

NLP for Social Media

Lecture 5: Direct Processing of Social Media Data

Monojit Choudhury

Microsoft Research Lab, monojitc@microsoft.com

Recap

Lecture 1: Introduction to NLP for Social Media

- Challenges & Opportunities
- Approaches

Lecture 2: Normalization & Spelling Variation

- Characteristics of Social Media text
- Types of Normalization

Lecture 3: Normalization using Noisy Channel

- Noisy channel model
- Estimating language model and channel model

Lecture 4: Transliteration and Word Embeddings

- Certain phenomena specific to Indic languages
- Transliteration
- Word Embedding techniques

Recap

Lecture 1: Introduction to NLP for Social Media

- Challenges & Opportunities
- Approaches

Approach

- Normalization



- Systems/techniques specifically built for SMD.



In the next 3 lectures...

- Lecture 5 (Monday): Direct Processing of Social Media Text
 - POS Tagging
 - Sentiment Detection
- Lecture 6 (Tuesday): Handling multilingual content
 - Language Detection
 - Processing code-mixed text
- Lecture 7 (Wednesday): Opportunities of SM Text
 - Understanding the individual
 - Interaction between individuals
 - Society and language

Developing SM-specific NLP systems

- Challenges
 - SM specific data creation
 - SM specific features
 - Experiments
- Potential Opportunities
 - Leveraging existing (for std. language) techniques, knowledge & resources as much as possible
 - Leveraging characteristics of social media



