Name Entity Recognition (NER) in microblogs

Supervisor: Diego Esteves

Omid Najaee

BONN

Terminologies

- ➔ **Entity Recognition:** the discovery entities such as people, locations, organizations and products in text.
  - ◆ example: microsoft_ORG, Merkel_PERSON

Modern NER models

- ➔ Early experiments on state-of-the-art algorithms which are mostly trained on news datasets, demonstrates they have 30-50% accuracy on tweets, in contrast to 80-90% accuracy on longer and well-written text.
- ➔ Major obstacles
    - ◆ **Short messages:** as a source of lack of context
    - ◆ **Noisy content:** social media content often has unusual spelling (e.g. 2moro).
    - ◆ **User-generated content:** users are rich source of information about the user, e.g. demographics, friendships.
    - ◆ **Capitalization:** e.g. *eat an apple* vs. *Apple Inc*.
    - ◆ **Multilingual:** social media content is strongly multilingual.

BONN

Modern NER models: accuracy experiment

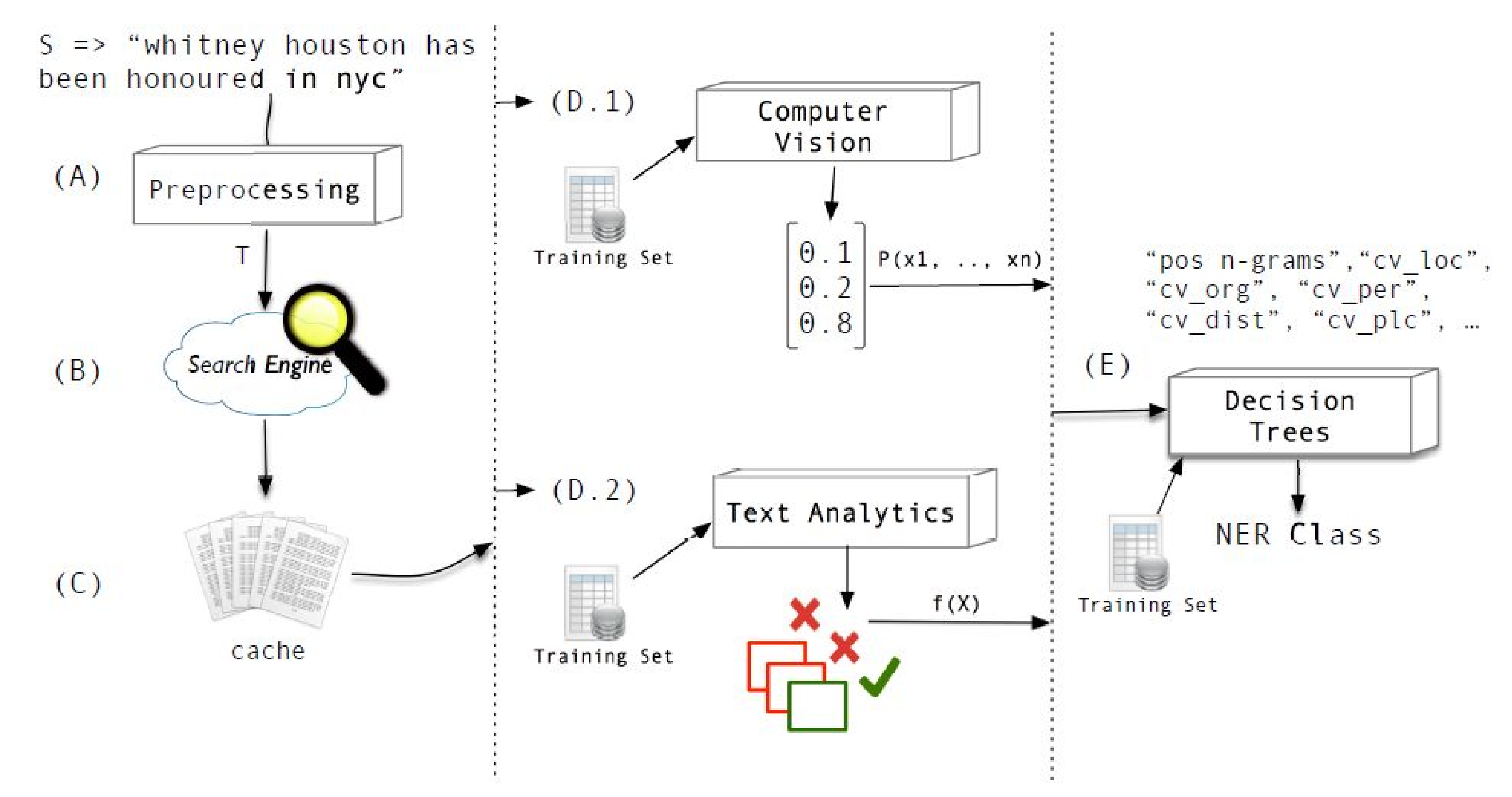
- A study has been ran on 9 different NER systems
- Each of the proposed approaches employ a different approach
- The experiment is ran on three different datasets

