NLP for Social Media

Lecture 1: What, Why and How?

Monojit Choudhury

Microsoft Research Lab, *monojitc@microsoft.com*

What will we learn?

- What is NLP?
- Why NLP for Social Media?
- What are the challenges & opportunities?
- What are the approaches?

What will we learn?

- What is NLP?
- Why NLP for Social Media?
- What are the challenges & opportunities?
- What are the approaches?

What is NLP?

- Natural Language Processing is
  - Making computers understand what we say
  - Making computers understand what we write
  - Making computers write
  - Making computers speak
  - Making computers learn our tongues

- Well, if a 4-year old can do that effortlessly, it should not be that difficult

A weird conversation!

Rahul: *I saw a monkey with a banana.*

Computer: *That's gruesome!*

Rahul: *Why? What's so gruesome about seeing a monkey?*

Computer: *Oh I see! What else did you see with the banana?*

Rahul: *Come on! Do you expect the monkey to share it with others?*

Computer: *I wonder, how did you manage to get its banana?*

In Natural Languages, AMBIGUITY is the rule rather than an exception

What makes NLP hard?

Underlying form ⟷ Surface form

$\forall x.Loves(x, wife(x))$

⇓

Every man loves his wife.

⇓

$\exists y.\forall x.Loves(x, wife(y))$

Natural Languages are inherently ambiguous. There's almost always many-to-many mappings between the surface and underlying forms

What makes NLP hard?

- Text ←→ Pronunciation

Read → /rid/ or /red/ → Red

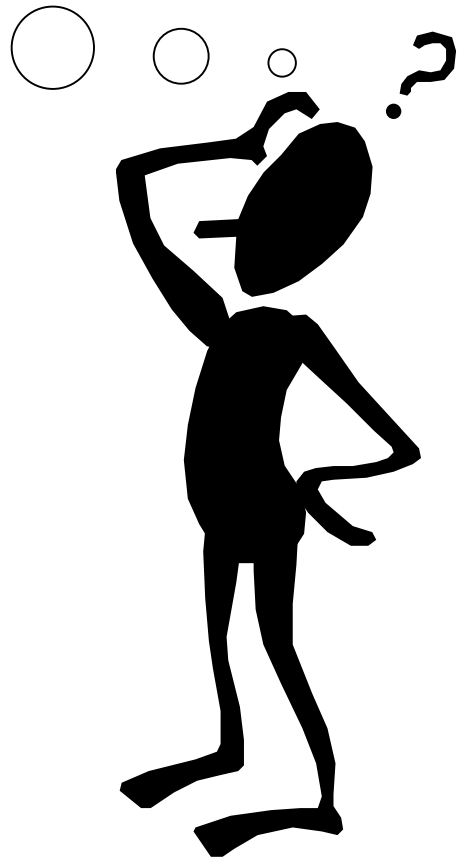
- Word ← → Meaning

Bank, rose, head

- Sentence ← → Intention

It’s hot out here.

I can think of only one… probably 2 meanings! But they say it has many!!



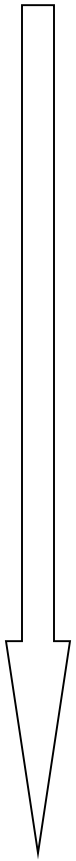
I made her duck.

And NLP is harder because

It is resource intensive

- There are more than 140,000 words in English
- The number of phrases (*take up, carry on*, etc.) is just twice that number
- The number of multiword expressions (*traffic light, Herculean task, bolt from the blue*, etc.) is unknown
- Language understanding requires a shared context: World-knowledge & common-sense.