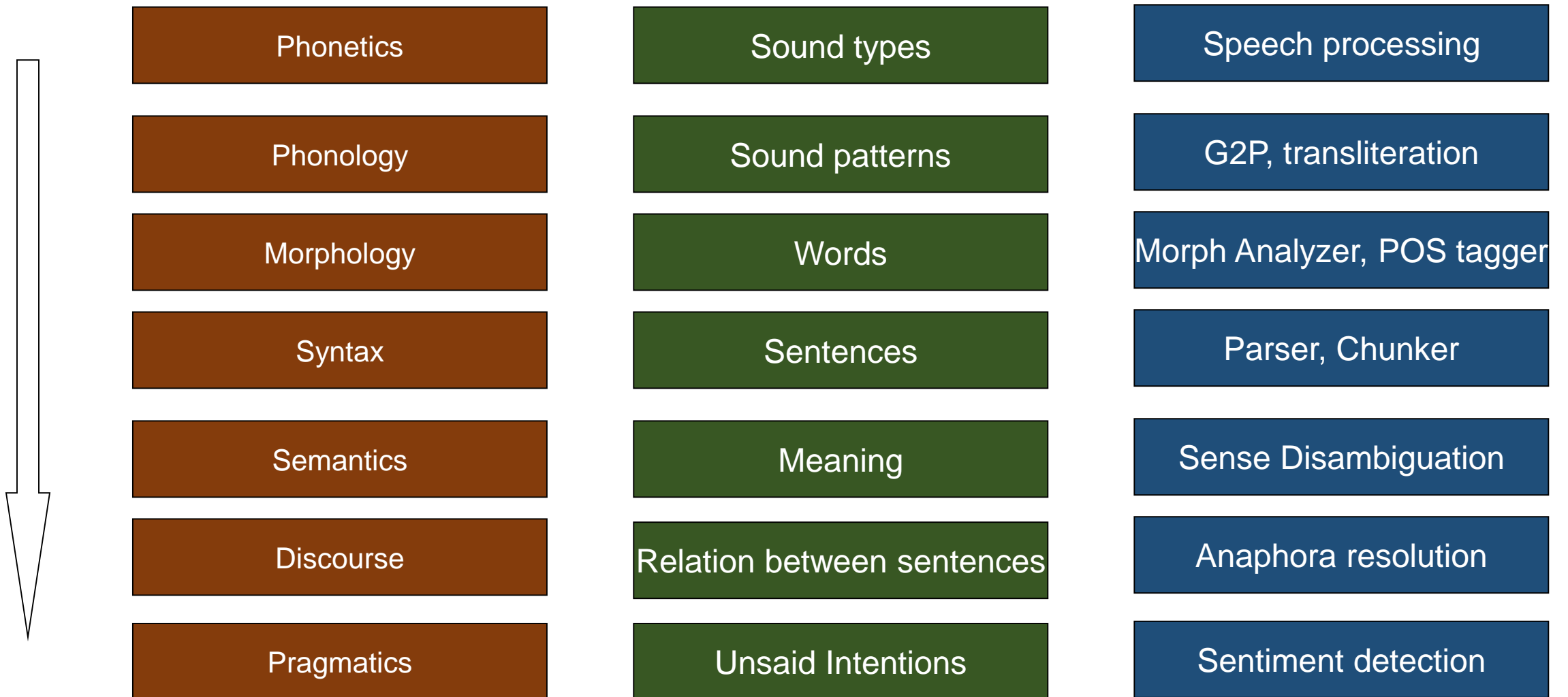
Agenda for Lecture 5

- Case-study 1: POS tagging
 - Basics of POS tagging
 - Gimpel et al. (2011)
 - Extensions & Other problems with a similar flavor:

Break (?)

- Case-study 2: Sentiment Analysis
 - Basics of Sentiment Analysis
 - Pak and Paroubek (2010)
 - Extensions & Other problems with similar flavor

Analyzing Language: A Reductionist Approach



Parts-of-Speech Tagging

Input: The panda eats shoots and leaves.

Output: Det NN VBS NNS CC NNS PUNC

Input: What is your name?

Output: WHP AUX PP NN PUNC

How many POS tags are
there in English?

Can we use the same
POS tagset for all languages?

Modeling POS Taggers

- Sequence Labeling Problem:



$$T^* = \delta(S) = \underset{T}{\operatorname{argmax}} \operatorname{Pr}(T|S)$$

$$= \underset{S}{\operatorname{argmax}} \operatorname{Pr}(S|T)\operatorname{Pr}(T)$$

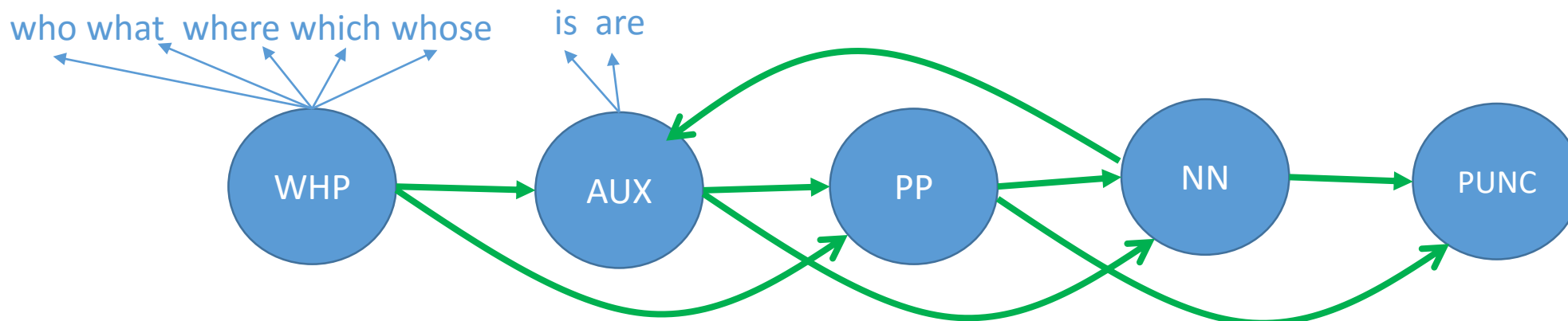
Tag to word
model

Tag sequence
Model

Some commonly made assumptions

$$\begin{aligned} T^* &= \operatorname{argmax} P(S|T)P(T) \\ &= \operatorname{argmax} [\prod_{i=1}^n P(w_i | t_i)]P(T) \\ &= \operatorname{argmax} [\prod_{i=1}^n P(w_i | t_i)] [\prod_{i=1}^n P(t_i | t_{i-1}t_{i-2}t_{i-3}\dots)] \\ &= \operatorname{argmax} \prod_{i=1}^n [P(w_i | t_i)P(t_i | t_{i-1}t_{i-2}t_{i-3}\dots)] \\ &= \operatorname{argmax} \prod_{i=1}^n [P(w_i | t_i)P(t_i | t_{i-1})] \end{aligned}$$

Hidden Markov Models
(HMMs)



