Sarcasm Detection: A Computational and Cognitive Study

Pushpak Bhattacharyya
CSE Dept.,
IIT Bombay and IIT Patna

California

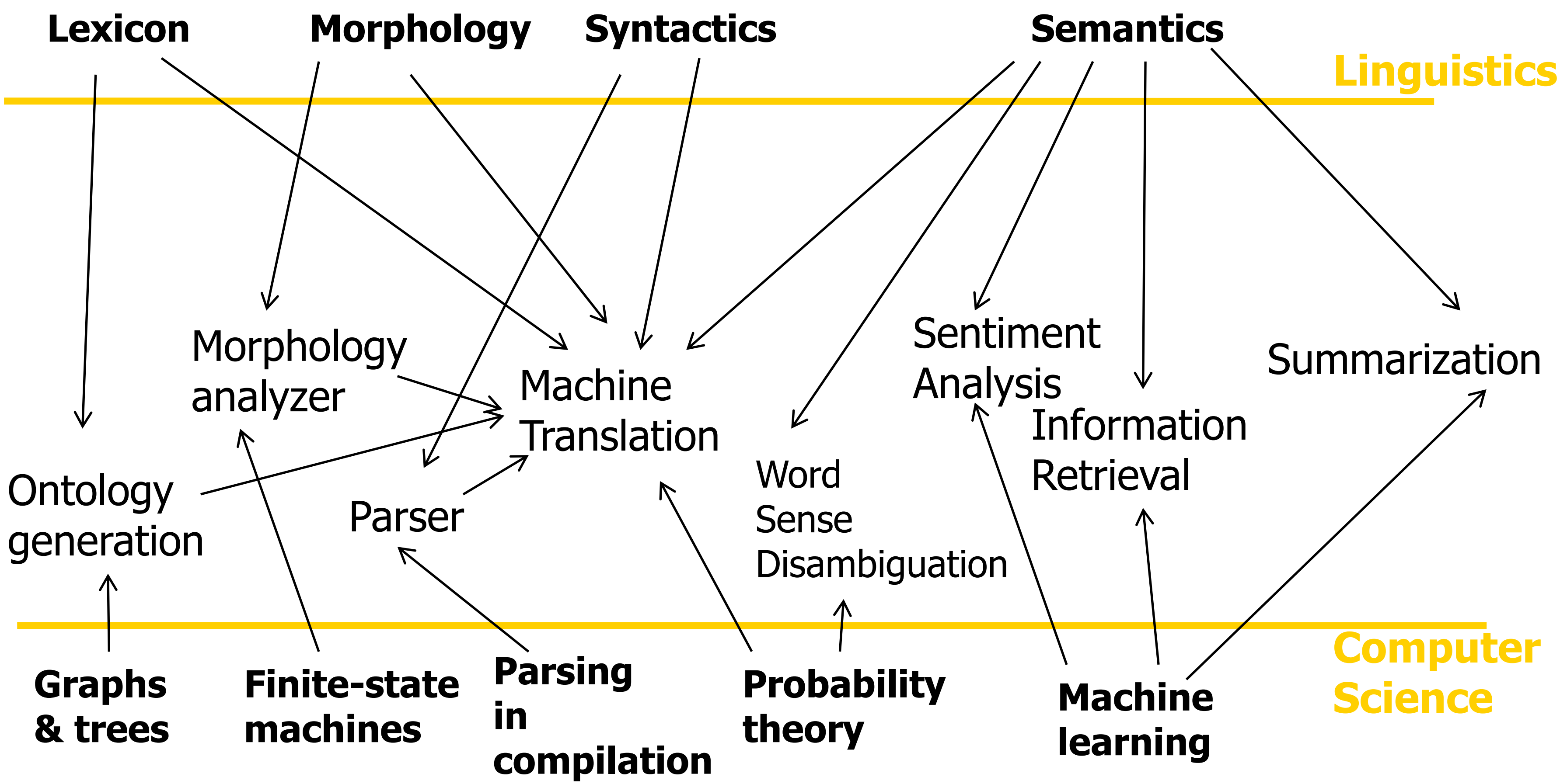
Jan 2018

Acknowledgment: *Aditya, Raksha, Abhijit, Kevin, Lakshya, Arpan, Vaibhav, Prerana, Vinita, Shad* and many, many others


Homepage
Inbox (105,041) - pushpakbh
Inbox (18,664) - pushpakdir
Resource Centre for Indian L
www.cfilt.iitb.ac.in
Search
Government of India
शब्द ब्रह्म अनु संधान
Center For Indian Language Technology
Indian Institute of Technology, Bombay
Home
Join us
People
Downloads
News
Publications
Tools
Different Guidelines
Resources
Events
ML - Linguistics Lectures
Local
Events
Google Custom Sea
CFILT, IIT Bombay
Center for Indian Language Technology (CFILT) was set up with a generous grant from the Department of Information Technology (DIT), Ministry of Communication and Information Technology, Government of India in 2000 at the Department of Computer Science and Engineering, IIT Bombay. Prior to this the Natural Language Processing (NLP) activity of the CSE Department, IIT Bombay took off in 1996 with a grant from the United Nations University, Tokyo to create a multilingual information exchange system for the web. The project called Universal Networking Language (UNL; www.undl.org) was participated in by 15 research groups across continents.
At any point of time about 30 research members work in CFILT, which includes PhD , masters and bachelor students, faculty members, linguists and lexicographers.
Deep semantics and multilinguality has throughout played a pivotal role in the activities of CFILT. The stress on semantics has led to research in the following fronts:
• Lexical Resources: Multilingual wordnets and ontologies and their linking
• Lexical and Structural Disambiguation: Resolve word and attachment ambiguities
• Shallow Parsing: Identifying correct parts of speech, named entities and non-recursive noun phrases for Marathi and Hindi
• Cross Lingual Information Retrieval: Indian language query to English and Hindi Retrieval
• Machine Translation: Automatic translation involving Marathi, Hindi and English
• Text Entailment: Testing if a piece text (hypothesis) is inferable from another (text)
100%
ENG
IN
3:20 PM
11-Jul-17


Homepage
Inbox (105,041) - pus
Inbox (18,664) - push
Resource Centre for I
Download Lectures: AI-NL
Welcome to AI-NLP-ML G
www.iitp.ac.in/~ai-nlp-ml/
Search
सत्यमेव जयते
AI-NLP-ML GROUP
Department of Computer Science and Engineering
Indian Institute of Technology Patna
Home
People
Downloads
Publications
Events
Talks
Gallery
Contact Us
Google Custom Search
Search
Latest News
Current Notifications
ACM Transaction on Asian
One paper has been accepted in Knowledge-Based Systems, Elsevier,
Three papers have been accepted in NLDB-2017.
One paper has been accepted in CICLING-2017.
Shweta Yadav, Research
AI-NLP-ML Group
Department of CSE, IIT Patna
The Artificial Intelligence-Natural Language Processing-Machine Learning (AI-NLP-ML) group at Department of Computer Science and Engineering, IIT Patna has started its official journey in June, 2015. The group is dedicated to explore the frontiers of Artificial Intelligence, Machine Learning and Natural Language Processing under the able guidance of Prof. Pushpak Bhattacharyya. The group also consists of other two faculty members, Dr. Asif Ekbal and Dr. Sriparna Saha, and around 30 members including research scholars, research engineers, lexicographers, B.Tech & M.Tech students. Several industry sponsored projects are currently being undertaken.
Elsevier, the renowned scientific literature publishing company has set up the Elsevier Centre of Excellence for Natural Language Processing to
ENG
IN
3:23 PM
11-Jul-17

NLP: At the confluence of linguistics & computer science



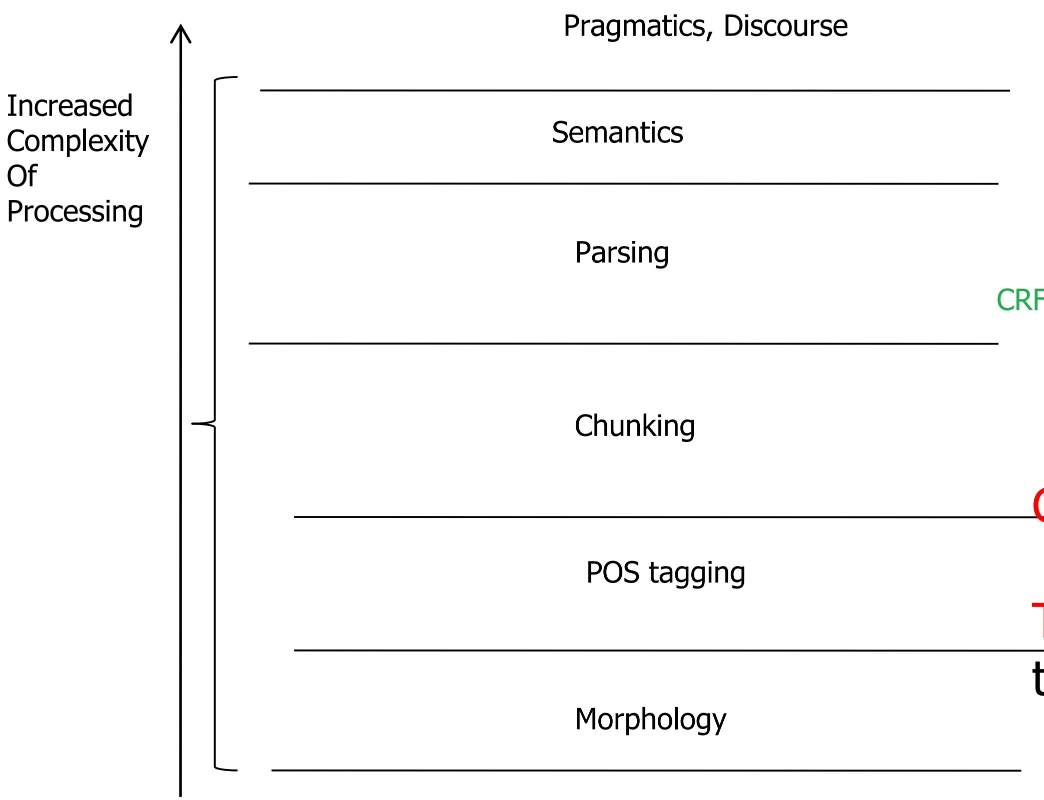
The chain

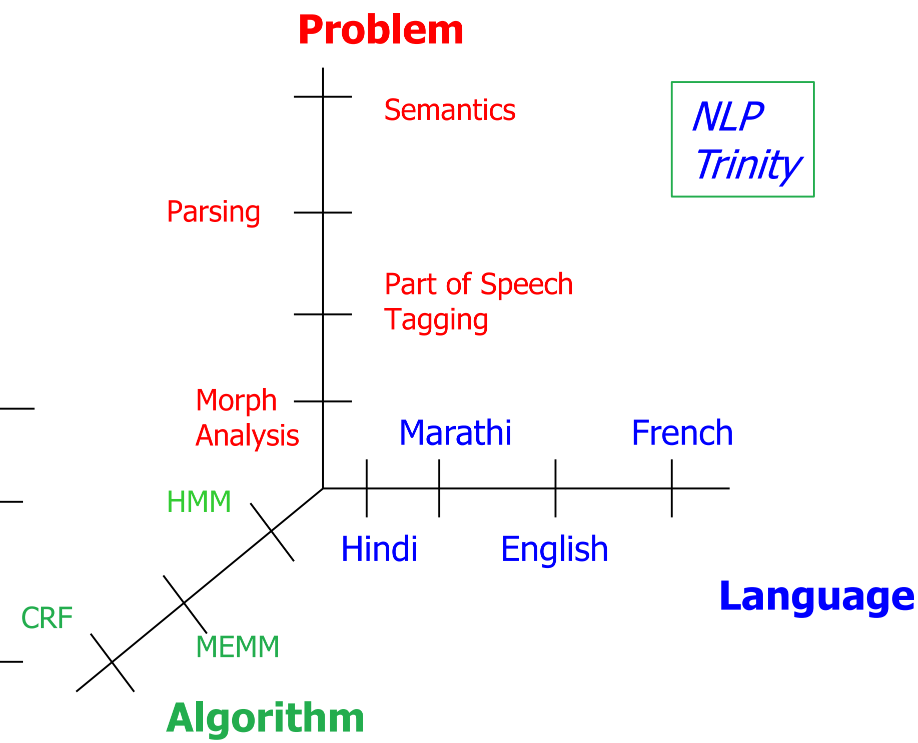
AI → NLP → Sentiment → Sarcasm → Numerical Sarcasm

Roadmap

- NLP and Ambiguity
- Sentiment Analysis
- Sarcasm
  - Features and ML
  - Numerical Sarcasm
  - Cognitive dimension
- Conclusions and future work

NLP: multilayered, Multi dimensional





GharaaSamorChyaaNe malaa sangitle

The one who is in front of the house told me

Need for NLP

- Humongous amount of language data in electronic form
- Unstructured data (like free flowing text) will grow to 40 zetabytes (1 zettabyte= 10^{21} bytes) by 2020.
- How to make sense of this huge data?
- Example-1: e-commerce companies need to know **sentiment** of online users, sifting through 1 lakh e-opinions per week: needs NLP
- Example-2: **Translation** industry to grow to $37 billion business by 2020

Machine Learning

- Automatically learning rules and concepts from data

Learning the concept of table.