System	Ritter Dataset P	R	F1
ANNIE	36.14	16.29	22.46
DBpedia Spotlight	34.70	28.35	31.20
Lupedia	38.85	18.62	25.17
NERD-ML	52.31	**50.69**	**51.49**
Stanford	**59.00**	32.00	41.00
Stanford-Twitter	54.39	44.83	49.15
TextRazor	36.33	38.84	37.54
Zemanta	34.94	20.07	25.49

NER in Twitter using Images and Text

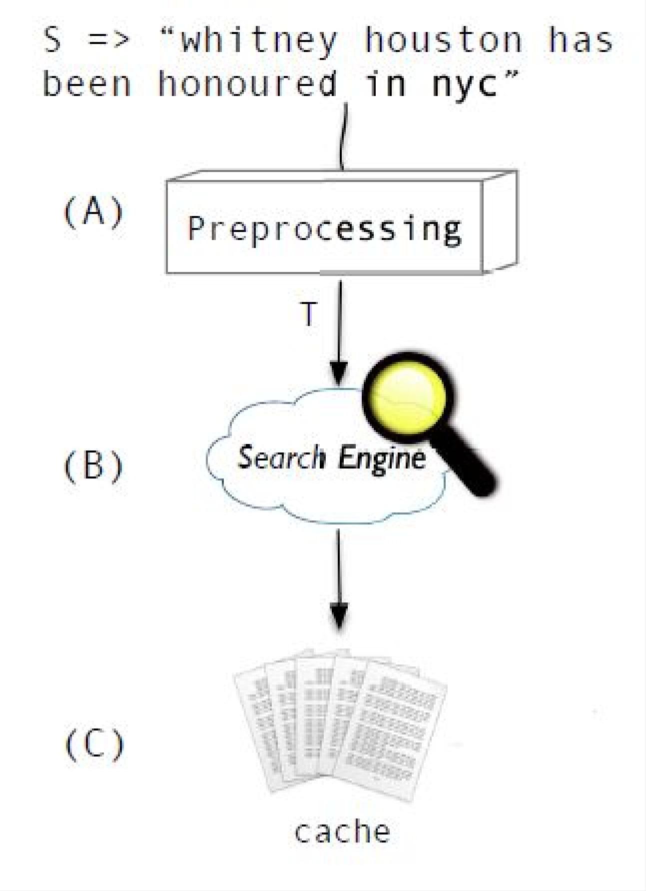
- A hybrid multi-level approach to discover named entities
- combes text and image features with a final classifier based on a *decision tree* model
- This model intents to produce biased indicators to certain classes (LOC, PER, and ORG)
- The proposed features for each class
    - **LOC:** building, suburb, street, city, country, mountain, highway, forest, coast, map
    - **ORG:** company logo
    - **PER:** human face