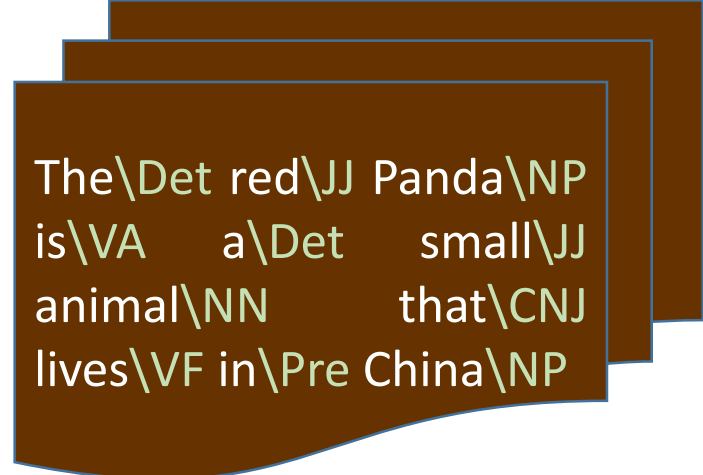
Estimating Probabilities

- Emission Probabilities $P(\text{word} | \text{tag})$

$$P(\text{word} | \text{tag}) = \frac{\text{count}(\text{word} - \text{tag})}{\text{count}(\text{tag})}$$

- Transition Probabilities $(\text{tag}_i | \text{tag}_{i-1})$

$$P(\text{tag}_i | \text{tag}_{i-1}) = \frac{\text{count}(\text{tag}_{i-1} \text{tag}_i)}{\text{count}(\text{tag}_{i-1})}$$



The\Det red\JJ Panda\NP
is\VA a\Det small\JJ
animal\NN that\CNJ
lives\VF in\Pre China\NP

State-of-the-art in POS Tagging

- Technology:
 - HMMs
 - Conditional Random Fields
 - Max-Ent, SVM, Neural Nets
- Data
 - Order of millions of words annotated for English
 - Order of 100s of thousands in many languages around the world
- Accuracy
 - 98% for English
 - 90%+ for most other languages

Why do we need POS Taggers

- As a syntactic preprocessing step
- As features for many other applications:
 - Translation
 - Information Retrieval
 - Named Entity Recognition
 - Sentiment Analysis

Case Study I: POS Tagging for Twitter

Gimpel, Kevin, et al. "Part-of-speech tagging for twitter: Annotation, features, and experiments." *ACL* 2011.

“Our contributions are as follows:

- we developed a POS tagset for Twitter,
- we manually tagged 1,827 tweets,
- we developed features for Twitter POS tagging and conducted experiments to evaluate them, and
- we provide our annotated corpus and trained POS tagger to the research community.

POS Tagset

- (a) @Gunservatively@ obozo^ will_V go_V nuts_A
when_R PA^ elects_V a_D Republican_A Governor_N
next_P Tue^ ., Can_V you_O say_V redistricting_V ?,
- (b) Spending_V the_D day_N withhh_P mommma_N !,
- (c) lmao! ..., s/o_V to_P the_D cool_A ass_N asian_A
officer_N 4_P #1_\$ not_R runnin_V my_D license_N and_&
#2_\$ not_R takin_V dru_N boo_N to_P jail_N ., Thank_V
u_O God^ ., #amen#

Twitter/online-specific

- # hashtag (indicates topic/category for tweet)
- @ at-mention (indicates another user as a recipient of a tweet)
- ~ discourse marker, indications of continuation of a message across multiple tweets
- U URL or email address
- E emoticon

Tagging Convention

Hashtags and at-mentions can also serve as words or phrases within a tweet; e.g. Is #qadaffi going down?. When used in this way, we tag hashtags with their appropriate part of speech, i.e., as if they did not start with #. Of the 418 hashtags in our data, 148 (35%) were given a tag other than #: 14% are proper nouns, 9% are common nouns, 5% are multi-word expressions (tagged as **G**), 3% are verbs, and 4% are something else. We do not apply this procedure to at-mentions, as they are nearly always proper nouns.

than for Standard English text. For example, apostrophes are often omitted, and there are frequently words like *ima* (short for *I'm gonna*) that cut across traditional POS categories. Therefore, we opted not to split contractions or possessives, as is common in English corpus preprocessing; rather, we introduced four new tags for combined forms: {nominal, proper noun} × {verb, possessive}.⁵

