively.

Table 9. Comparison of SNEIT system performance with baseline

Method	Label	P	R	F
Meij et al.	0	0.92	0.97	0.94
	1	**0.22**	**0.09**	**0.13**
SNEIT	0	0.73	0.98	0.84
	1	**0.62**	**0.09**	**0.16**

The performance of SNEIT system in whole tweet linking is 0.62 in terms of precision, while that of baseline is 0.22. The corresponding improvement in F-measure is 0.03. Interestingly there was no difference in recall achieved between SNEIT and Meij *et al.* systems. These results are statistically significant when tested with t-test (1-tail 95% value is 1.645 and 2-tail 95% value is 1.96).

5 Discussion

In this section we analyze the experiments and results followed by a discussion on some of the design decisions in the paper.

5.1 Analysis

Definition: In defining the salient entity identification task for the scope of this paper, we define salient entity of tweet as "Given an option, if the tweet author will attach image of this entity to the tweet, then it is the salient entity of the tweet". This definition satisfies only imaged tweets. We use only imaged tweets in CWC15 creation. So this definition satisfies this paper's purpose. For the text-only tweet (we do not use image), we identify salient entity by the SNEIT model using textual features.

Annotation trivia: A total of 10,938 tweets were manually annotated by volunteers. Out of these, 4272 were marked duplicate(D), 507 were marked as sarcastic(S), 1838 were marked as pointless(P) and 307 were marked N as they had no $\text{SNE}_{candidate}$ identified. For 368 tweets, no Wikipedia entry was identified. After discounting these, we are left with **3646** tweets, which became CWC15 dataset.

Table 10. Common errors made by SNE Identifier

Gold	Identified	Error
B-SNE	I-SNE	0.014
B-SNE	O-SNE	0.447
I-SNE	B-SNE	0.025
I-SNE	O-SNE	0.621
O-SNE	B-SNE	0.025
O-SNE	I-SNE	0.009

Table 11. Performance on $\text{SNE}_{candidate}$ Type

Type of $\text{SNE}_{candidate}$		P	R	F
Person	B-SNE	0.758	0.683	0.715
	I-SNE	0.726	0.692	0.702
Location	B-SNE	0.748	0.569	0.63
	I-SNE	0.405	0.317	0.303
Org. &Others	B-SNE	0.628	0.447	0.513
	I-SNE	0.524	0.483	0.461

Error Analysis: Table 10 lists the common errors made by SNE Identifier. For each case we list the fraction of times the gold tag is misclassified as identified tag, by SNE Identifier. Here we see that false negatives are higher than false positives,

as tags B-SNE and I-SNE get identified as O-SNE. To some extent this could be attributed to the distribution of tags (B-SNE 9.6%, I-SNE 3.5% and O-SNE 86.8%). This also reflects in the higher precision and lower recall of SNE Identifier.

Table 11 compares the SNE Identification performance for different $\text{SNE}_{candidate}$ types like Person, Location, Organization and Others. We see that SNEIT system performs best when $\text{SNE}_{candidate}$ is of type Person.

One of the limitations of SNEIT method is that it finds salient NEs rather than salient entities. This is because we use the presence of entity in image as evidence of salience in the training dataset (CWC15) creation. NEs like 'Barack Obama' or 'Paris' can be recognized in image whereas entities like 'Entropy' or 'Human Rights' will be difficult to be recognized in image.

5.2 Design decisions

SNEIT works on Text-only and Imaged Tweets: In the illustrated tweet example in Section 3.2, we see the difficulty of SNE identification problem, due to the presence of multiple NEs. A quick hack to identify salience is to use image feature (presence of entity in image) as salience detecting feature. But this would make our method applicable to imaged tweets only. To make our method generic and applicable for tweets with and without accompanying images, we used textual features alone in our salience identifier. We derive the right combination of textual features which is able to identify SNE from $\text{SNE}_{candidate}$ with a F measure of 0.63.