What is "tableness"

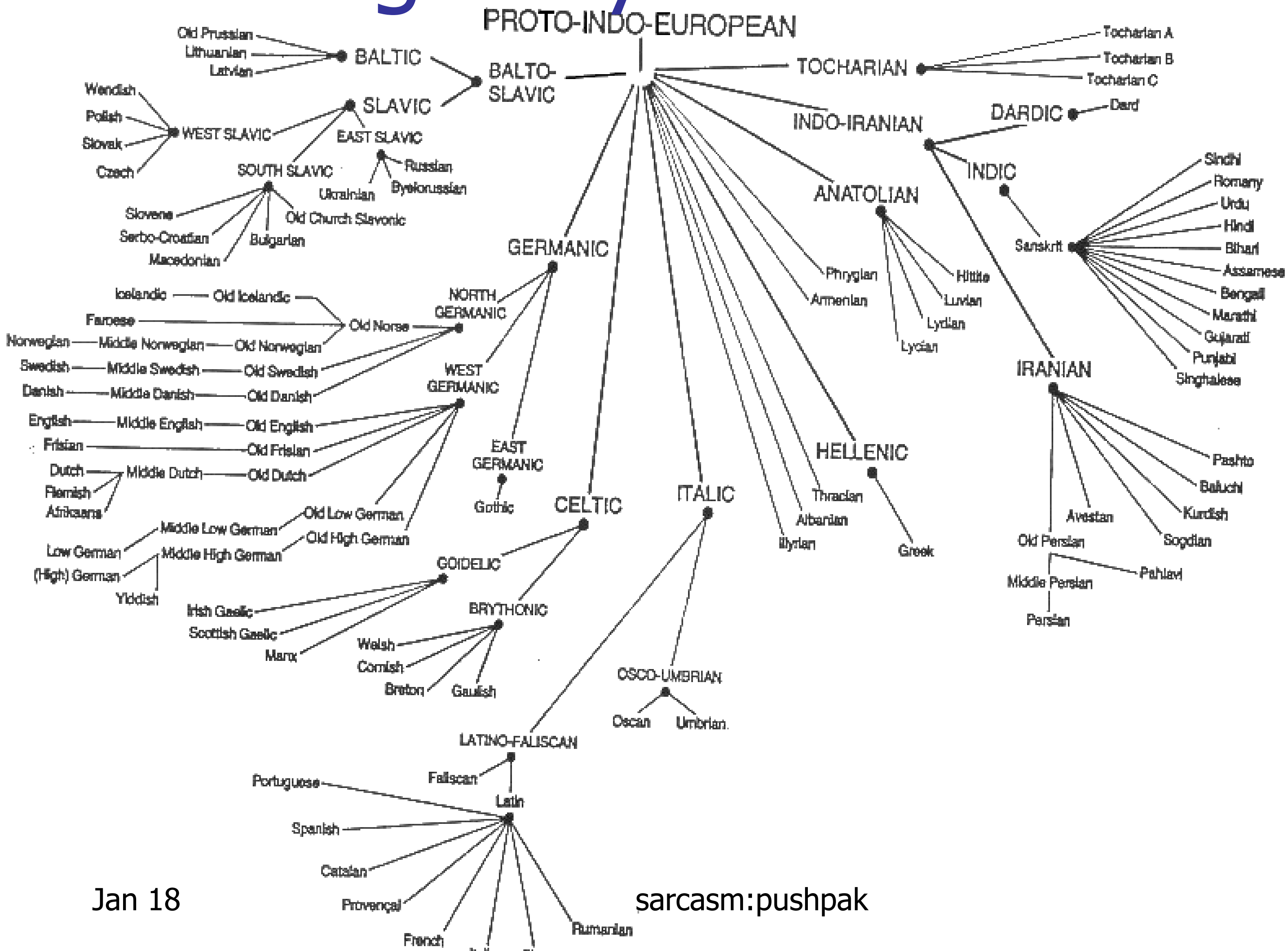
Rule: a flat surface with 4 legs (approx.: to be refined gradually)

NLP-ML marriage

NLP= Ambiguity Processing

- Lexical Ambiguity
  - *Present (Noun/Verb/Adjective; time/gift)*
- Structural Ambiguity
  - *1 and 2 bed room flats live in ready*
- Semantic Ambiguity
  - *Flying planes can be dangerous*
- Pragmatic Ambiguity
  - *I love being ignored* (after a party, while taking leave of the host)

Another challenge of NLP: multilinguality



Rules: when and when not

- When the phenomenon is understood AND expressed, rules are the way to go

- "Do not learn when you know!!"

- When the phenomenon "seems arbitrary" at the current state of knowledge, DATA is the only handle!
  - *Why do we say "Many Thanks" and not "Several Thanks"!*
  - *Impossible to give a rule*

Impact of probability: Language modeling

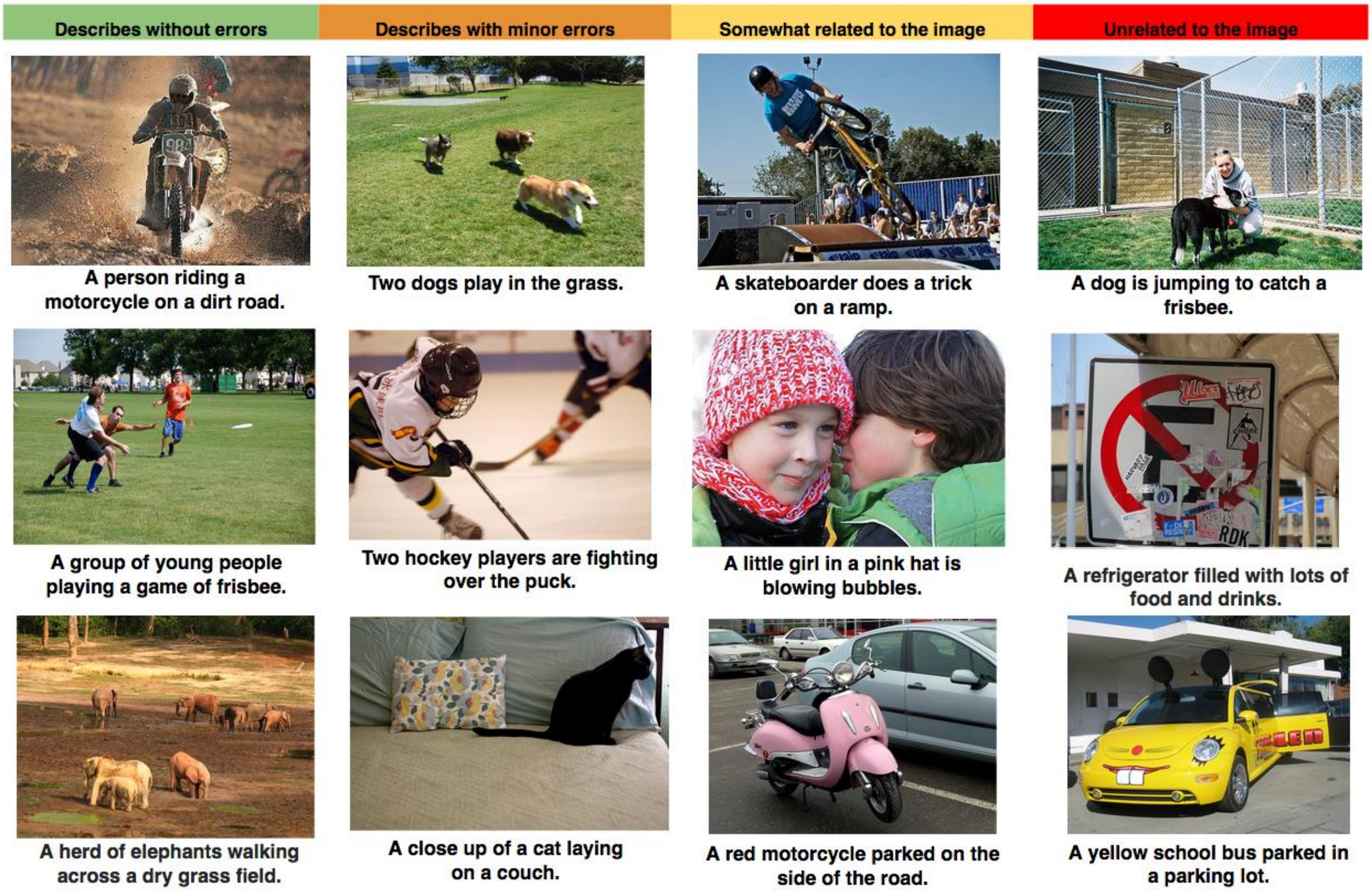
Probabilities computed in the context of corpora

1. P("The sun rises in the east")
2. P("The sun rise in the east")
   - Less probable because of grammatical mistake.
3. P(The svn rises in the east)
   - Less probable because of lexical mistake.
4. P(The sun rises in the west)
   - Less probable because of semantic mistake.

Power of Data- Automatic image labeling
(Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, 2014)

Automatically captioned: "Two pizzas sitting on top of a stove top oven"

Automatic image labeling (cntd)

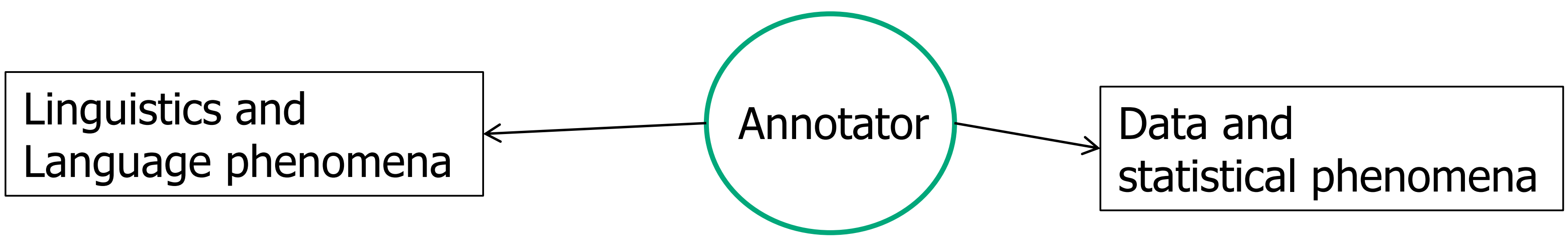


Main methodology

- Object A: extract parts and features
- Object B which is in correspondence with A: extract parts and features
- LEARN mappings of these features and parts
- Use in NEW situations: called DECODING

Linguistics-Computation Interaction

- Need to understand BOTH language phenomena and the data
- An annotation designer has to understand BOTH linguistics and statistics!