Twitter Specific Features

- **TWORTH**: *Twitter orthography features*
 - several regular expression-style rules that detect at-mentions, hashtags, URLs.
- **NAMES**: *Frequently-capitalized tokens.*
 - How often a token is capitalized.
- **TAGDICT**: Traditional tag dictionary.
- **DISTSIM**: Distributional similarity.
 - used 1.9 million tokens from 134,000 unlabeled tweets to construct distributional features from the successor and predecessor probabilities for the 10,000 most common terms
- **METAPH**: Phonetic normalization using metaphones

Experiments & Results

- Train: 1000 Tweets (14.5k Tokens)
- Dev: 327 Tweets (4.8k Tokens)
- Test: 500 Tweets (7.1k Tokens)

	Dev.	Test
Our tagger, all features	88.67	89.37
independent ablations:		
–DISTSIM	87.88	88.31 (–1.06)
–TAGDICT	88.28	88.31 (–1.06)
–TWORDTH	87.51	88.37 (–1.00)
–METAPH	88.18	88.95 (–0.42)
–NAMES	88.66	89.39 (+0.02)
Our tagger, base features	82.72	83.38
Stanford tagger	85.56	85.85
Annotator agreement	92.2	

Related Problems

- Entity Recognition:
 - Named Entity: Names of people, places, organization
 - Date & time
 - How can you model it as a sequence labeling problem?
- Event Recognition

Case Study II: Sentiment Analysis

Pak, Alexander, and Patrick Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." *LREC*. Vol. 10. 2010.

What is Sentiment?

funkeybrewster: @redeyechicago I think Obama's visit might've sealed the victory for Chicago. Hopefully the games mean good things for the city.

vcurve: I like how Google celebrates little things like this: Google.co.jp honors Confucius Birthday — Japan Probe

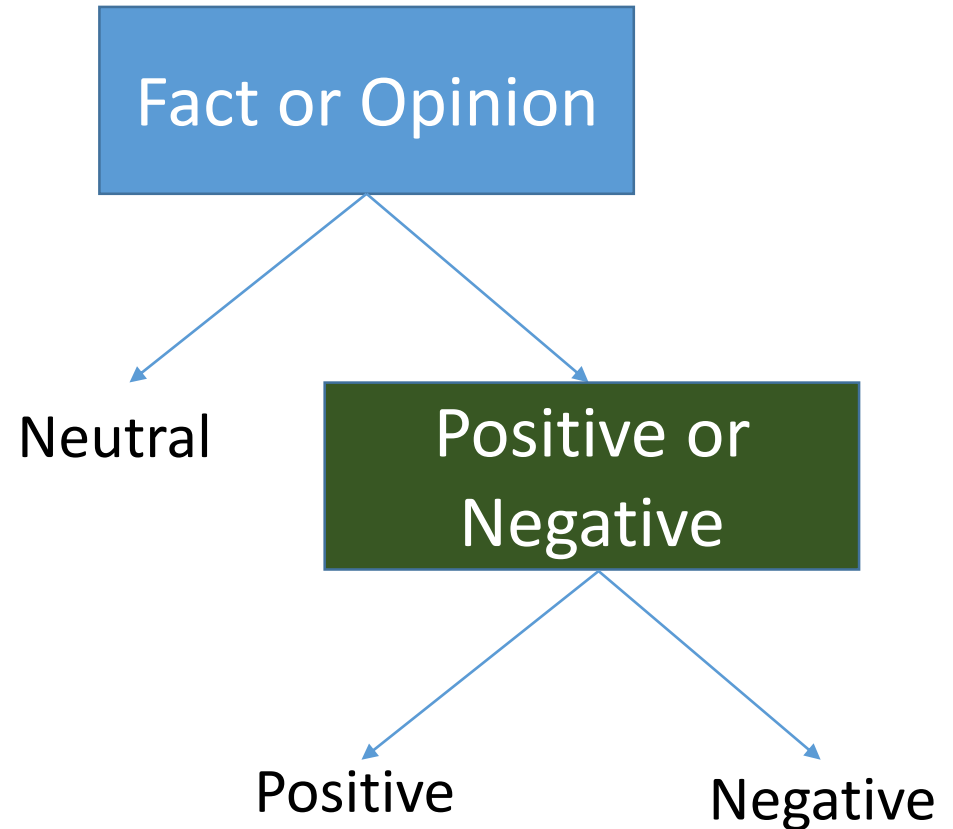
mattfellows: Hai world. I hate faulty hardware on remote systems where politics prevents you from moving software to less faulty systems.

brrooklyn: I love the sound my iPod makes when I shake to shuffle it. Boo bee boo