NER as $\text{SNE}_{candidate}$ Generator: This paper is based on the premise that salient entity of a tweet can be solely determined by how the entity is presented in the tweet. By assuming the source of salience to be local to a tweet, we limit the search space to those entities in the tweet. Hence, a system that is capable of identifying NEs in the tweet would serve as a candidate generator for a SNE identification system. We test this assumption in the CWC15 dataset. While creating the CWC15 dataset, annotators mark a tweet as N if no $\text{SNE}_{candidate}$ was listed. We counted all the tweets marked N and found it to be 2.8% of total annotated tweets. In other words, in 97% of the annotated tweets, at least one of the salient entities is in the candidate entity set identified by the NER. Therefore it is reasonable to use the NER as a SNE candidate generator. We analyze the performance of the three NERs [16], [20], [27] as $\text{SNE}_{candidate}$ generator, in terms of how $\text{SNE}_{candidate}$ suggested by a NER is selected as SNE (accounting for only unique suggestions by a NER in this calculation). 36% of SNEs are selected by annotators from $\text{SNE}_{candidate}$ suggested by [27] and 40% by [20]. 24% of SNEs were selected from $\text{SNE}_{candidate}$ suggested by [16]. This is in line with results of Table 5 and confirms our choice of merging NER results.

Randomness of Tweet Sample: Tweets containing images are increasing. Zhao *et al.* [35] did a statistical analysis on image and multimedia content in social media. They expect the proportion of imaged tweets to increase. Chen *et al.* [5] note that imaged tweets constitute over 45% of overall traffic in (Chinese) Weibo. With increasing number of imaged tweets, a sample of tweets with only imaged tweets has higher chances of being a random sample. Further, among the 10938 tweets we annotated, 507 were sarcastic tweets. This indicates the eclectic nature of tweet authors. The higher number of sources also increases chances of a random sample.

In their study on a sample of 2000 tweets, Pear Analytics [1] classified 40% as containing "pointless babble", with another 37.55% as merely conversational. They found that only a small fraction contains topics of general interest. In the 10,938 tweets we annotated, we found similar figures with only 33% of the tweets as being useful. The similar distribution of non-informative and informative content across sample containing all tweets and sample containing only image tweets, shows our tweet sample is a random sample.

Choice of Sports Event: Our choice of the sports event is guided by amount of social media content generated by the event and the popularity of that event among our annotators. The latter ensures that the annotators have a good knowledge about the NEs to be annotated. Cricket World Cup 2015 (CWC2015) held in Australia and New Zealand created unprecedented levels of

online and social media interest. As per the official statistics of the CWC 2015 web page [11], there were over 26 million unique visitors to this website making up over 225 million page views.

On the event's Facebook page, 36 million people generated 341 million interactions. On Twitter, the discussion around #cwc15 was huge, with over 8 million tweets sent around the tournament, with over 800 Million live tweet impressions from the group stages. Thus the CWC 2015 was one of the most engaging topics on social media and twitter, producing significant amounts of data. The event was popular among our annotators too convincing us to create a dataset around this topic.

6 Conclusion

We started with the aim of filtering salient content from social media text. We saw the challenges in defining and identifying the salient entity from the entities in social media content. We built a supervised machine learning model to identify SNEs in tweets. Our model has a rich set of textual features, which have been tuned to identify SNE in tweets with an F-measure of 0.63. We find that SNE when used in filtering application gives better result than the regular named entities. Our method outperformed two of the three baseline methods.

We have made the annotated CWC15 dataset, including 507 sarcastic tweets and the source code of the SNEIT system, publicly available[12]. While salient entity datasets have been released in the recent times for web pages and documents, CWC15 dataset is the first publicly available SNE dataset for tweets. CWC15 dataset contains four types of entities. As can be seen from our analysis, accuracy of entity salience identification varies based on the type of entity. Entity type person gives higher accuracy than other entity types. In our analysis we also found that word features outperformed NER features like Chunk, POS and entity. We plan to jointly model SNE identification and NER in tweets, in our future work.

Research in entity salience is still in its infancy. Salient entities improve the richness of the semantic network in "Web of Thing" paradigm [10]. Our experiments show that only 43% of named entities in a tweet are salient, thereby making it important to filter them from non-salient entities. CWC15 dataset could also be very interesting in a RepLab-like setting, where brands (as entities) are interested in checking their reputation in social media and in particular in images circulating in social media. We plan to use CWC15 dataset in the task of labeling content of an image in a tweet based on the text of a tweet. We also plan to enhance SNEIT to identify SNEs in Facebook posts.