Sentiment Analysis

Definition (Liu 2010)

(Liu, 2010) defines a sentiment or opinion as a quintuple-

$$< o_j , f_{jk} , so_{ijkl} , h_i , t_l >,$$

where

o_j *is a target object,*

f_{jk} *is a feature of the object* o_j*,*

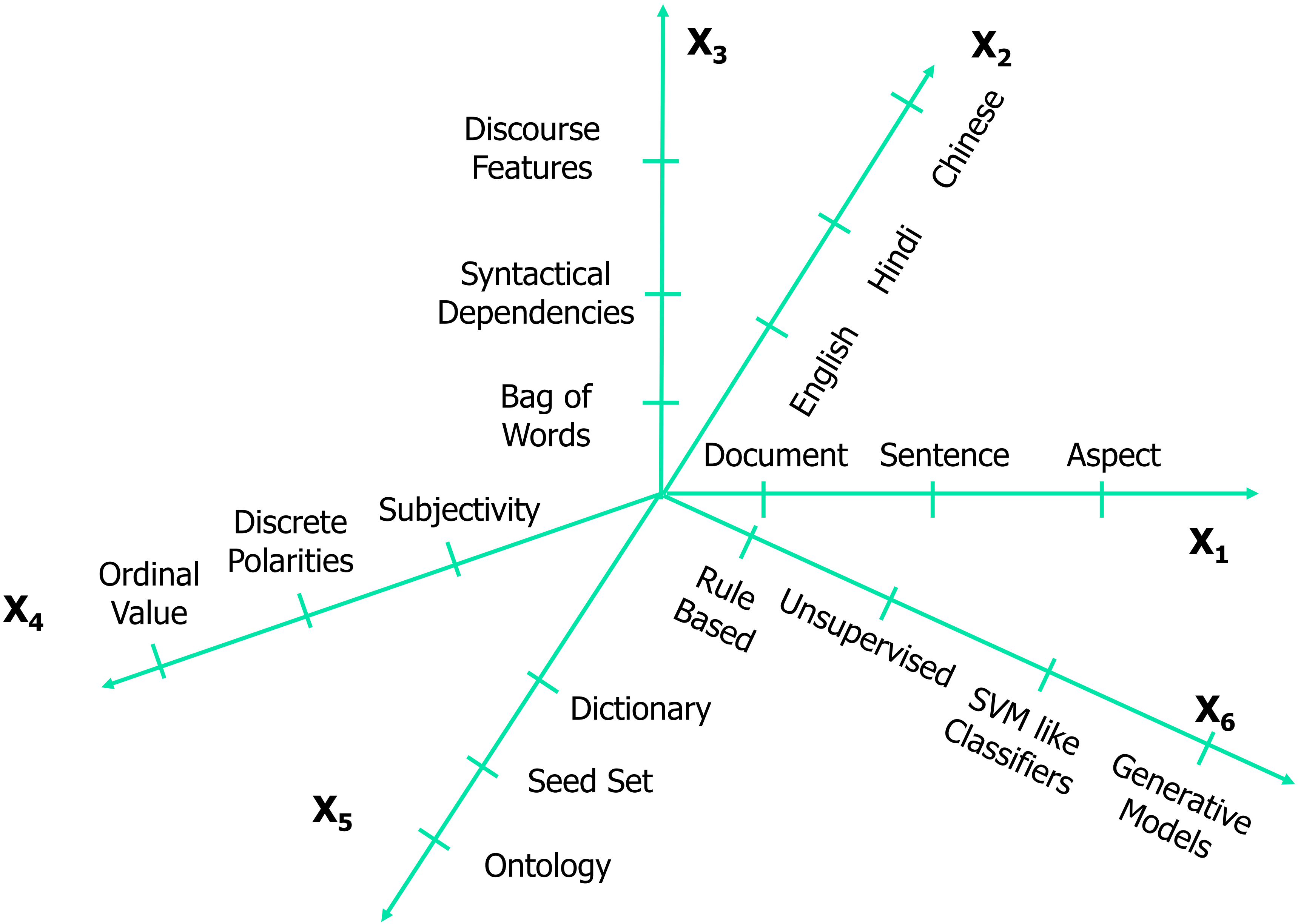
so_{ijkl} *is the sentiment value of the opinion*

of the opinion holder h_i

on feature f_{jk}

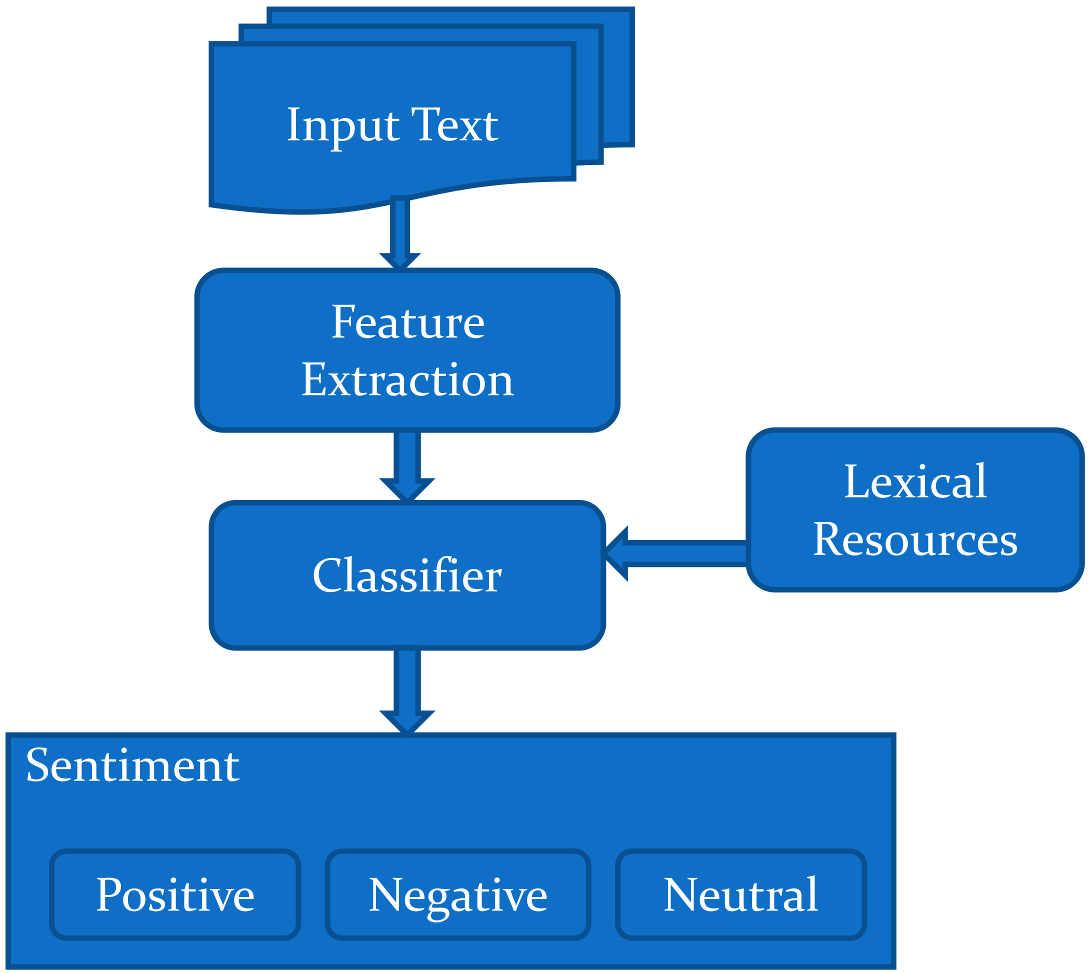
of object o_j

at time t_l



Dimensions of Sentiment Analysis

Block diagram



Challenges

`I suggest you wear your perfume with windows and doors shut! #sarcasm'

Sarcasm

`... falls 284 runs short of what would have been a fourth first-class triple-century'.

www.cricinfo.co

Implicit knowldege

`The movie may have the nicest actors, a talented music director of worldwide acclaim and the most expensive set one has ever seen but it fails to impress'.

Thwarting

`keeps you on the edge of your seat'

`Tim Tam. \m/'

Nature of text

`He is a deadly football player'

`You may have deadly snakes at the camp site at night`

Balamurali et al [2011

Domain specificity

Representative figures for SA Accuracy

Features	# of features	Frequency or Presence?	NB	ME	SVM
Unigrams	16165	Freq.	**78.7**	N/A	72.8
Unigrams	16165	Pres.	81.0	80.4	**82.9**
Unigrams+bigrams	32330	Pres.	80.6	80.8	**82.7**
Bigrams	16165	Pres.	77.3	**77.4**	77.1
Unigrams+POS	16695	Pres.	81.5	80.4	**81.9**
Adjectives	2633	Pres.	77.0	**77.7**	75.1
Top 2633 unigrams	2633	Pres.	80.3	81.0	**81.4**
Unigrams+position	22430	Pres.	81.0	80.1	**81.6**

Sarcasm

Etymology

- Greek: *'sarkasmós'*: 'to tear flesh with teeth'
- Sanskrit: *'vakrokti'*: 'a twisted (vakra) utterance (ukti)'

Definition- Foundation is *Irony*

Mean opposite of what is on surface

"A form of irony that is intended to express contempt or ridicule." The Free Dictionary	"Verbal irony that expresses negative and critical attitudes toward persons or events." (Kreuz and Glucksberg, 1989)
"The use of irony to mock or convey contempt." Oxford Dictionary	"Irony that is especially bitter and caustic" (Gibbs, 1994)

Types of Sarcasm

Sarcasm (Camp, 2012)

Propositional	Embedded	Like-prefixed	Illocutionary
A proposition that is intended to be sarcastic.	Sarcasm is embedded in the meaning of words being used.	'Like/As if' are common prefixes to ask rhetorical questions.	Non-speech acts (body language, gestures) contributing to the sarcasm
'This looks like a perfect plan!'	*'I love being ignored'*	*'Like you care'*	*'(shrugs shoulders) Very helpful indeed!'*

Tuple Representation for Sarcasm Ivanko and Pexman (2003)

<S, H, C, U, p, p'>

S	Speaker
H	Hearer
C	Context
U	Utterance
p	Literal Proposition
p'	Intended Proposition

"I love being ignored!"

S	The person referred to as by 'I'
H	The listener (say, host of a party)
C	Context
U	'I love being ignored'
p	'I love being ignored'
p'	'I do not like being ignored'

Impact on Sentiment Analysis (SA) (1/2)

Two SA systems:

MeaningCloud: https://www.meaningcloud.com/

NLTK (Bird, 2006)

Two datasets:

Sarcastic tweets by Riloff et al (2013)

Sarcastic utterances from our dataset of TV transcripts (Joshi et al 2016b)

Impact on Sentiment Analysis (2/2)

	Precision (Sarc)	Precision (Non-sarc)
Conversation Transcripts		
MeaningCloud[1]	20.14	49.41
NLTK (Bird, 2006)	38.86	81
Tweets		
MeaningCloud[1]	17.58	50.13
NLTK (Bird, 2006)	35.17	69

[1] www.meaningcloud.com

Clues for Sarcasm

- Use of laughter expression
  - *haha, you are very smart xD*
  - *Your intelligence astounds me. LOL*

- Heavy Punctuation
  - *Protein shake for dinner!! Great!!!*

- Use of emoticons
  - *i LOVE it when people tweet yet ignore my text X-(*

- Interjections
  - *3:00 am work YAY. YAY.*

- Capital Letters
  - *SUPER EXCITED TO WEAR MY UNIFORM TO SCHOOL TOMORROW ! ! :D lol.*

Incongruity: at the heart of things!

- *I love being ignored*
- *3:00 am work YAY. YAY.*
- *Up all night coughing. yeah me!*
- *No power, Yes! Yes! Thank you storm!*
- *This phone has an awesome battery back-up of 2 hour (Sarcastic)*

Two kinds of incongruity

- **Explicit incongruity**
  - Overtly expressed through sentiment words of both polarities
  - Contribute to almost 11% of sarcasm instances

    'I <u>love</u> being <u>ignored</u>'
- **Implicit incongruity**
  - Covertly expressed through phrases of implied sentiment

    'I <u>love</u> this paper so much that I <u>made a doggy bag out of it</u>'

Sarcasm Detection Using Semantic incongruity

Aditya Joshi, Vaibhav Tripathi, Kevin Patel, Pushpak Bhattacharyya and Mark Carman, *Are Word Embedding-based Features Useful for Sarcasm Detection?*, **EMNLP 2016**, Austin, Texas, USA, November 1-5, 2016.

Also covered in: How Vector Space Mathematics Helps Machines Spot Sarcasm, MIT Technology Review, 13th October, 2016.

www.cfilt.iitb.ac.in/sarcasmsuite/

Feature Set

Lexical	
Unigrams	Unigrams in the training corpus
Pragmatic	
Capitalization	Numeric feature indicating presence of capital letters
Emoticons & laughter expressions	Numeric feature indicating presence of emoticons and 'lol's
Punctuation marks	Numeric feature indicating presence of punctuation marks
Implicit Incongruity (Based on Riloff et al	
Implicit Sentiment Phrases	Boolean feature indicating phrases extracted from the implicit phrase extraction step
Explicit Incongruity (Based on Ramteke et al	
#Explicit incongruity	Number of times a word is followed by a word of opposite polarity
Largest positive /negative subsequence	Length of largest series of words with polarity unchanged
#Positive words	Number of positive words
#Negative words	Number of negative words
Lexical Polarity	Polarity of a tweet based on words present

Datasets

Name	Text-form	Method of labeling	Statistics
Tweet-A	Tweets	Using sarcasm-based hashtags as labels	5208 total, 4170 sarcastic
Tweet-B	Tweets	Manually labeled (Given by Riloff et al(2013))	2278 total, 506 sarcastic
Discussion-A	Discussion forum posts (IAC Corpus)	Manually labeled (Given by Walker et al (2012))	1502 total, 752 sarcastic

Results

Features	P	R	F
Original Algorithm by Riloff et al. (2013)			
Ordered	0.774	0.098	0.173
Unordered	0.799	0.337	0.474
Our system			
Lexical (**Baseline**)	0.820	0.867	0.842
Lexical+Implicit	0.822	0.887	0.853
Lexical+Explicit	0.807	0.985	0.8871
All features	0.814	0.976	**0.8876**

Tweet-A

Approach	P	R	F
Riloff et al. (2013) (**best reported**)	0.62	0.44	0.51
Maynard and Greenwood (2014)	0.46	0.38	0.41
Our system (all features)	**0.77**	**0.51**	**0.61**