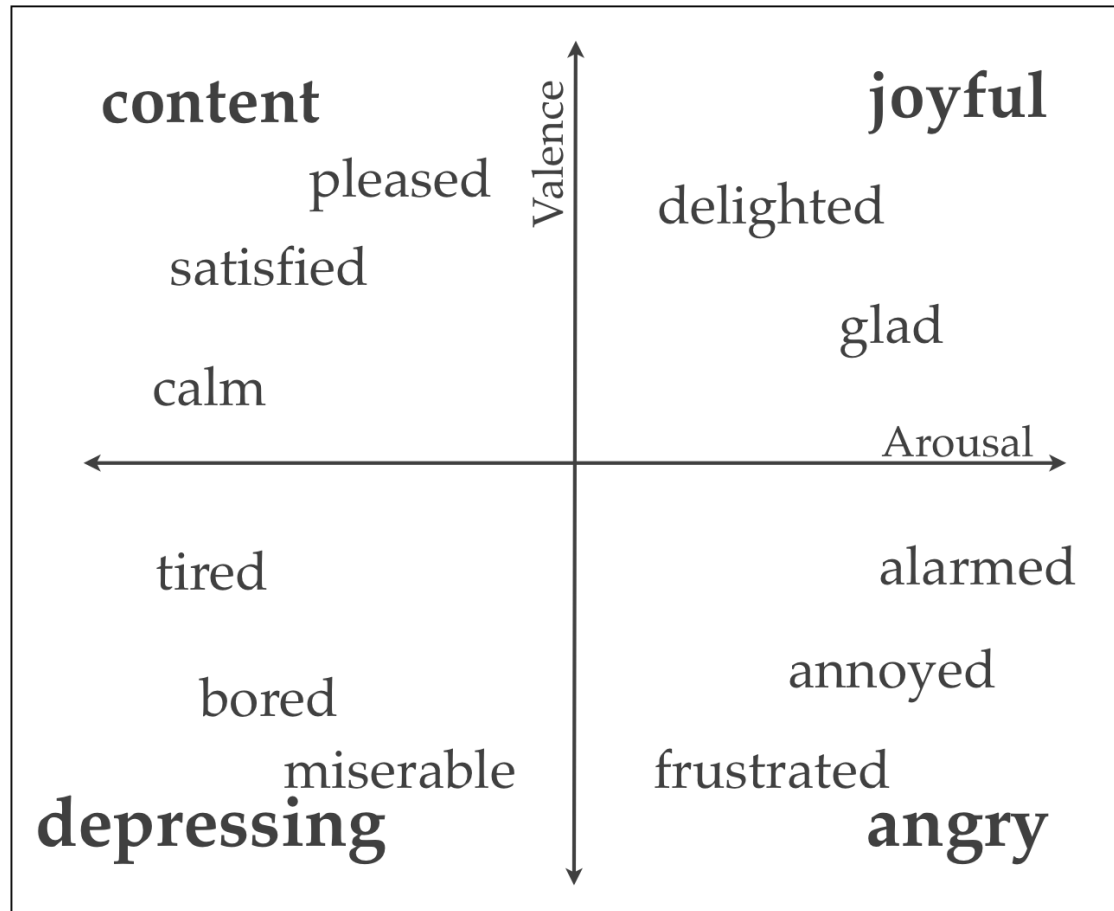
MeganWilloughby: Such a Disney buff. Just found out about the new Alice in Wonderland movie. Official trailer: <http://bit.ly/131Js0> I love the Cheshire Cat.

Sentiment Analysis

- A three way classification:
 - Positive
 - Neutral
 - Negative



Emotion Detection



Reader vs. Writer's
emotion

Input Unit

- Documents
- Blogs
- Sentences
- Phrases
- Words
- Tweets

Contributions...

1. We present a method to collect a corpus with positive and negative sentiments, and a corpus of objective texts. Our method allows to collect negative and positive sentiments such that no human effort is needed for classifying the documents. Objective texts are also collected automatically. The size of the collected corpora can be arbitrarily large.
2. We perform statistical linguistic analysis of the collected corpus.
3. We use the collected corpora to build a sentiment classification system for microblogging.
4. We conduct experimental evaluations on a set of real microblogging posts to prove that our presented technique is efficient and performs better than previously proposed methods.

Data Creation

- Happy emoticons: “:-)””, “:)””, “=)””, “:D” etc.
- Sad emoticons: “:-(””, “:(””, “=(””, “;(”” etc.