Architecture



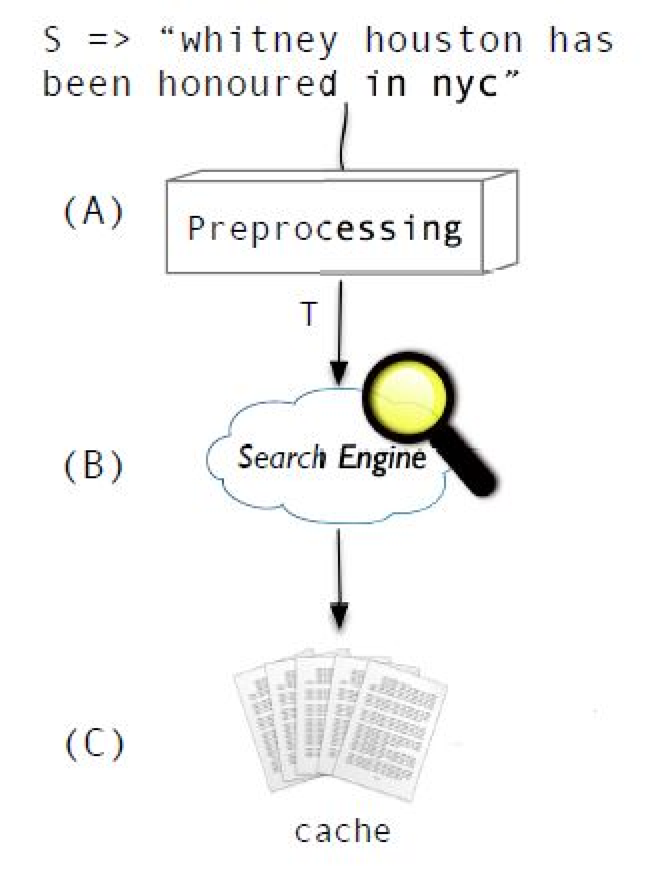
Architecture

(A) *POS Tagging* and *Shallow Parsing* to filter out tokens.
(B) Query from search engine (bing)
(C) Cache top *N* images for each term in the *S*.



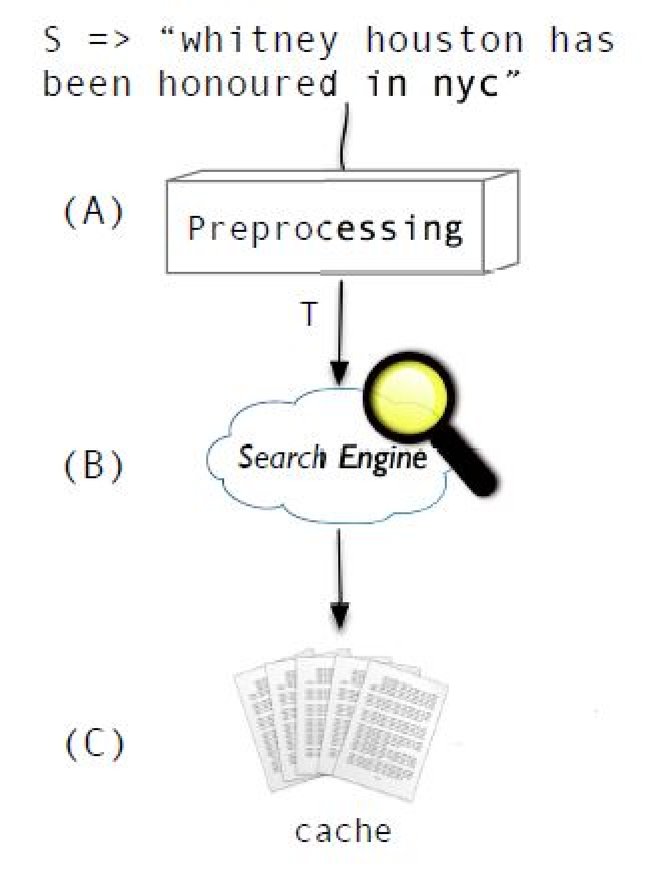
Architecture

(A) *POS Tagging* and *Shallow Parsing* to filter out tokens.
(B) Query from search engine (bing)
(C) Cache top *N* images for each term in the *S*.