Tweet-B

Features	P	R	F
Lexical (**Baseline**)	0.645	0.508	0.568
Lexical+Explicit	0.698	0.391	0.488
Lexical+Implicit	0.513	0.762	0.581
All features	0.489	0.924	**0.640**

Discussion-A

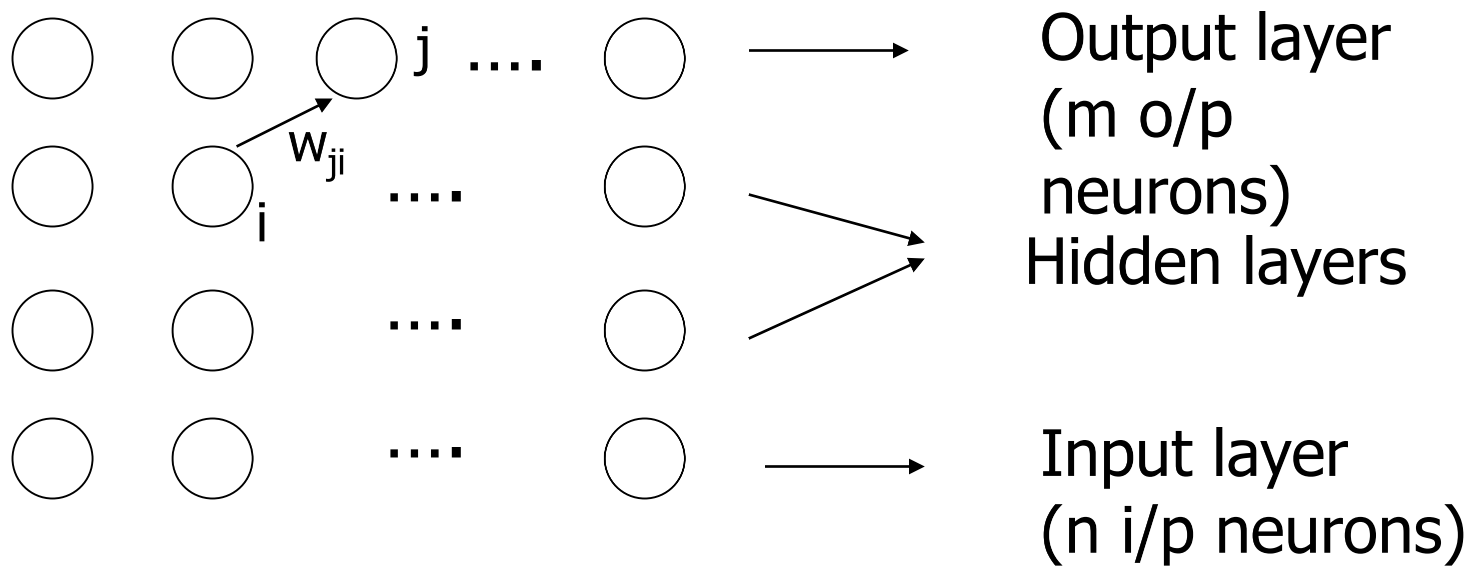
Inter-sentential incongruity

- Incongruity may be expressed between sentences.
- We extend our classifier for Discussion-A by considering posts before the target post. These posts are ‘elicitor posts’.

- Precision rises to 0.705 but the recall falls to 0.274.
  - Possible reason: Features become sparse since only 15% posts have elicitor posts

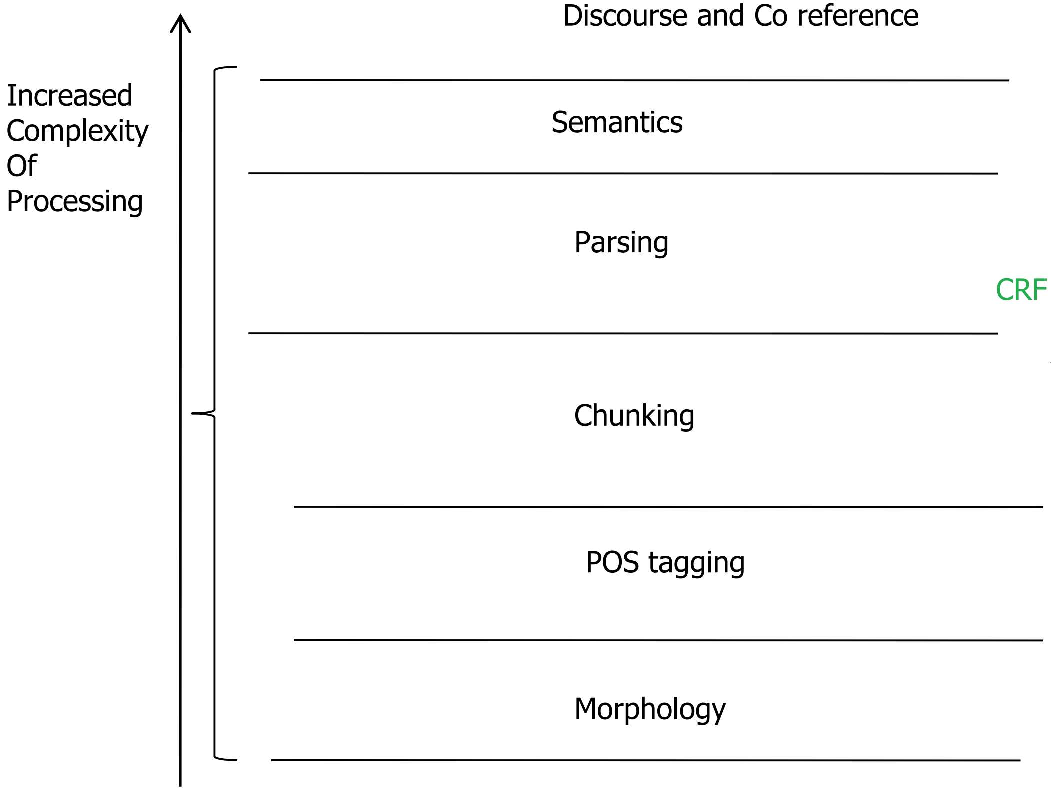
Sentiment and Deep Neural Nets

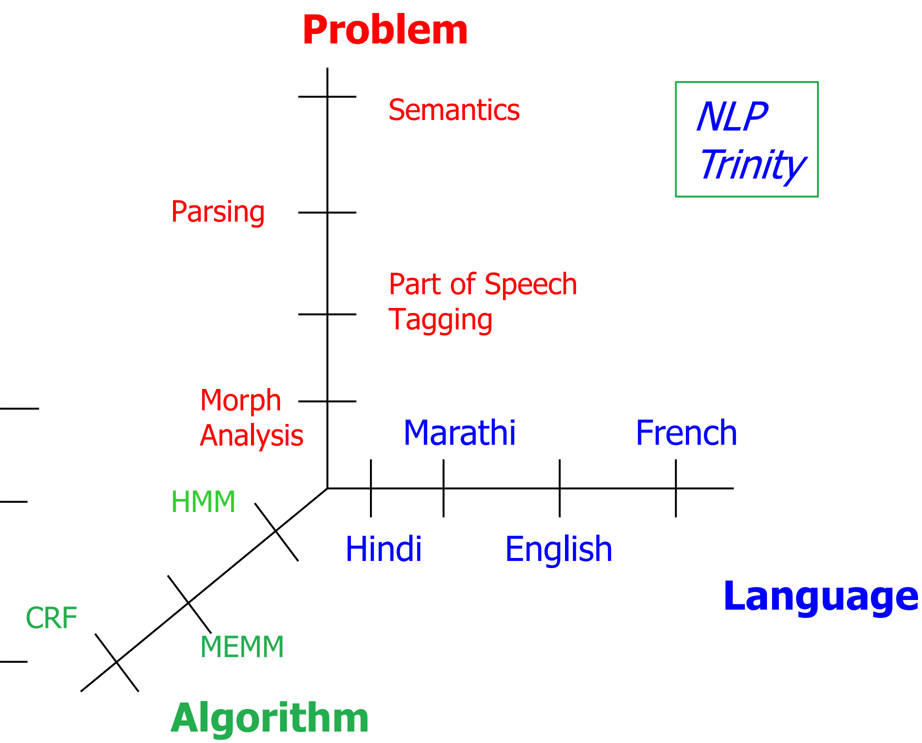
Deep neural net



- NLP pipeline ←→ NN layers
- Discover bigger structures bottom up, starting from character?
- Words, POS, Parse, Sentence, Discourse?

NLP: layered, multidimensional





Captuirng Incongruity Using Word Vectors

Some incongruity may occur without the presence of sentiment words

This can be captured using word embedding-based features, **in addition to other features**

"A man needs a woman like a fish needs bicycle."

Word2Vec similarity(man,woman) = 0.766
Word2Vec similarity(fish, bicycle) = 0.131

Word embedding-based features

Unweighted similarity features (S):
For every word and word pair,
1) Maximum score of most similar word pair
2) Minimum score of most similar word pair
3) Maximum score of most dissimilar word pair
4) Minimum score of most dissimilar word pair

Distance-weighted similarity features (WS): 4 S features weighted by linear distance between the two words

Both (S+WS): 8 features

Experiment Setup

- Dataset: 3629 Book snippets (759 sarcastic) downloaded from GoodReads website
- Labelled by users with tags
- Five-fold cross-validation
- Classifier: SVM-Perf optimised for F-score
- Configurations:
  - Four prior works (augmented with our sets of features)
  - Four implementations of word embeddings (Word2Vec, LSA, GloVe, Dependency weights-based)

Results (1/2)

Features	**P**	**R**	**F**
	Baseline		
Unigrams	67.2	78.8	72.53
S	64.6	75.2	69.49
WS	67.6	51.2	58.26
Both	67	52.8	59.05

	LSA			**GloVe**			**Dependency Weights**			**Word2Vec**		
	P	R	F	P	R	F	P	R	F	P	R	F
L	73	79	75.8	73	79	75.8	73	79	75.8	73	79	75.8
+S	81.8	78.2	**79.95**	81.8	79.2	**80.47**	81.8	78.8	80.27	80.4	80	**80.2**
+WS	76.2	79.8	77.9	76.2	79.6	77.86	81.4	80.8	81.09	80.8	78.6	79.68
+S+WS	77.6	79.8	78.68	74	79.4	76.60	82	80.4	**81.19**	81.6	78.2	79.86
G	84.8	73.8	78.91	84.8	73.8	78.91	84.8	73.8	**78.91**	84.8	73.8	**78.91**
+S	84.2	74.4	**79**	84	72.6	77.8	84.4	72	77.7	84	72.8	78
+WS	84.4	73.6	78.63	84	75.2	**79.35**	84.4	72.6	78.05	83.8	70.2	76.4
+S+WS	84.2	73.6	78.54	84	74	78.68	84.2	72.2	77.73	84	72.8	78
B	81.6	72.2	76.61	81.6	72.2	76.61	81.6	72.2	76.61	81.6	72.2	76.61
+S	78.2	75.6	**76.87**	80.4	76.2	**78.24**	81.2	74.6	**77.76**	81.4	72.6	76.74
+WS	75.8	77.2	76.49	76.6	77	76.79	76.2	76.4	76.29	81.6	73.4	77.28
+S+WS	74.8	77.4	76.07	76.2	78.2	77.18	75.6	78.8	77.16	81	75.4	**78.09**
J	85.2	74.4	79.43	85.2	74.4	79.43	85.2	74.4	79.43	85.2	74.4	79.43
+S	84.8	73.8	78.91	85.6	74.8	79.83	85.4	74.4	79.52	85.4	74.6	**79.63**
+WS	85.6	75.2	**80.06**	85.4	72.6	78.48	85.4	73.4	78.94	85.6	73.4	79.03
+S+WS	84.8	73.6	78.8	85.8	75.4	**80.26**	85.6	74.4	**79.6**	85.2	73.2	78.74

Table 3: Performance obtained on augmenting word embedding features to features from four prior works, for four word embeddings; L: Liebrecht et al. (2013), G: González-Ibánez et al. (2011a), B: Buschmeier et al. (2014) , J: Joshi et al. (2015)

Results (2/2)

	Word2Vec	LSA	GloVe	Dep. Wt.
+S	0.835	0.86	0.918	**0.978**
+WS	**1.411**	0.255	0.192	1.372
+S+WS	**1.182**	0.24	0.845	0.795

Table 4: Average gain in F-Scores obtained by using intersection of the four word embeddings, for three word embedding feature-types, augmented to four prior works; Dep. Wt. indicates vectors learned from dependency-based weights

Word Embedding	Average F-score Gain
LSA	0.452
Glove	0.651
Dependency	1.048
Word2Vec	1.143

Table 5: Average gain in F-scores for the four types of word embeddings; These values are computed for a subset of these embeddings consisting of words common to all four

Numerical Sarcasm

About 17% of sarcastic tweets have origin in number

- This phone has an awesome battery back-up of 38 hours (Non-sarcastic)
- This phone has an awesome battery back-up of 2 hour (Sarcastic)
- This phone has a terrible battery back-up of 2 hours (Non-sarcastic)

Numerical Sarcasm

- waiting 45 min for the subway in the freezingcold is so much fun.
- well 3 hrs of sleep this is awesome.
- gotta read 50 pages and do my math before tomorrow i'm so excited.
- -28 c with the windchill fantastic 2 weeks.
- woooo when you're up to 12:30 finishing you're english paper.