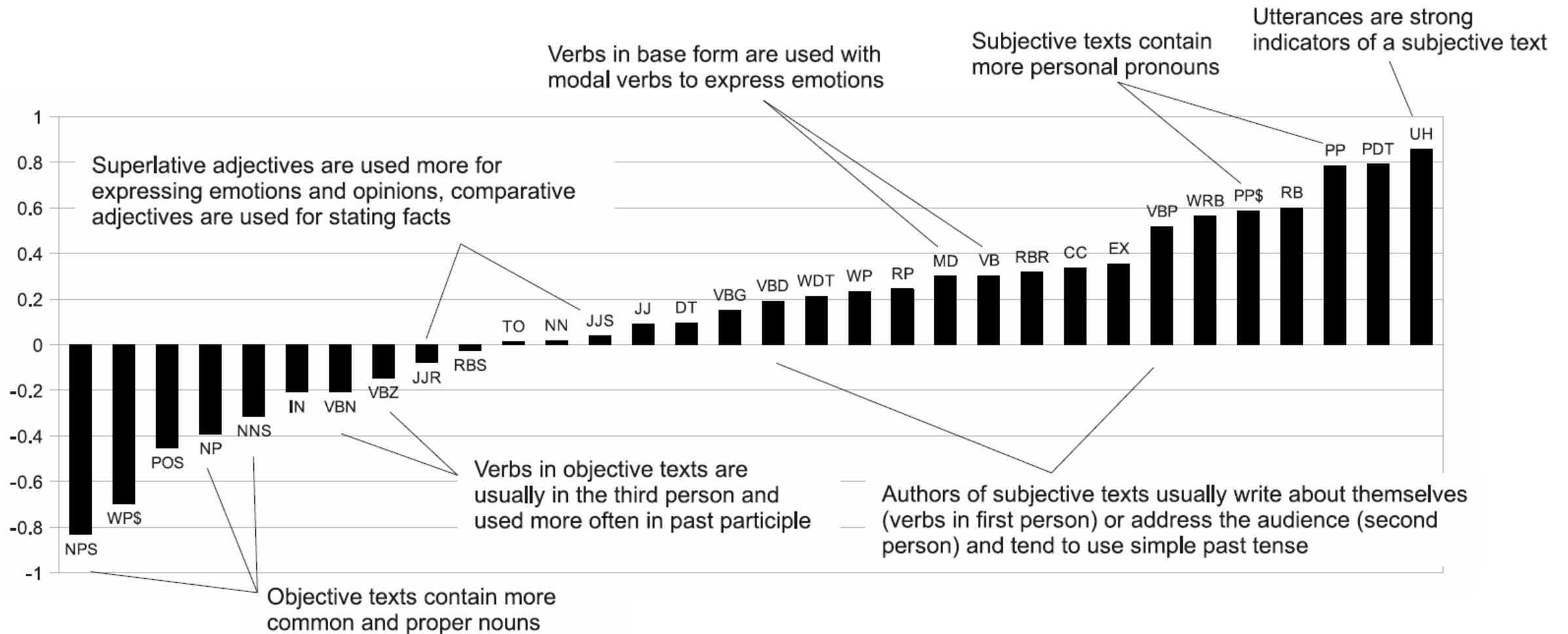
In order to collect a corpus of objective posts, we retrieved text messages from Twitter accounts of popular newspapers and magazines , such as “New York Times”, “Washington Posts” etc. We queried accounts of 44 newspapers to collect a training set of objective texts.