Architecture

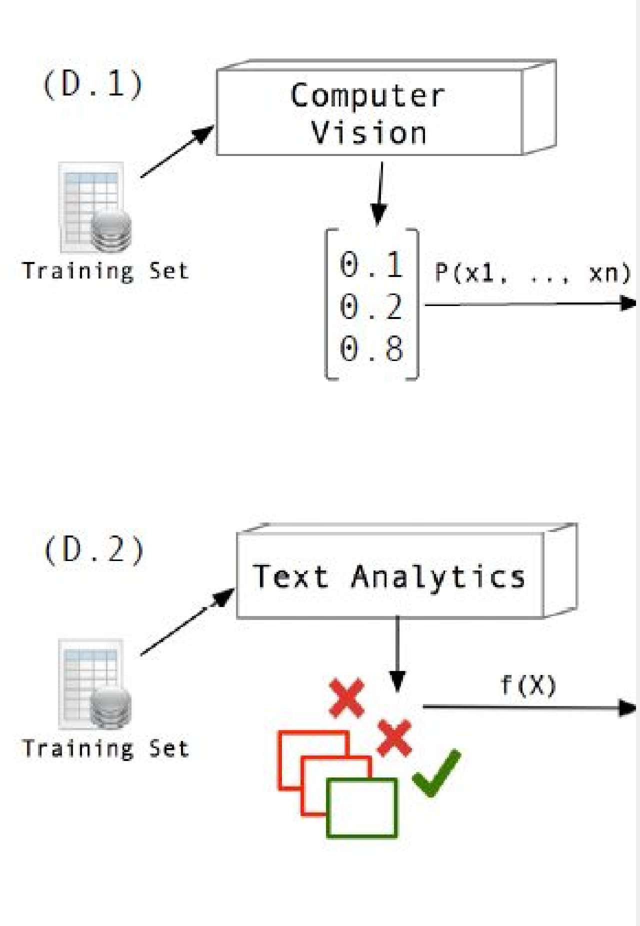
(A) *POS Tagging* and *Shallow Parsing* to filter out tokens.
(B) Query from search engine (bing)
(C) Cache top *N* images for each term in the *S*.



Architecture

(D. 1) detect objects in each picture with a probability P(x_i)

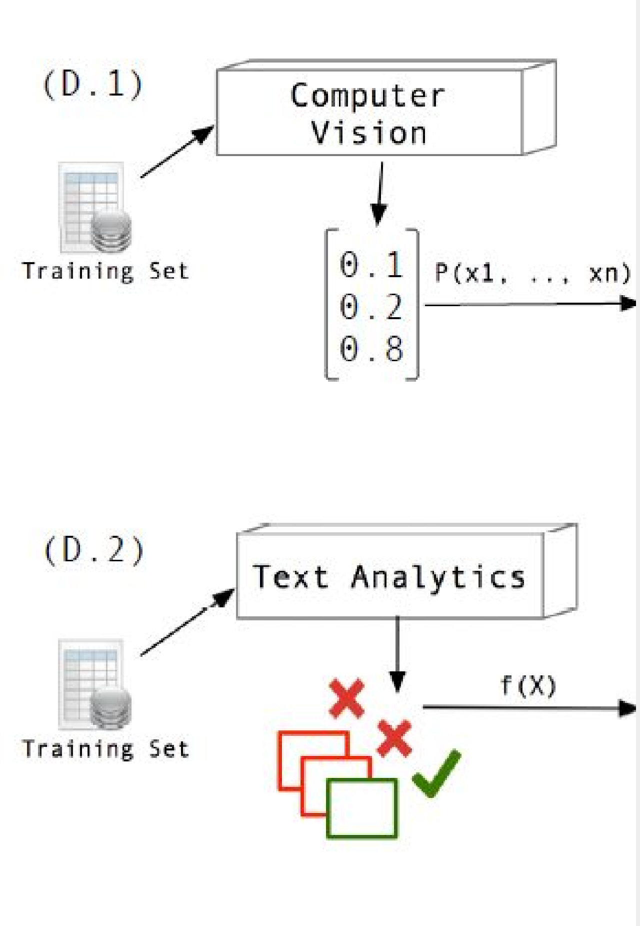
(D. 2) we perform clustering to group texts together that are "distributively" similar.