Numerical Sarcasm Dataset

Dataset-1	100000 (Sarcastic)	250000 (Non-Sarcastic)
Dataset-2	8681 (Num Sarcastic)	8681 (Non-Sarcastic)
Dataset-3	8681 (Num Sarcastic)	42107 (Non-Sarcastic)
Test Data	1843 (Num Sarcastic)	8317 (Non-Sarcastic)

- To create this dataset, we extract tweets from Twitter-API (https://dev.twitter.com).
- Hashtags of the tweets served as labels *#sarcasm #sarcastic* etc.
- Dataset-1 contains normal sarcastic + numeric sarcastic and non-sarcastic tweets.
- Rest all the other dataset contains numeric sarcastic and non-sarcastic tweets only.

Systems for Numerical Sarcasm Detection

- Rule-based System
- Machine Learning System
- Deep Learning System

Rule-based System (Matching of NPs)

- Two repositories:

- Sarcastic and non-sarcastic using a training dataset

- Each tuple in the repository is of the format:

  (Tweet No., Noun Phrase list, Number, Number Unit)

Rule-based System (NP-Exact Matching)

- Extract noun phrases in the tweet, using a nltk parser
- Select the word in the tweet POS tagged as 'CD' as the number and the word in the tweet following the number as the number unit[1]

[1] In case there are more than one numbers in the tweet, we randomly select one.

Example

"This phone has an awesome battery back-up of 2 hours",

```
(S
  This/DT
  (NP (NBAR phone/NN))
  has/VBZ
  an/DT
  (NP (NBAR awesome/JJ battery/NN backup/NN))
  of/IN
  2/CD
  (NP (NBAR hours/NNS)))
```

Example (cntd.)

- Noun Phrases:

  ['phone', 'awesome', 'battery', 'backup', 'hours']

- Addition to sarcastic repository:

  (Tweet No., ['phone', 'awesome', 'battery', 'backup', 'hours'], 2, 'hours')

Algorithm (match sarcastic respository)

- Consult the sarcastic tweet repository
- Match words in the noun phrase list between the test tweet and entries in the repository
- Select the most similar entry from the sarcastic repository
- If numbers are close, *sarcastic* else *non-sarcastic*

Algorithm (match non-sarcastic respository)

- Search and do as in case of sarcastic reposirtory
- Get most similar tweet
- If numbers are ***FAR APART***, sarcastic else non-sarcastic

Rule-based System (NP-Exact Matching) (Cont'd)

- Test Tweet: ‘I love writing this paper at 9 am‘
- Matched Sarcastic Tweet: ‘I love writing this paper daily at 3 am‘
- 9 ***NOT*** close to 3

test tweet is ***non-sarcastic***

Example (sarcastic case)

- Test Tweet: 'I am so productive when my room is 81 degrees'
- Matched Non-sarcastic Tweet: 'I am very much productive in my room as it has 21 degrees'
- Absolute difference between 81 and 21 is high

Hence test tweet is Sarcastic

Comparing this simple approach

Approaches	**Overall Precision**	**Overall Recall**	**Overall F1-Score**
Buschmeier et.al.	0.84	0.24	0.16
Gonzalez-Ibanez et.al.	0.83	0.23	0.15
Liebrecht et.al.	0.85	0.24	0.17
Joshi et.al.	0.86	0.29	0.25
Exact-NP-Matching (Rule-based)	**0.81**	**0.83**	**0.82**

Machine Learning based approach: classifiers and features

- SVM, KNN and Random Forest classifiers
- Sentiment-based features
  - Number of
    - positive words
    - negative words
    - highly emotional positive words,
    - highly emotional negative words.
- Positive/Negative word is said to be highly emotional if it's POS tag is one amongst : 'JJ', 'JJR', 'JJS', 'RB', 'RBR', 'RBS', 'VB', 'VBD', 'VBG', 'VBN', 'VBP', 'VBZ'.

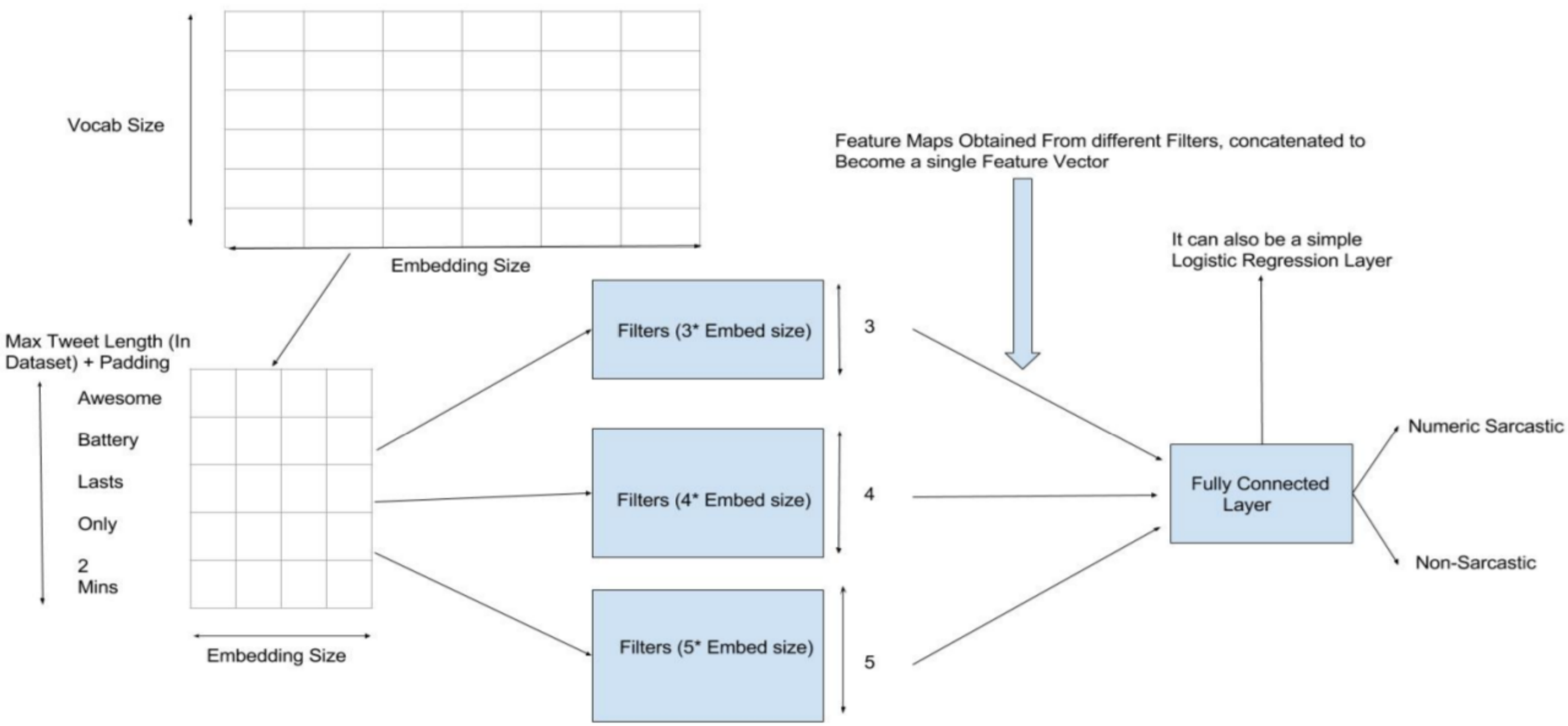
Emotion Features

- Positive emoticon
- Negative emoticon
- Boolean feature that will be one if both positive and negative words are present in the tweet.
- Boolean feature that will be one when either positive word and negative emoji is present or vice versa.

Punctuation features

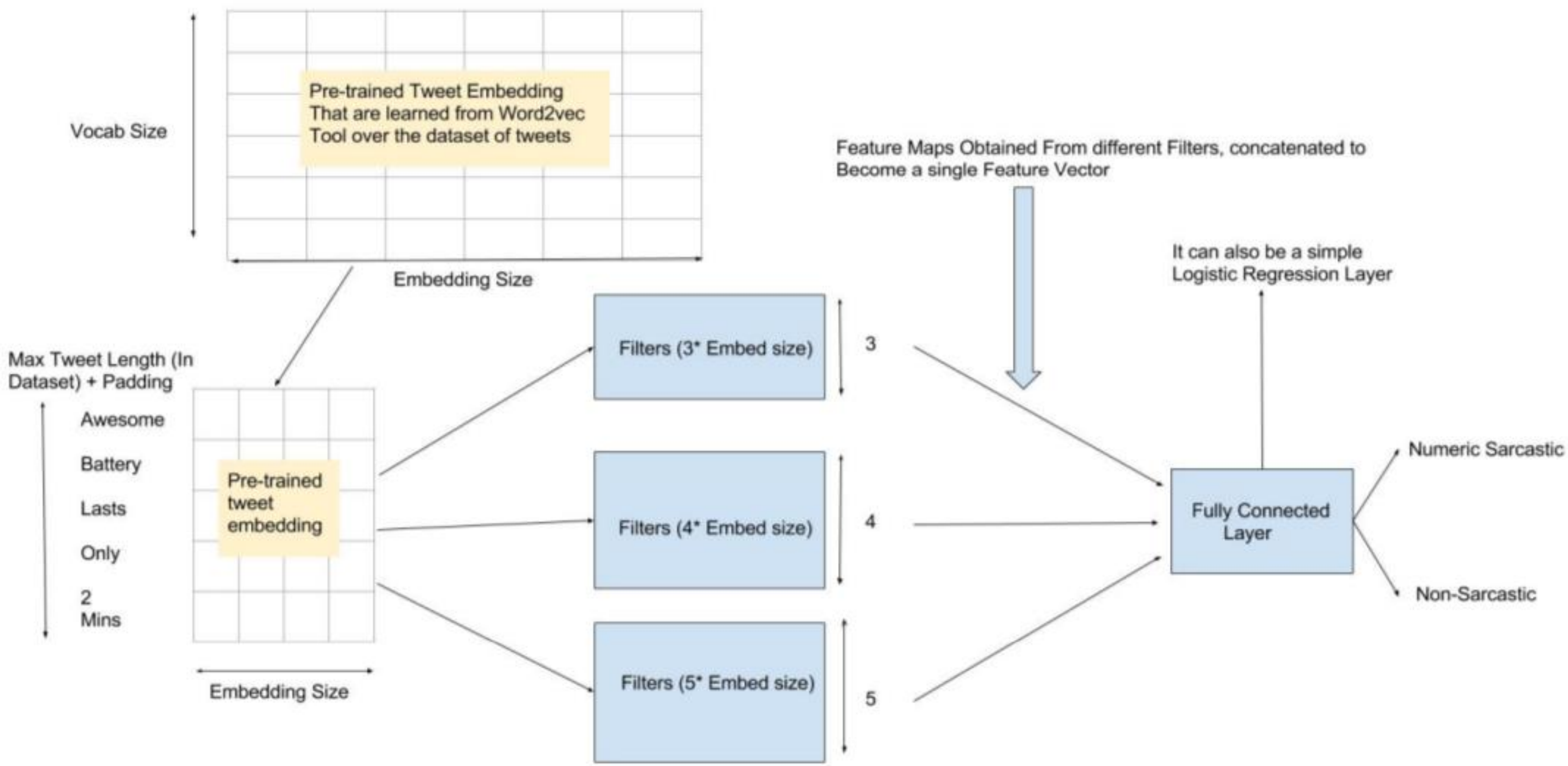
- number of exclamation marks.
- number of dots
- number of question mark.
- number of capital letter words.
- number of single quotations.
- Number in the tweet: This feature is simply the number present in the tweet.
- Number unit in the tweet : This feature is a one hot representation of the type of unit present in the tweet. Example of number unit can be hour, minute, etc.