Analyzing Language: A Reductionist Approach

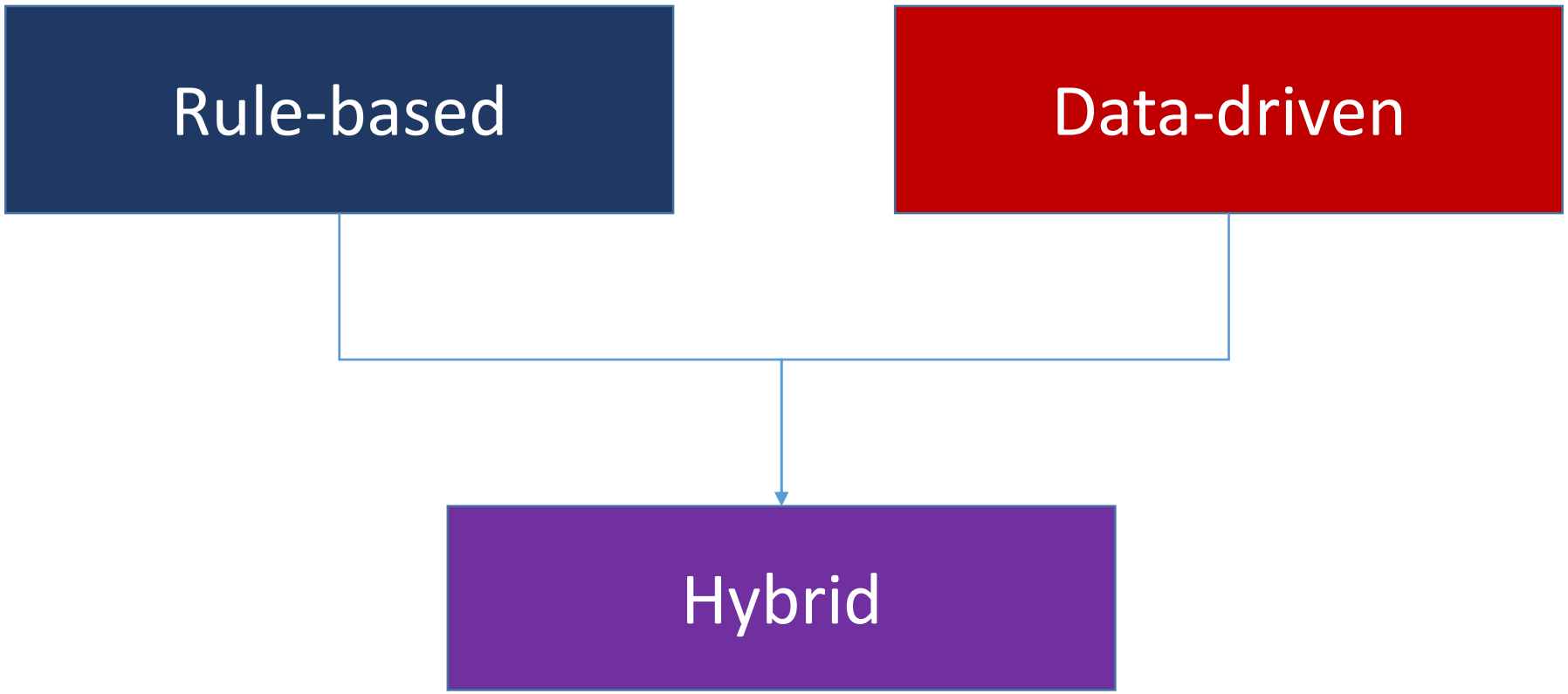


Phonetics	Sound types	Speech processing
Phonology	Sound patterns	G2P, transliteration
Morphology	Words	Morph Analyzer, POS tagger
Syntax	Sentences	Parser, Chunker
Semantics	Meaning	Sense Disambiguation
Discourse	Relation between sentences	Anaphora resolution
Pragmatics	Unsaid Intentions	Sentiment detection

Computational Linguists also study

- Language learning
- Language dynamics: Change & Evolution
- Socio-linguistics: Interaction between aspects of language and society.
- Literary analysis
- Speech Pathology

Approaches



What will we learn?

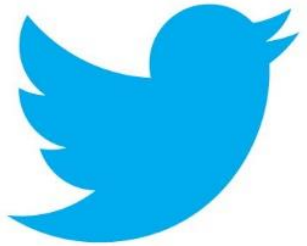
- What is NLP?
- Why NLP for Social Media?
- What are the challenges & opportunities?
- What are the approaches?