Architecture

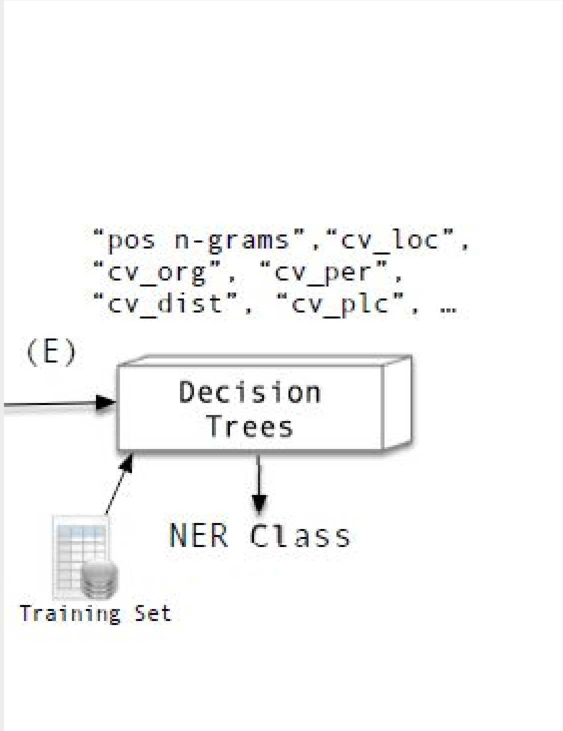
(D. 1) detect objects in each picture with a probability P(x_i)

(D. 2) we perform clustering to group texts together that are “distributively” similar.



Architecture

(E) A decision tree classifier for inferring from the data features.



Experiments

➔ Performance obtained from this approach on ritter dataset with 4-fold cross-validation.

NER System	Description	Precision	Recall	F-measure
Bontcheva et al., 2013	Stanford-twitter	0.54	0.45	0.49
Etter et al., 2013	SVM-HMM	0.65	**0.49**	0.54
this approach	Cluster (images and texts) + DT	**0.82**	0.46	**0.59**

BONN

Topic Modeling

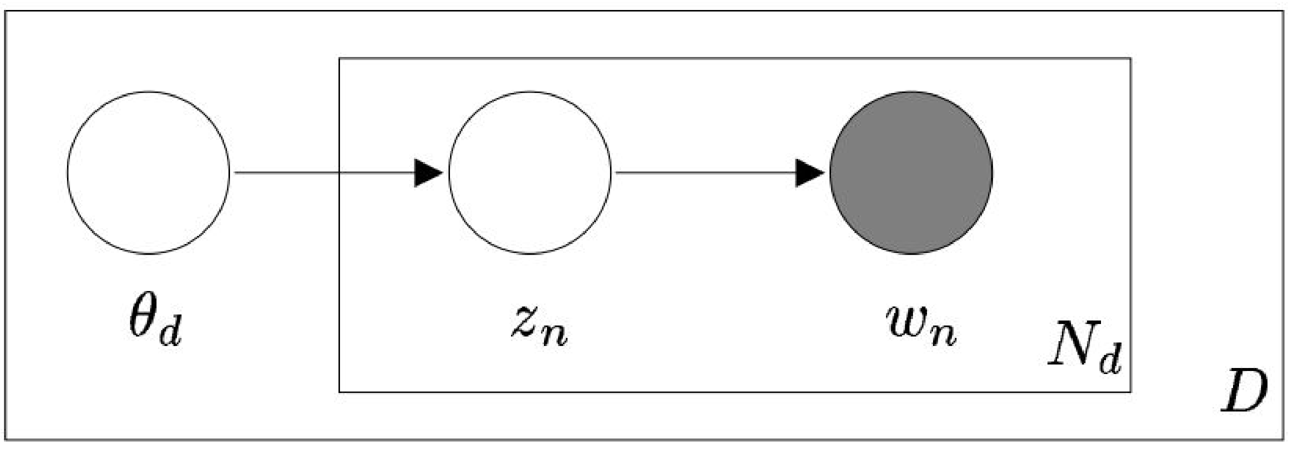
- **Motivation:** organize, search and understand vast quantities of information.
- **Definition:** a method for finding a group of words (aka topic) from a collection of documents that best represents the information in the collection.
- **Capacity**
  - Discovering hidden topical patterns that are present across the collection.
  - Annotating documents according to these topics.
  - Using these annotations to organize, search and summarize texts.