Deep Learning based approach: CNN-FF Model



Deep Learning based approach (Cont'd)

- EmbeddingSize of 128
- Maximum tweet length 36 words
- Padding used
- Filters of size 3, 4, 5 used to extarct features



Vocab Size
Pre-trained Tweet Embedding
That are learned from Word2vec
Tool over the dataset of tweets
Embedding Size
Feature Maps Obtained From different Filters, concatenated to
Become a single Feature Vector
It can also be a simple
Logistic Regression Layer
Max Tweet Length (In
Dataset) + Padding
Awesome
Battery
Lasts
Only
2
Mins
Pre-trained
tweet
embedding
Embedding Size
Filters (3* Embed size)
3
Filters (4* Embed size)
4
Filters (5* Embed size)
5
Fully Connected
Layer
Numeric Sarcastic
Non-Sarcastic

Comparison of results (1: sarcastic, 0: non-sarcastic)

Approaches	Precision			Recall			F-score		
	P(1)	**P(0)**	**P(avg)**	**R(1)**	**R(0)**	**R(avg)**	**F(1)**	**F(0)**	**F(avg)**
Past Approaches									
Buschmeier et.al.	0.19	0.98	0.84	0.99	0.07	0.24	0.32	0.13	0.16
Liebrecht et.al.	0.19	1.00	0.85	1.00	0.07	0.24	0.32	0.13	0.17
Gonzalez et.al.	0.19	0.96	0.83	0.99	0.06	0.23	0.32	0.12	0.15
Joshi et.al.	0.20	1.00	0.86	1.00	0.13	0.29	0.33	0.23	**0.25**
Rule-Based Approaches									
Approach-1	0.53	0.87	0.81	0.39	0.92	0.83	0.45	0.90	**0.82**
Approach-2	0.44	0.85	0.78	0.28	0.92	0.81	0.34	0.89	0.79
Machine-Learning Based Approaches									
SVM	0.50	0.95	0.87	0.80	0.82	0.82	0.61	0.88	**0.83**
KNN	0.36	0.94	0.84	0.81	0.68	0.70	0.50	0.79	0.74
Random Forest	0.47	0.93	0.85	0.74	0.81	0.80	0.57	0.87	0.82
Deep-Learning Based Approaches									
CNN-FF	**0.88**	**0.94**	**0.93**	**0.71**	**0.98**	**0.93**	**0.79**	**0.96**	**0.93**
CNN-LSTM-FF	0.82	0.94	0.92	0.72	0.96	0.92	0.77	0.95	0.92
LSTM-FF	0.76	0.93	0.90	0.68	0.95	0.90	0.72	0.94	0.90

Case Studies Examples

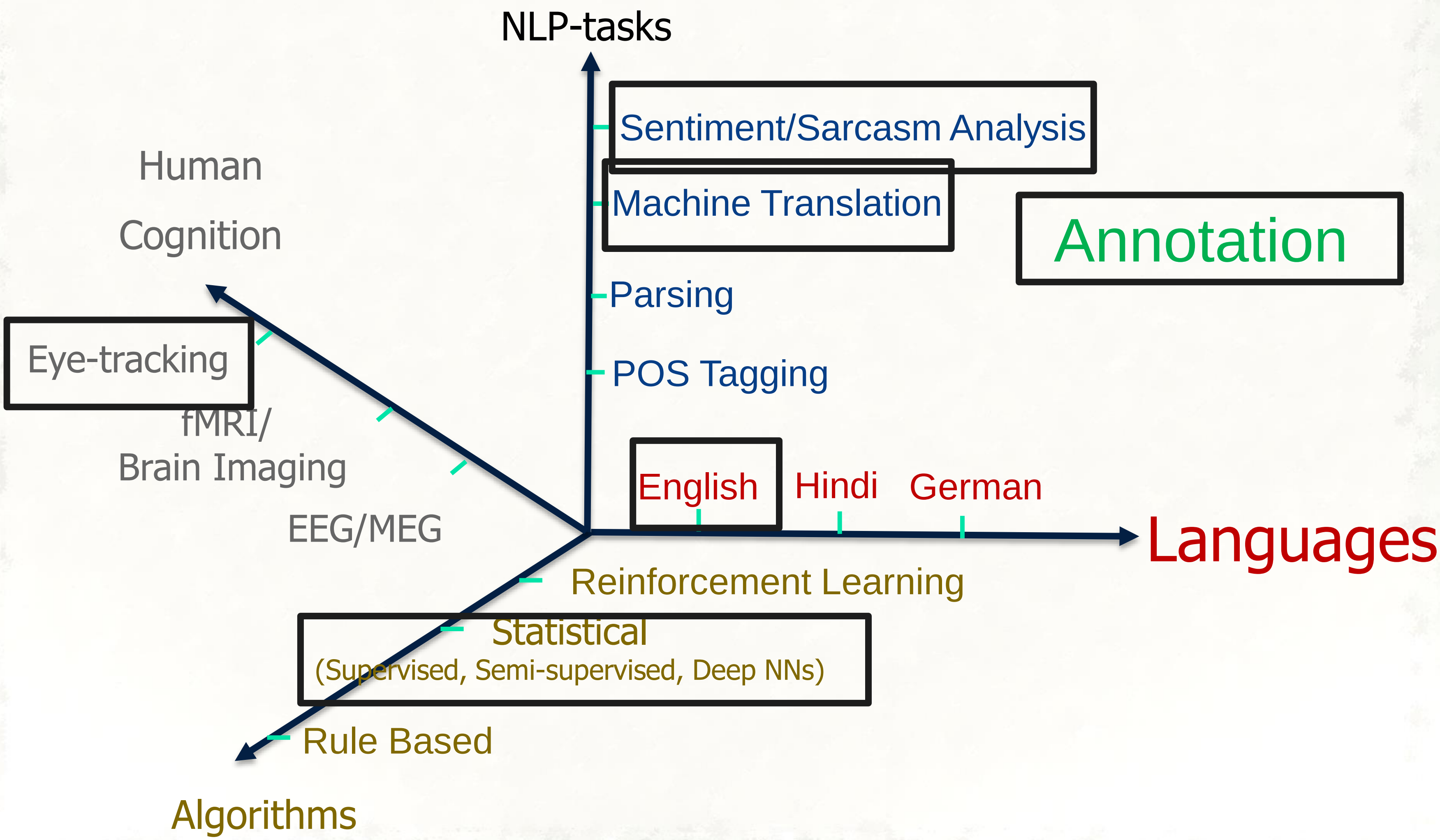
- “waiting 45 min for the subway in the freezing cold is so much fun iswinteroveryet”
  - Classified as Numeric Sarcastic only by Deep learning based classifier
- “unspeakably excited to take a four hour practice act for the 4th time.”
  - Classified as Numeric Sarcastic by both the CNN architectures only.
- "yeah wasted $3 to go two stops thanks for the service ttc crapservice.”
  - Classified as Numeric Sarcastic only by Deep learning based classifier.

Failure Examples

- “my mother has the talent of turning a 10 minute drive into a 25 minute drive needforspeed”.
- “arrived at school 6:30 this morning yeah we have an easy life we work 8-3 @ john h”.
- “woke up to hrs ago and i can barely keep my eyes open best part of my day i don't get home til 7 pm”.
- “hey airlines i really appreciate you canceling my direct flight home and sending me 1000 miles out of the way to connect”.

Enter cognition

NLP-trinity



NLP-tasks
Sentiment/Sarcasm Analysis
Machine Translation
Parsing
POS Tagging
Annotation
Human
Cognition
Eye-tracking
fMRI/
Brain Imaging
EEG/MEG
English
Hindi
German
Languages
Reinforcement Learning
Statistical
(Supervised, Semi-supervised, Deep NNs)
Rule Based
Algorithms

Eye-tracking Technology

Invasive and non-invasive eye-trackers

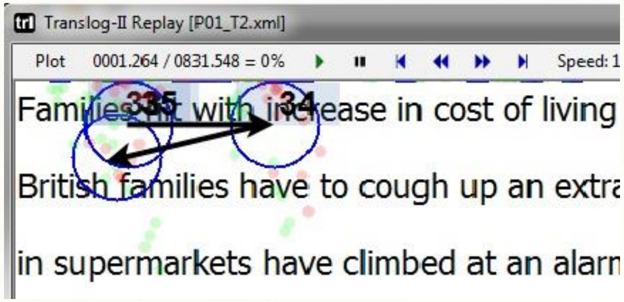
(image - sources: http://www.tobii.com/)

For linguistic studies non-invasive eye-trackers are used

- **Data delivered by eye-trackers**
  - Gaze co-ordinates of both eyes (binocular setting) or single eye (monocular setting)
  - Pupil size
- **Derivable data**
  - Fixations, Saccades, Scanpaths, Specific patterns like progression and regression.

Nature of Gaze Data

- **Gaze Point:** Position (co-ordinate) of gaze on the screen
- **Fixations :** A long stay of the gaze on a particular object on the screen
- **Saccade:** A very rapid movement of eye between the positions of rest.
  - Progressive Saccade / Forward Saccade / Progression
  - Regressive Saccade / Backward Saccade / Regression
- **Scanpath:** A path connecting a series of fixations.



Eye-movement and Cognition

- Eye-Mind Hypothesis (Just and Carpenter, 1980)

When a subject is views a word/object, he or she also processes it cognitively, for approximately the same amount of time he or she fixates on it.

- Considered useful in explaining theories associated with reading (Rayner and Duffy,1986; Irwin, 2004; von der Malsburg and Vasishth, 2011)

- Linear and uniform-speed gaze movement is observed over texts having simple concepts, and often non-linear movement with non-uniform speed over more complex concepts (Rayner, 1998)

Harnessing Cognitive Features for Sarcasm Detection (Mishra and Bhattacharyya, ACL 2016)

Features for Sarcasm: Augmented with cognitive

Textual

(1) Unigrams (2) Punctuations
(3) Implicit incongruity
(4) Explicit Incongruity
(5) Largest +ve/-ve subsequences
(6) +ve/-ve word count
(7) Lexical Polarity
(8) Flesch Readability Ease,
(9) Word count

Simple gaze

(1) Average Fixation Duration,
(2) Average Fixation Count,
(3) Average Saccade Length,
(4) Regression Count,
(5) Number of words skipped,
(6) Regressions from second half to first half,
(7) Position of the word from which the largest regression starts

Complex gaze

(1) Edge density,
(2) Highest weighted degree
(3) Second Highest weighted degree
(With different edge-weights)

Experiment Setup

- **Dataset**:
  - 994 text snippets : 383 positive and 611 negative, 350 are sarcastic/ironic
  - Mixture of Movie reviews, Tweets and sarcastic/ironic quotes
  - Annotated by 7 human annotators
  - Annotation accuracy: **70%-90%** with Fleiss kappa IAA of **0.62**
- **Classifiers**:
  - Naïve Bayes, SVM, Multi Layered Perceptron
  - Feature combinations:
    - Unigram Only
    - Gaze Only (Simple + Complex)
    - Textual Sarcasm Features (Joshi et., al, 2015) (Includes unigrams)
    - Gaze+ Sarcasm
- **Compared with** : **Riloff, 2013 and Joshi, 2015**