Why NLP?

- Aids communication between two humans
  - Machine translation
  - Speech-to-speech translation
  - Speech-to-text & text-to-speech
  - Editorial aids (spelling & grammar checkers)
- Aids communication between human and machine
  - Personal assistants
  - Interactive Voice Response systems
- Aids communication between two machines

Why NLP for Social Media?

- Social Media generates BIG UNSTRUCTURED NATURAL LANGUAGE DATA
    - Volume: 1.3 Billion monthly active FB users
    - Velocity: 5700 Tweets/sec. 2500 FB-msg/sec
    - Variety: scripts, languages, style, topic, …



- Today's world resides in social media
- It is impossible to process (consume, understand or summarize) this information manually.

Instagram

Why NLP for Social Media

- Trending Topic Detection
- Information Retrieval & Extraction
- Information Summarization
- Sentiment Detection
- Rumor Detection
- Adult Content Filtering

It is one of the hottest emerging research sub-area in NLP.

What will we learn?

- What is NLP?
- Why NLP for Social Media?
- What are the challenges & opportunities?
- What are the approaches?

A New Recipe for Language!

Trying our chicken in Penang Curry

Conversation is speech-like

Umair Z Ahmed, Deep Chakravarti, Abhishek Padmanabh and 5 others like this.

View 2 more comments

Monojit Choudhury @Mallar Bangalore-e. tui kothay?
November 18, 2011 at 8:49pm · Like

Sayan Bhattacharya chaliye jao guru.....
November 18, 2011 at 11:35pm · Like

Sandeep Peethamber abae whats going on??? even i want to
November 19, 2011 at 12:03pm · Like

Moushumi Goswami Kothakar Master chef contest ?
November 19, 2011 at 8:18pm · Like

Monojit Choudhury contest naa, class
November 19, 2011 at 8:36pm · Like

Deep Chakravarti well done Monojit. This is impressive
November 19, 2011 at 8:42pm · Like

Write a comment...

A New Recipe for Language!

Trying our chicken in Penang Curry

Non-standard spellings

Umair Z Ahmed, Deep Chakravarti, Abhishek Padmanabh and 5 others like this.

View 2 more comments

Monojit Choudhury @Mallar Bangalore-e. tui kothay?
November 18, 2011 at 8:49pm · Like

Sayan Bhattacharya chaliye jao guru.....
November 18, 2011 at 11:35pm · Like

Sandeep Peethamber abae whats going on??? even i want to 🙂
November 19, 2011 at 12:03pm · Like

Moushumi Goswami Kothakar Master chef contest ?
November 19, 2011 at 8:18pm · Like

Monojit Choudhury contest naa, class 😬
November 19, 2011 at 8:36pm · Like

Deep Chakravarti well done Monojit. This is impressive
November 19, 2011 at 8:42pm · Like

Write a comment...

A New Recipe for Language!

Trying our chicken in Penang Curry

Tags, emoticons

Umair Z Ahmed, Deep Chakravarti, Abhishek Padmanabh and 5 others like this.

View 2 more comments

Monojit Choudhury @Mallar Bangalore-e. tui kothay?
November 18, 2011 at 8:49pm · Like

Sayan Bhattacharya chaliye jao guru.....
November 18, 2011 at 11:35pm · Like

Sandeep Peethamber abae whats going on??? even i want to
November 19, 2011 at 12:03pm · Like

Moushumi Goswami Kothakar Master chef contest ?
November 19, 2011 at 8:18pm · Like

Monojit Choudhury contest naa, class
November 19, 2011 at 8:36pm · Like

Deep Chakravarti well done Monojit. This is impressive
November 19, 2011 at 8:42pm · Like

Write a comment...

A New Recipe for Language!

Trying our chicken in Penang Curry

Code-mixing

Umair Z Ahmed, Deep Chakravarti, Abhishek Padmanabh and 5 others like this.