Latent Dirichlet Allocation (LDA)

- One of the techniques for *Topic Modeling*
- It tells what topics are present in any given document by observing all the words in it and producing a topic distribution
- Developed in 2013, David M. Blei, Andrew NG, Michael Jordan, University of California, Berkeley
- **Gensim**, a python-based freely available implementation by Radim Rehurek and Petr Sojka

LDA Model

- The LDA model for words, topics and documents



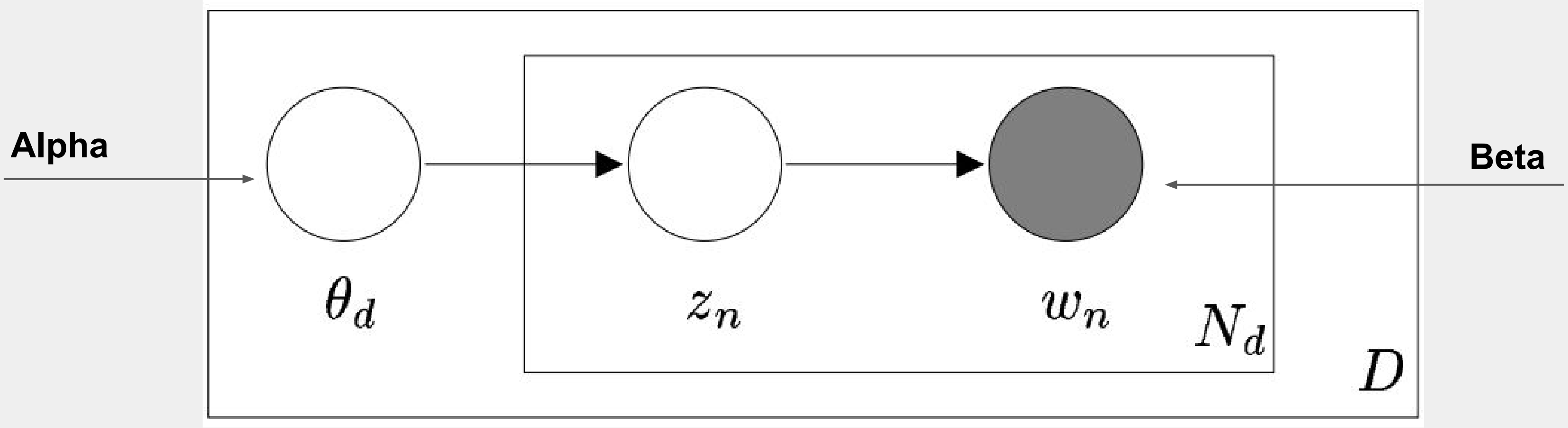
w_n: observed words in a document i

z_n: the topic for j-th word in document i

theta_d: the topic distribution for document i

LDA Model

- Alpha and Beta hyperparameters of the model



Alpha: a parameter that sets the prior on the per-document topic distrib
Beta: a parameter that sets the prior on the per-topic word distrib

Thank you for your attention.

Questions?

References

- Isabelle Augenstein, Leon Derczynski, Kalina Bontcheva, **Generalisation in Named Entity Recognition: A Quantitative Analysis**, *University of Sheffield, 2016*
- Diego Esteves, Rafael Peres, Jens Lehmann, and Giulio Napolitano, **Named Entity Recognition in Twitter using Images and Text**, *University of Bonn, 2017*
- David M. Blei and Andrew Y. Ng and Michael I. Jordan, **Latent Dirichlet Allocation**, *Journal of Machine Learning Research*, 2013