Acknowledgments

We would like to thank Edgar Meij, Johannes Leveling, Johannes Hoffart, Niloy Ganguly, Partha Pratim Talukdar and Aruna Chaluvadi for their advice and contributions to the review process. We are also indebted to our annotators for their help in creating CWC15 dataset.

References

1. **Amigo, E., Carrillo de Albornoz, J., Chugur, I., Corujo, A., Gonzalo, J., Martin, T., Meij, E., de Rijke, M., & Spina, D. (2013).** Overview of RepLab 2013: Evaluating Online Reputation Monitoring Systems. *Proceedings of the Fourth International Conference of the CLEF initiative*, pp. 333–352.
2. **Bansal, R., Panem, S., Radhakrishnan, P., Gupta, M., & Varma, V. (2014).** Linking entities in #microposts. *4th Workshop on Making Sense of Microposts (#Microposts2014)*.
3. **Boguraev, B. & Kennedy, C. (1997).** Salience-based content characterisation of text documents. *Advances in Automatic Text Summarization*, The MIT Press, pp. 2–9.
4. **Chen, T., Lu, D., Kan, M., & Cui, P. (2013).** Understanding and classifying image tweets. *Proceedings of the 21st ACM international conference on Multimedia*, MM '13, ACM, New York, USA, pp. 780–784.

[11] http://www.icc-cricket.com/cricket-world-cup/news/2015/media-releases/87040/icc-cricket-world-cup-2015-following-soars-to-record-levels-onlin.

[12] Source code and dataset are available at `https://github.com/priyaradhakrishnan0/SNEIT`.

5. **Chisholm, A. & Hachey, B. (2015).** Entity disambiguation with web links. *Transactions of the Association for Computational Linguistics*, Vol. 3, pp. 145–156.
6. **Cohen, J. (1960).** A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, Vol. 20, No. 1, pp. 37–46.
7. **Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011).** Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, Vol. 12, pp. 2493–2537.
8. **Daiber, J., Jakob, M., Hokamp, C., & Mendes, P. N. (2013).** Improving efficiency and accuracy in multilingual entity extraction. *Proceedings of the 9th International Conference on Semantic Systems (I-Semantics)*.
9. **Dalvi, N., Kumar, R., Pang, B., Ramakrishnan, R., Tomkins, A., Bohannon, P., Keerthi, S., & Merugu, S. (2009).** A web of concepts. *Proceedings of the Twenty-eighth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '09, ACM, New York, NY, USA, pp. 1–12.
10. **Derczynski, L., Maynard, D., Aswani, N., & Bontcheva, K. (2013).** Microblog-genre noise and impact on semantic annotation accuracy. *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, HT '13, ACM, New York, NY, USA, pp. 21–30.
11. **Deschacht, K. & Moens, M. (2007).** Text analysis for automatic image annotation. In *The 45th Annual Meeting of the Association for Computational Linguistics*. ACL.
12. **Deschacht, K., Moens, M., & Robeyns, W. (2007).** Cross-media entity recognition in nearly parallel visual and textual documents. *Large Scale Semantic Access to Content (Text, Image, Video, and Sound)*, RIAO '07, pp. 133–144.
13. **Fernando, P., Pinto, D., Cardiff, J., & Rosso, P. (2011).** On the difficulty of clustering microblog texts for online reputation management. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA 11. ACL, pp. 146–152.
14. **Ferragina, P. & Scaiella, U. (2010).** TAGME: On-the-fly Annotation of Short Text Fragments (by Wikipedia Entities). *Proc. of the* 19^{th} *ACM Intl. Conf. on Information and Knowledge Management (CIKM)*, pp. 1625–1628.
15. **Finkel, J. R., Grenager, T., & Manning, C. D. (2005).** Incorporating non-local information into information extraction systems by gibbs sampling. **for Computer Linguistics, T. A.**, editor, *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pp. 363–370.
16. **Gamon, M., Yano, T., Song, X., Apacible, J., & Pantel, P. (2013).** Identifying salient entities in web pages. *Proceedings of the 22Nd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '13, ACM, New York, NY, USA, pp. 2375–2380.
17. **Gattani, A., Lamba, D. S., Garera, N., Tiwari, M., Chai, X., Das, S., Subramaniam, S., Rajaraman, A., Harinarayan, V., & Doan, A. (2013).** Entity extraction, linking, classification, and tagging for social media: A wikipedia-based approach. *Proceedings of VLDB Endowment*, Vol. 6, No. 11, pp. 1126–1137.
18. **Gillick, D. & Dunietz, J. (2014).** A new entity salience task with millions of training examples. *Proceedings of the European Association for Computational Linguistics*.
19. **Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., & Smith, N. A. (2011).** Part-of-speech tagging for twitter: Annotation, features, and experiments. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, ACL, Stroudsburg, PA, USA, pp. 42–47.
20. **Lafferty, J., McCallum, A., & Pereira, F. (2001).** Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML 2001, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 282–289.
21. **Liu, X., Li, Y., Wu, H., Zhou, M., Wei, F., & Lu, Y. (2013).** Entity linking for tweets. *The Annual Meeting of the Association for Computational Linguistics*, ACL.
22. **Manning, C. D., Raghavan, P., & Schtze., H. (2008).** *Introduction to Information Retrieval*. Cambridge University Press, New York, USA.
23. **McParlane, P. J., McMinn, A. J., & Jose, J. M. (2014).** "picture the scene...";: Visually summarising social media events. *Proceedings of the 23rd ACM International Conference on Conference on*