View 2 more comments

Monojit Choudhury @Mallar Bangalore-e. tui kothay?
November 18, 2011 at 8:49pm · Like

Sayan Bhattacharya chaliye jao guru.....
November 18, 2011 at 11:35pm · Like

Sandeep Peethamber abae whats going on??? even i want to 🙂
November 19, 2011 at 12:03pm · Like

Moushumi Goswami Kothakar Master chef contest ?
November 19, 2011 at 8:18pm · Like

Monojit Choudhury contest naa, class 😬
November 19, 2011 at 8:36pm · Like

Deep Chakravarti well done Monojit. This is impressive
November 19, 2011 at 8:42pm · Like

Write a comment...

A New Recipe for Language!

Trying our chicken in Penang Curry

Transliteration

Umair Z Ahmed, Deep Chakravarti, Abhishek Padmanabh and 5 others like this.

View 2 more comments

Monojit Choudhury @Mallar Bangalore-e. tui kothay?
November 18, 2011 at 8:49pm · Like

Sayan Bhattacharya chaliye jao guru.....
November 18, 2011 at 11:35pm · Like

Sandeep Peethamber abae whats going on??? even i want to
November 19, 2011 at 12:03pm · Like

Moushumi Goswami Kothakar Master chef contest ?
November 19, 2011 at 8:18pm · Like

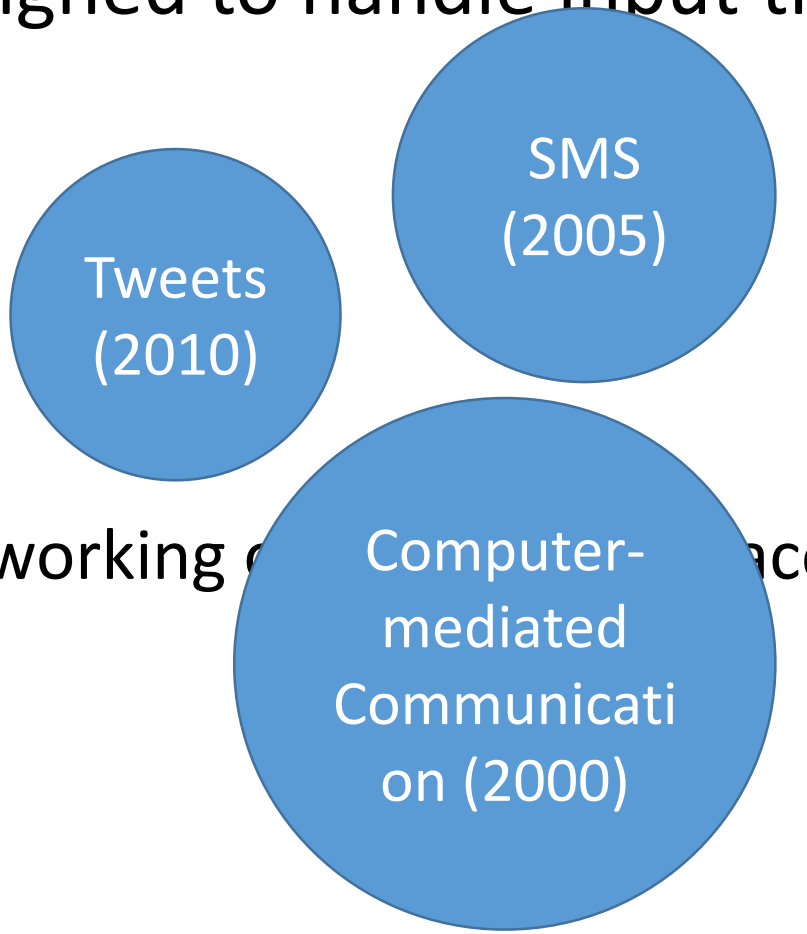
Monojit Choudhury contest naa, class
November 19, 2011 at 8:36pm · Like

Deep Chakravarti well done Monojit. This is impressive
November 19, 2011 at 8:42pm · Like

