SNEIT : Salient Named Entity Identification in Tweets

Priya Radhakrishnan, Ganesh Jawahar, Dr. Manish Gupta, Dr. Vasudeva Varma

February, 2016

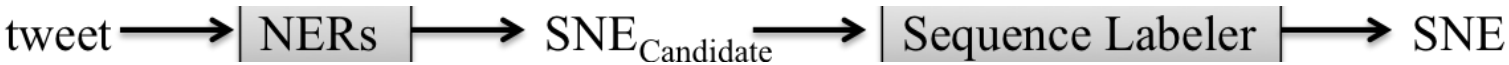
Identifying Salient Named Entities

- Information Extraction is about analyzing the textual content to extract the important or salient information from it.
- An NE is salient when it is *central* to the document.
- We propose a method to identify salient named entities in tweets, assigning a salience score to each NE in a tweet.
- Approach : A supervised labeler learns salience from an annotated dataset.

Definition - Identifying SNE

- 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.
- Entity in the image accompanying the text is most likely the salient entity of te text[DM07]
- **SNE** "*that entity present in the image accompaning the tweet.*"
- We look at the Named Entities in the tweet and pick the salient one, based on the features of the tweet.

Approach - Identifying SNE



- Identify the NE (using 3 NERs) called $SNE_{candidate}$.
- Train a learner to select the SNE from the $SNE_{candidate}$.
- Learn using a dataset of tweets and their salient entities.
- The dataset is created using imaged tweets. Tweets are annotated with entity in the image as its SNE.

Approach - Identifying SNE

- Learner associates a salience score for each $\text{SNE}_{candidate}$.
- Salience score is the probability of a $\text{SNE}_{candidate}$ being SNE.
- identifying salience of $\text{SNE}_{candidate}$ considers context (neighboring words) of the $\text{SNE}_{candidate}$.
- Model the learning task as a supervised sequence labeling task, implemented using Conditional Random Fields (CRF).
- CRF features include word features, POS tag, Chunk POS tag and Entity tag.
- 15 textual features.
- Labeler's target labels are B-SNE, I-SNE and O-SNE.

Related Work - Identifying SNE

- Salient entities
- Imaged tweets
- Named Entity Recognition in tweets

Dataset - Identifying SNE

Select the annotator (by name) hari

Select the collection (by name) largeseimptrainingset

Tweet 1202 of 10938 update

Clear | Submit | Prev | Next | Check | D | P | S

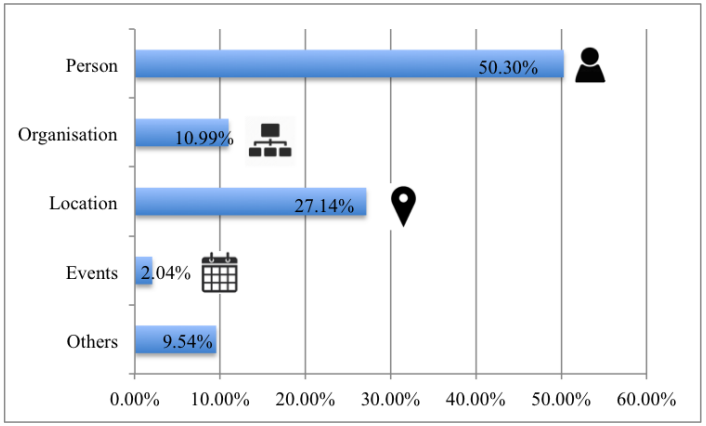
What is so un-subcontinental about #Sangakkara, something similar to Sachin Tendulkar? #SLvSA http://t.co/8Kv5Q7Jm1B http://t.co/JLfWGd40vt#578019877811724289

Select only the NE's you see in the picture. Select also its Knowledge Base entry.

- Sachin Tendulkar (ALAN_RITTER,ARK_TWEET)
- Sangakkara (ALAN_RITTER,ARK_TWEET)
- Tendulkar (STANFORD_CRF)
- Sachin (STANFORD_CRF)

Comments

Dataset - Identifying SNE



- Salient entities are represented in the image accompanying the tweet.
- 3646 tweets manually annotated with the salient named entity and the Wikipedia page that describes it
- Select NE as salient if it is present in the accompanying image
- D = 4272, S = 507, P = 1838, N = 307, no KB = 368
- Tweets on Cricket World Cup 2015

Results & Evaluation

- *Identify SNE in tweet with F-measure of* 0.63.
- Intrinsic evaluation : Using the RepLab 2013 filtering task, performs better than median of tweet filtering task results.[RL13]
- Baseline comparison : Outperforms two of the three baseline methods[GY12][GD14][MW12].

Error Analysis & Design decisions

- SNEIT performs best when SNE$_c$*andidate* is of type Person.
- SNE Identifier misclassifying the gold tag : tags B-SNE and I-SNE get identified as O-SNE

- SNEIT works on text-only and imaged tweets.
- Choice of sporting event - popularity, annotator's interest and background knowledge

References

DM07 K. Deschacht and M. Moens. Text analysis for automatic image annotation. ACL 2007.

GY12 M.Gamon, T.Yano, X.Song, J.Apacible, and P.Pantel. Identifying salient entities in web pages. CIKM'13

GD14 D.Gillick and J.Dunietz. A new entity salience task with millions of training examples. EACL'14.

MW12 E.Meij, W.Weerkamp, and M.d.Rijke. Adding semantics to microblog posts. WSDM'12

RL13 E. Amigo, J. Carrillo de Albornoz, I. Chugur, A. Corujo, J. Gonzalo, T. Martin, E. Meij, M. de Rijke, and D. Spina. Overview of RepLab 2013: Evaluating Online Reputation Monitoring Systems. CLEF 2013.