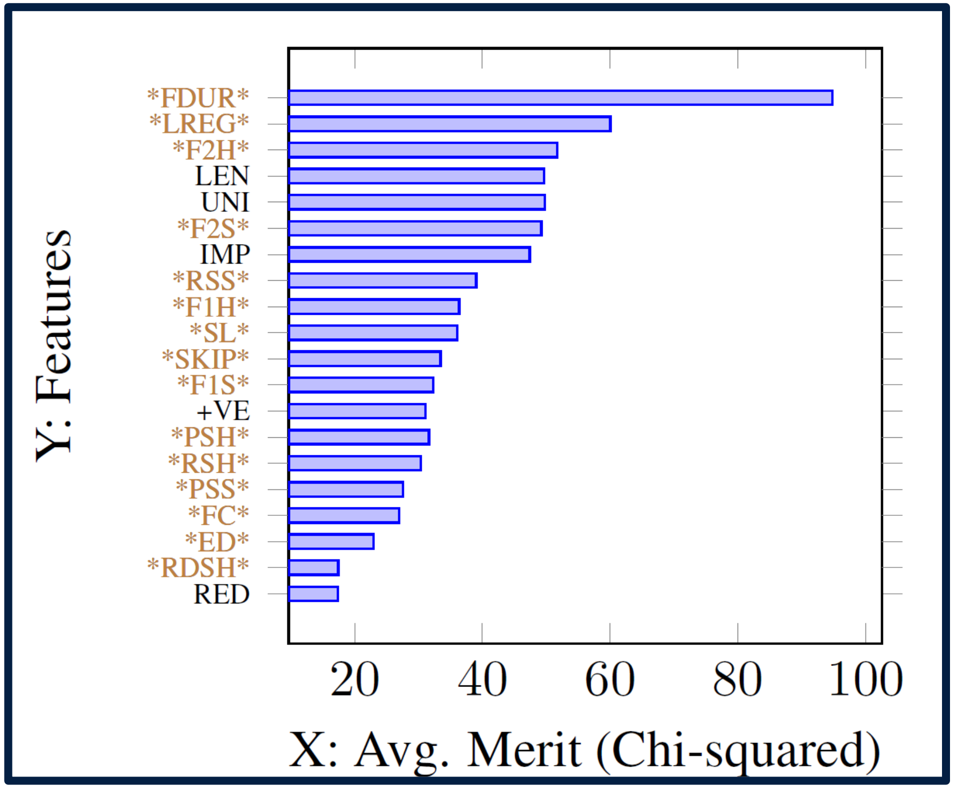
Results

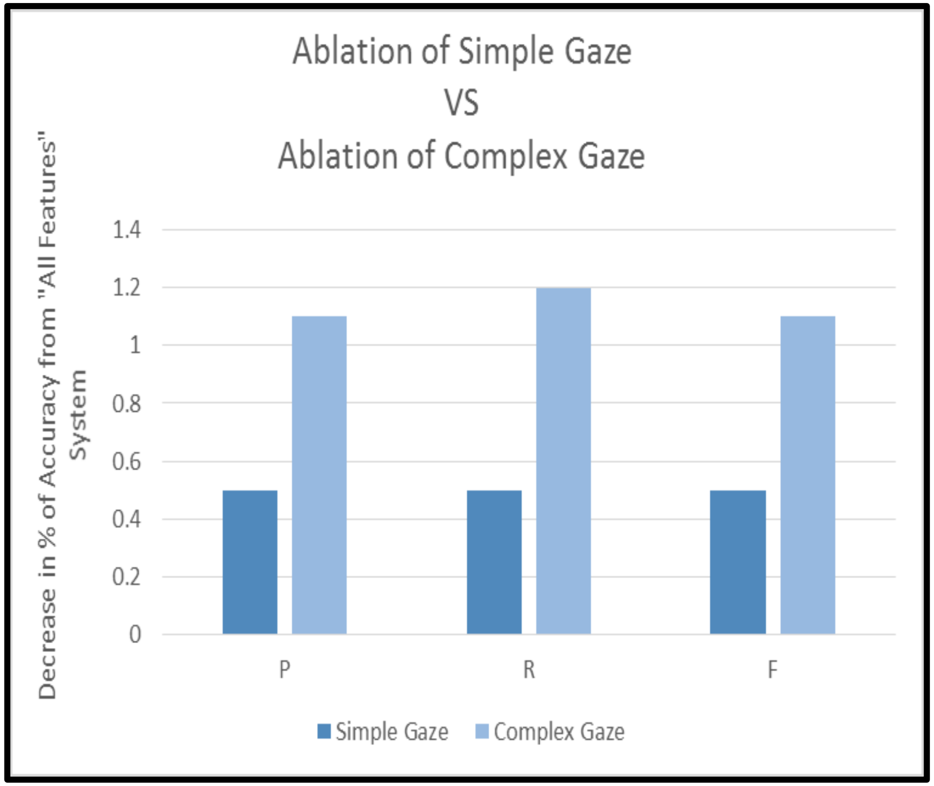
Features	P(1)	P(-1)	P(avg)	R(1)	R(-1)	R(avg)	F(1)	F(-1)	F(avg)
Multi Layered Neural Network									
Unigram	53.1	74.1	66.9	51.7	75.2	66.6	52.4	74.6	66.8
Sarcasm (Joshi et. al.)	59.2	75.4	69.7	51.7	80.6	70.4	55.2	77.9	69.9
Gaze	62.4	76.7	71.7	54	82.3	72.3	57.9	79.4	71.8
Gaze+Sarcasm	63.4	75	70.9	48	84.9	71.9	54.6	79.7	70.9
Näive Bayes									
Unigram	45.6	82.4	69.4	81.4	47.2	59.3	58.5	60	59.5
Sarcasm (Joshi et. al.)	46.1	81.6	69.1	79.4	49.5	60.1	58.3	61.6	60.5
Gaze	57.3	82.7	73.8	72.9	70.5	71.3	64.2	76.1	71.9
Gaze+Sarcasm	46.7	82.1	69.6	79.7	50.5	60.8	58.9	62.5	61.2
Original system by Riloff et.al. : Rule Based with implicit incongruity									
Ordered	60	30	49	50	39	46	54	34	47
Unordered	56	28	46	40	42	41	46	33	42
Original system by Joshi et.al. : SVM with RBF Kernel									
Sarcasm (Joshi et. al.)	73.1	69.4	70.7	22.6	95.5	69.8	34.5	80.4	64.2
SVM Linear: with default parameters									
Unigram	56.5	77	69.8	58.6	75.5	69.5	57.5	76.2	69.6
Sarcasm (Joshi et. al.)	*59.9*	*78.7*	*72.1*	*61.4*	*77.6*	*71.9*	*60.6*	*78.2*	*72*
Gaze	**65.9**	75.9	72.4	49.7	86	73.2	56.7	80.6	72.2
Gaze+Sarcasm	63.7	79.5	74	61.7	80.9	74.1	62.7	80.2	74
Multi Instance Logistic Regression: Best Performing Classifier									
Gaze	65.3	77.2	73	53	**84.9**	73.8	58.5	**80.8**	73.[illegible]
Gaze+Sarcasm	62.5	**84**	**76.5**	**72.6**	76.7	**75.3**	67.2	80.2	**75.[illegible]**

p=0.01

p=0.03

Feature Significance





Predicting Readers' Sarcasm Understandability By Modeling Gaze Behavior (Mishra and Bhattacharyya, AAAI 2016)

Sarcasm, Cognition and Eye-movement

- Sarcasm often emanates from **context incongruity** (Campbell and Katz 2012), which, possibly, surprises the reader and enforces a re-analysis of the text.

- In the absence of any information, human brain would start processing the text in a sequential manner, with the aim of comprehending the literal meaning.

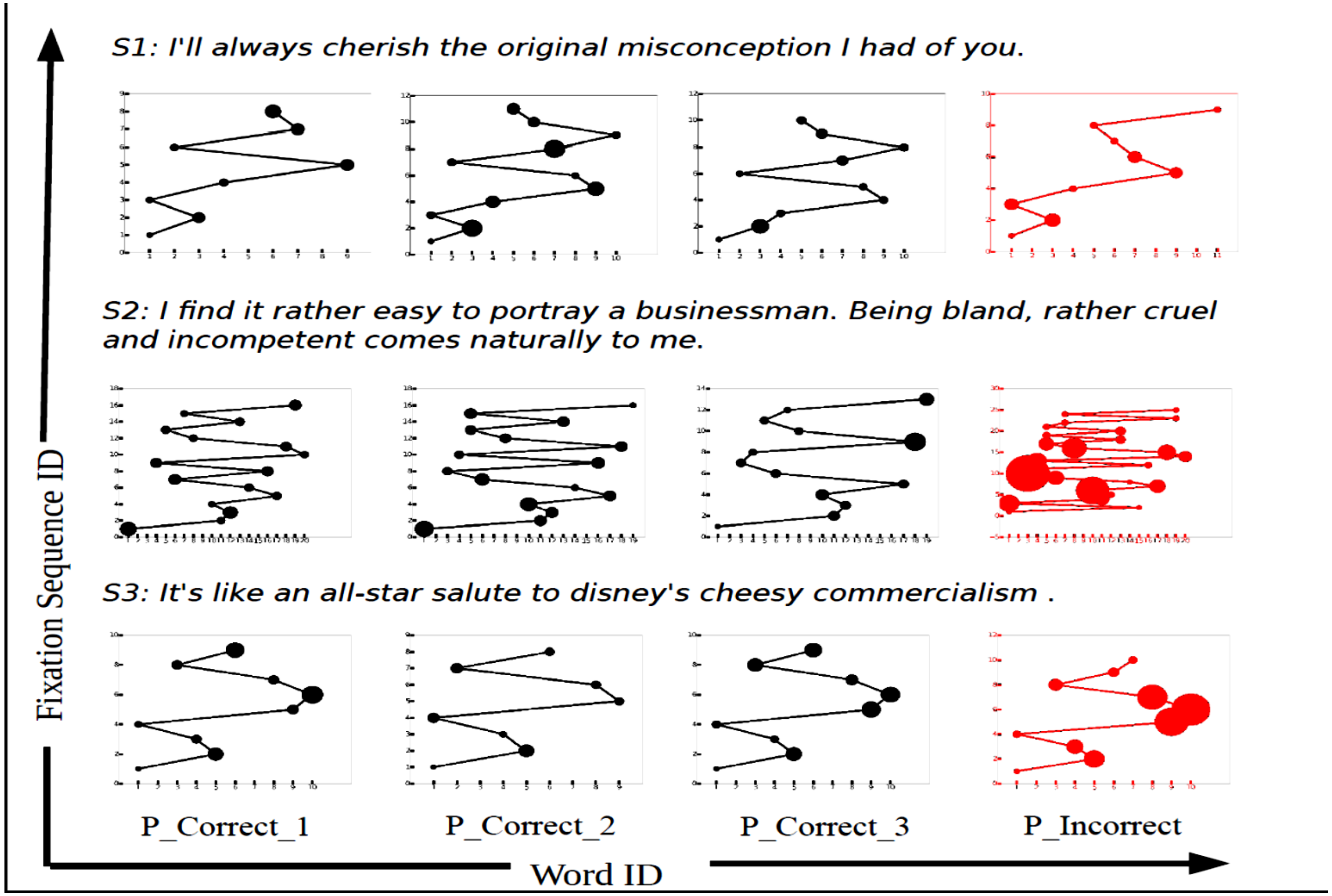
- When incongruity is perceived, the brain initiates a re-analysis to reason out such disparity (Kutas et al.,1980).

Hypothesis: Incongruity may affect the way eye-gaze moves through the text. Hence, distinctive eye-movement patterns may be observed when sarcasm is understood in contrast to an unsuccessful attempt.

Creation of Eye-movement Dataset

- **Document Description:**1000 short texts – Movie reviews, tweets and quotes, 350 sarcastic 650 non-sarcastic
- Ground truth verified by linguists. Grammatical mistakes corrected to avoid reading difficulties.
- **Participant Description:** 7 graduates from Engineering and Science background.
- **Task Description:** Texts annotated with sentiment polarity labels. Gaze data collected using Eye-link 1000 plus tracker following standard norms (Holmqvist et al. 2011)
- **Annotation Accuracy (IAA):** Highest- **90.29%,** Lowest- **72.57%,** Average- **84.64**% (Domain wise: Movie: **83.27%**, Quote: **83.6%**, Twitter: **84.88%)**

Sarcasm Understandability – Scanpath Representation



Analysis of Eye-movement Data

- **Variation in Basic Gaze attributes:** Average Fixation Duration and Number of Regressive Saccades significantly higher ($p<0.0001$ and $p<0.01$) when sarcasm is not understood than when it is.

- **Variation in Scanpaths:** For two incongruous phrases A and B, Regressive Saccades often seen from B to A when sarcasm is successfully realized. Moreover, Fixation duration is more on B than A.

- **Qualitative observations from Scanpaths:** Sarcasm not understood due to: (i) Lack of attention (ii) Lack of realization of context incongruity

Features for Sarcasm Understandability

Textual Features	Gaze Features
(1) # of interjections	(1) Avg. Fixation Duration (AFD)
(2) # of punctuations	(2) Avg. Fixation Count
(3) # of discourse connectors	(3) Avg. Saccade Length
(4) # of flips in word polarity	(4) # of Regressions
(5) Length of the Largest Pos/Neg Subsequence	(5) # of words skipped
(6) # of Positive words	(6) AFD on the 1st half of the text
(7) # of Negative words	(7) AFD on the 2nd half of the text
(8) Flecsh's reading ease score	(8) # of regressions from the 2nd half to the 1st half
(9) Number of Words	(9) Position of the word from which the longest regression happens.
	(10) Scanpath Complexity