Information and Knowledge Management, CIKM '14, ACM, New York, NY, USA, pp. 1459–1468.

24. **Meij, E., Weerkamp, W., & de Rijke, M. (2012).** Adding semantics to microblog posts. *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, ACM, New York, NY, USA, pp. 563–572.
25. **Pattabhiraman, T. & Cercone, N. (1990).** Selection: Salience, relevance and the coupling between domain-level tasks and text planning. *Proceedings of the Fifth International Workshop on Natural Language Generation*, pp. 7986.
26. **Pear Analytics (2009).** *Twitter study*.
27. **Ritter, A., Clark, S., Mausam, & Etzioni, O. (2011).** Named entity recognition in tweets: An experimental study. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, ACL, Stroudsburg, PA, USA, pp. 1524–1534.
28. **Saez-Trumper, D., Castillo, C., C., & Lalmas, M. (2013).** Social media news communities: gatekeeping, coverage, and statement bias. *Proceedings of the 22nd ACM international conference on on information and knowledge management*, CIKM '13, ACM, pp. 1679–1684.
29. **Shen, W., Wang, J., & Han, J. (2015).** Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 27, No. 2, pp. 443–460.
30. **Surender Reddy Yerva, K. A., Zoltan Miklos (2012).** Entity-based classification of twitter messages. *IJCSA*, Vol. 9, No. 2, pp. 88–115.
31. **Tran, T. A., Niederee, C., Kanhabua, N., Gadiraju, U., & Anand, A. (2015).** Balancing novelty and salience: Adaptive learning to rank entities for timeline summarization of high-impact events. *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, ACM, New York, NY, USA, pp. 1201–1210.
32. **Trani, S., Ceccarelli, D., Lucchese, C., Orlando, S., & Perego, R. (2016).** Sel: A unified algorithm for entity linking and saliency detection. *Proceedings of the 2016 ACM Symposium on Document Engineering*, DocEng '16, ACM, New York, NY, USA, pp. 85–94.
33. **Yamada, I., Ito, T., Usami, S., Takagi, S., Takeda, H., & Takefuji, Y. (2014).** Evaluating the helpfulness of linked entities to readers. *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, pp. 169–178.
34. **Yamada, I., Shindo, H., Takeda, H., & Takefuji, Y. (2016).** Joint learning of the embedding of words and entities for named entity disambiguation. *CoNLL*, pp. 250–259.
35. **Zhao, X., Zhu, F., Qian, W., & Zhou, A. (2013).** Impact of multimedia in sina weibo: Popularity and life span. *Springer Proceedings in Complexity Semantic Web and Web Science*, Springer New York.

Article received on 22/12/2016; accepted on 20/02/2017.
Corresponding author is Priya Radhakrishnan.