Features

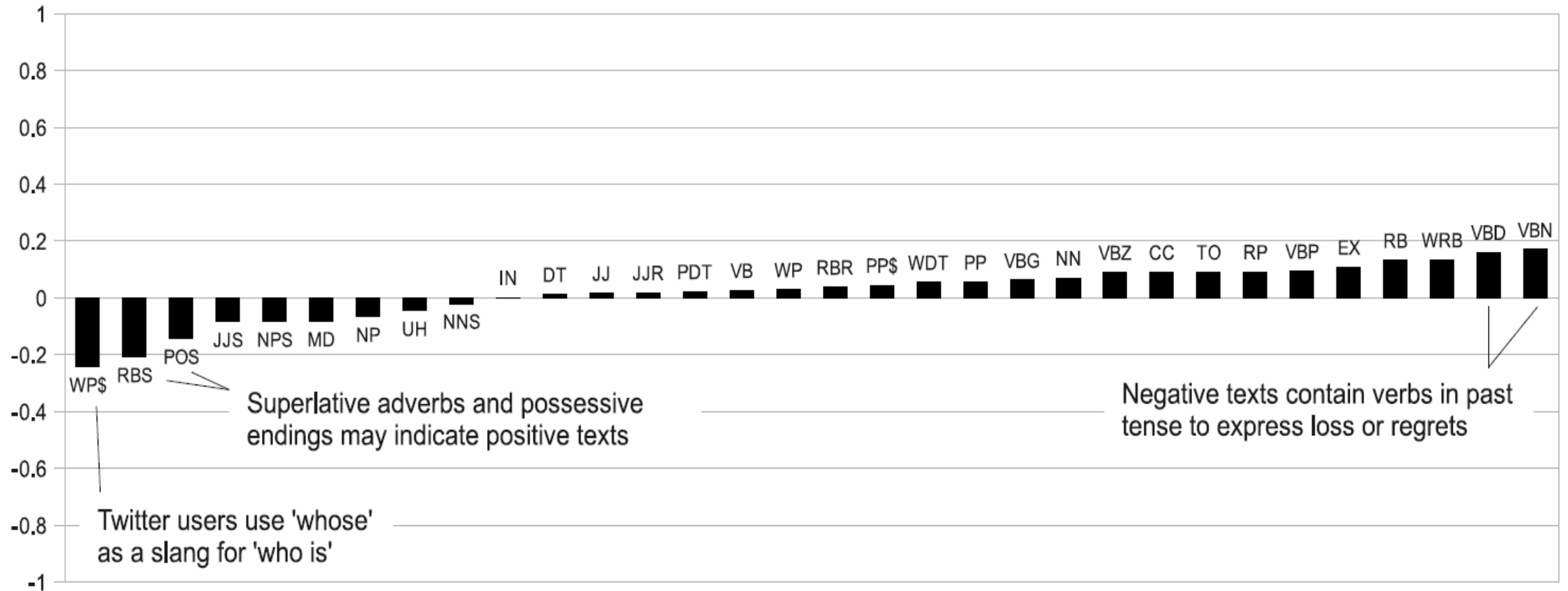
- Word n-grams
- Parts-of-speech tags

Constructing n-grams – we make a set of n-grams out of consecutive words. A negation (such as “no” and “not”) is attached to a word which precedes it or follows it. For example, a sentence “I do not like fish” will form two bigrams: “I do+not”, “do+not like”, “not+like fish”. Such a procedure allows to improve the accuracy of the classification since the negation plays a special role in an opinion and sentiment expression(Wilson et al., 2005).

POS Tag Distributions: Subjective vs. Objective



POS Tag Distribution: Positive vs. Negative



Results and Conclusions

- Accuracy reaches around 60 to 70% (where a random baseline will have 33% accuracy).
- Best performance with bigrams
- Attachment of “Not” and negation helps
- POS tags help

Some Remarks

- Unlike POS tagging, for sentiment analysis, the accuracy of the off-the-shelf tools for standard language is not that bad for social media data.
- Most of the work leverage on the same set of features as for standard language, but train on SM datasets.
- SM specific phenomena are handled during tokenization (removal of hashtags or mentions or URLs)

Problems Similar in Flavor

- Humor detection
- Sarcasm detection
- Politeness detection
- Drunk texting detection
- ...

Suggested Readings & References

POS Tagging:

- Gimpel, Kevin, et al. "Part-of-speech tagging for twitter: Annotation, features, and experiments." *ACL* 2011.
- Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., & Smith, N. A. (2013). Improved part-of-speech tagging for online conversational text with word clusters. Association for Computational Linguistics.

Sentiment Detection:

- Pak, Alexander, and Patrick Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." *LREC*. Vol. 10. 2010.
- Agarwal, Apoorv, et al. "Sentiment analysis of twitter data." *Proceedings of the Workshop on Languages in Social Media*. Association for Computational Linguistics, 2011.

Other References

- Ritter, A., Etzioni, O., & Clark, S. (2012, August). Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1104-1112). ACM.
- Ritter, A., Clark, S., & Etzioni, O. (2011, July). Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 1524-1534). Association for Computational Linguistics.

Resources:

Several Twitter tools from CMU: <http://www.ark.cs.cmu.edu/TweetNLP/>