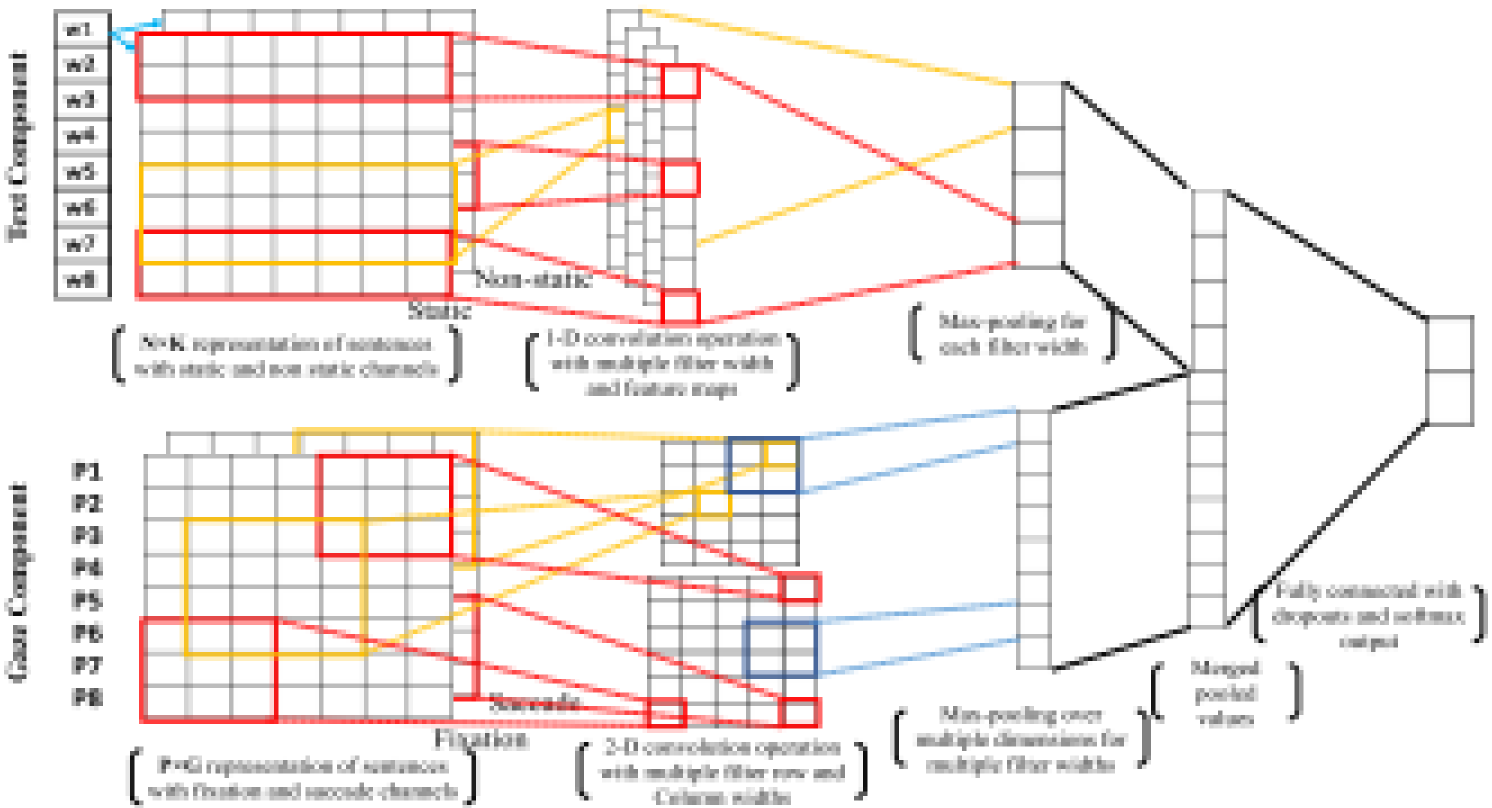
Experiment and Results

- **Classifier:** Multi-instance Logistic Regression (Xu and Frank 2004). Each training example corresponds to one sentence. Each example "bags" a maximum of 7 instances, one for each participant. Each instance is a combination of Gaze and Textual Features.

Class	sarcasm_miss			sarcasm_hit			Weighted Avg.			Kappa
	P	R	F	P	R	F	P	R	F	Avg.
	Baseline1: Classification based on class frequency									
All	16.1	15.5	15.7	86.5	87	86.7	85.9	86.71	86.3	0.014
	Baseline2: MILR Classifier considering time taken to read + textual features									
All	23.6	86.9	78.2	11.5	94.1	82.7	15.4	90.4	80	0.0707
	Our approach: MILR Classifier considering only gaze features									
All	82.6	36	50	89.9	98.7	94.1	88.8	89.4	87.5	0.4517
	Our approach: MILR Classifier considering gaze + textual features									
Quote	68.1	47.5	56.0	91.8	96.3	94.0	88.4	89.4	88.6	0.5016
Movie	42.9	36.6	39.5	88.6	91.0	89.8	81.4	82.5	81.9	0.293
Twitter	63.0	61.7	62.4	94.4	94.7	94.6	90.4	90.5	90.5	0.5695
All	**87.8**	**61**	**72**	**94.1**	**98.6**	**96.3**	**93.2**	**93.5**	**93**	**0.6845**

Abhijit Mishra, Kuntal Dey and Pushpak Bhattacharyya, *Learning Cognitive Features from Gaze Data for Sentiment and Sarcasm Classification Using Convolutional Neural Network*, **ACL 2017**, Vancouver, Canada, July 30-August 4, 2017.

CNN-FF combination



Results: Sarcasm Detection

	Configuration	Precision	Recall	F_Score
Gaze	Gaze-Fixation	74.39	69.62	**71.93**
	Gaze-Saccade	68.58	68.23	68.40
	Gaze-Multi-channel	67.93	67.72	67.82
Text	Text-static	67.17	66.38	66.77
	Text-non-static	84.19	87.03	**85.59**
	Text-Multi-channel	84.28	87.03	**85.63**
Gaze & Text	Text-static_Gaze-Fixation	72.38	71.93	72.15
	Text-static_Gaze-Saccade	73.12	72.14	72.63
	Text-static_Gaze-Multi-channel	71.41	71.03	71.22
	Text-non-static_Gaze-Fixation	87.42	85.2	**86.30**
	Text-non-static_Gaze-Saccade	84.84	82.68	83.75
	Text-non-static_Gaze-Multi-channel	84.98	82.79	83.87
	Text-Multi-channel_Gaze-Fixation	87.03	86.92	**86.97**
	Text-Multi-channel_Gaze-Saccade	81.98	81.08	81.53
	Text-Multi-channel_Gaze-Multi-channel	83.11	81.69	82.39

(a) Results with Deep CNNs

Configuration	Precision	Recall	F_Score
Gaze_NB	73.8	71.3	71.9
Gaze_SVM	**72.4**	**73.2**	**72.2**
Gaze_MLP	71.7	72.3	71.8

(b) CoNLL systems with Gaze Features

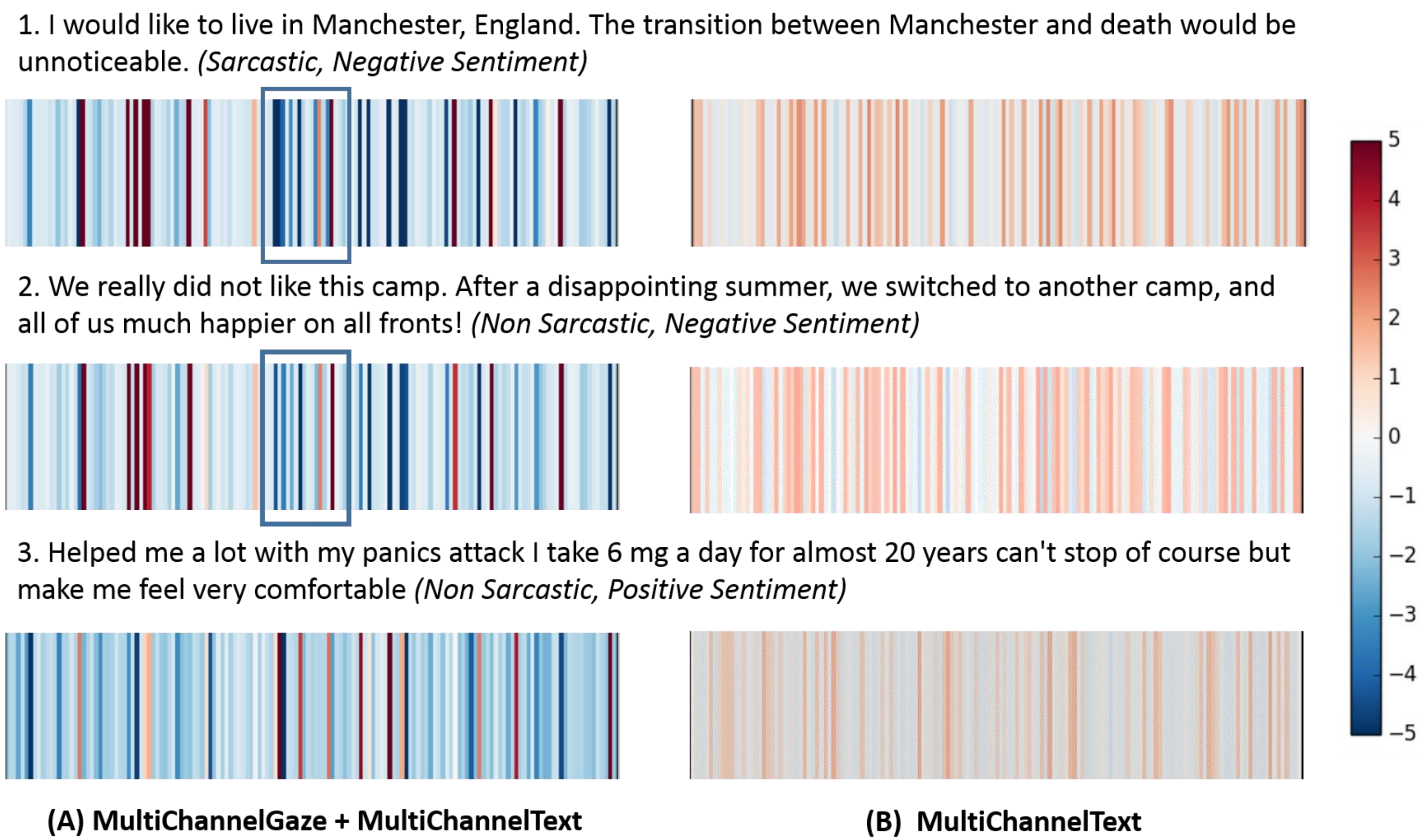
Configuration	Precision	Recall	F_Score
Gaze_Text_NB	**70.9**	71.9	71.2
Gaze_Text_SVM	**74**	**74.1**	**74**
Gaze_Text_MLP	70.9	71.9	70.9

(c) CoNLL systems with Gaze+Text Features

Observations - Sarcasm

- **Higher classification accuracy**
  - Clear differences between vocabulary of sarcasm and no-sarcasm classes in our dataset., Captured well by non-static embeddings.
- **Effect of dimension variation**
  - Reducing embedding dimension improves accuracy by a little margin.
- **Effect of fixation / saccade channels:**
  - Fixation and saccade channels perform with similar accuracy when employed separately.
  - Accuracy reduces with gaze multichannel (may be because the higher variation of both fixations and saccades across sarcastic and non-sarcastic classes, unlike sentiment classes).

Analysis of Features



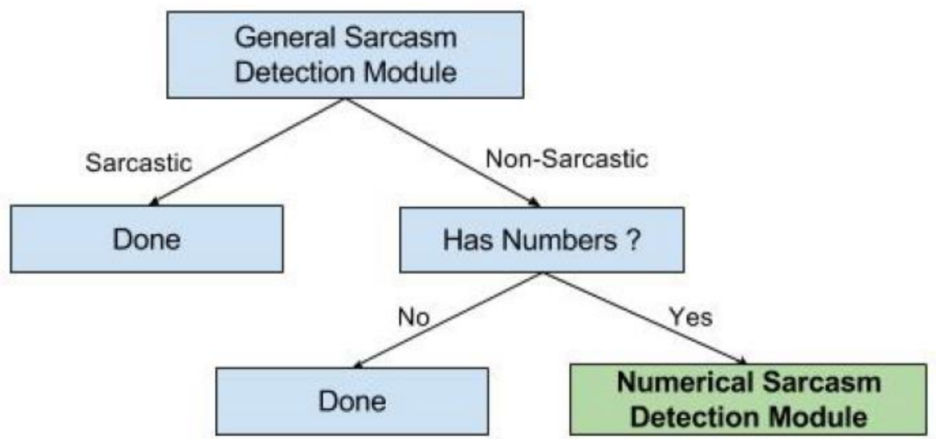
- Visualization of representations learned by two variants of the network. The output of the Merge layer (of dimension 150) are plotted in the form of colour-bars following Li et al. (2016)

Conclusions

- AI→NLP→SA→Sarcasm chain
- General SA does not work well for Sarcasm
- General Sarcasm does not work well for numerical sarcasm
- Rich feature set needed: surface to deeper intent incongruity
- Success from data and annotation
- Success from Deep Learning

Future Work

- Mine the web for more training data of numerical saracasm
- Explain features "discovered" in deep learning
- Perform large scale sentiment and sarcasm detection on social media, tweet, blogs etc.



Resources and Publications

- http://www.cfilt.iitb.ac.in
- http://www.cse.iitb.ac.in/~pb

Most recent and relevant:

Aditya Joshi,Pushpak Bhattacharyya and Mark Carman, *Automatic Sarcasm Detection: A Survey*, ACM Computing Survey (**ACM-CSUR**), Article No. 73, Volume 50 Issue 5, September 2017

THANK YOU