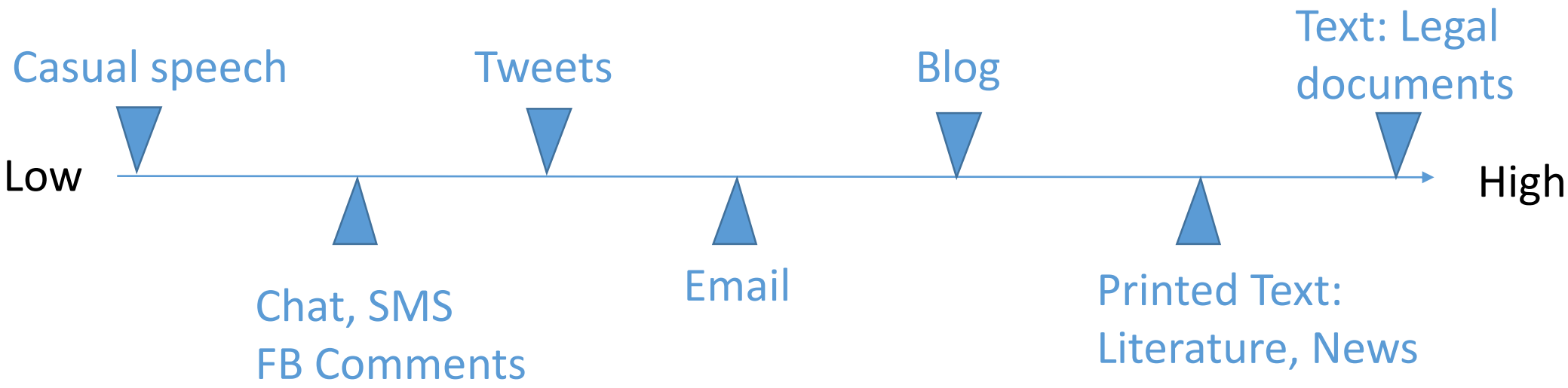
Write a comment...

Changing Landscapes

- Traditionally, NLP systems are designed to handle input that is
  - Grammatically correct
  - No spelling errors
  - Single language
  - The right script
  - Text-like and Formal (unless one is working o aces)



The Formality Continuum



But there are also opportunities

Data, DATA & more DAAAATTTTTTAAAAAAAAA

- Speech data is expensive; social media data is a good proxy
- Personal conversations
- Socially grounded data
- Language dynamics, e.g.,
  - Evolution of new hashtags, words
  - Spelling changes

Language usage &
- Topic
- Demography
- Social relationships
- Personal relationships

What will we learn?

- What is NLP?
- Why NLP for Social Media?
- What are the challenges & opportunities?
- What are the approaches?

How well does existing NLP tools work on Social Media Data?

System	Accuracy on Std. Language	Accuracy on Social media
Machine Translation En-Es (BLEU)	~35 [Moses, WMT12]	29 [Hassan & Menzes, 2014]
Parts-of-speech Tagging (word labeling accuracy)	98% [Stanford Tagger]	85% [Gimpel et al. 2011]
Sentiment detection (tweet labeling accuracy)	92% [Pang & Lee, 2004]	70-80% [Barbosa and Feng, 2010]

Approach

- Normalization

- Systems/techniques specifically built for SMD.

How well does existing NLP tools work on Social Media Data?

System	Accuracy on Std. Language	Accuracy on Social media
Machine Translation En-Es (BLEU)	~35 [Moses, WMT12]	29 [Hassan & Menzes, 2014]
Parts-of-speech Tagging (word labeling accuracy)	98% [Stanford Tagger]	85% [Gimpel et al.
Sentiment detection (tweet labeling accuracy)	92% [Pa 2004]	rbosa and Feng, 2010]

BLEU = 32 with normalization

How well does existing NLP tools work on Social Media Data?

System	Accuracy on Std. Language	Accuracy on Social media
Machine Translation En-Es (BLEU)	~35 [Moses, WMT12]	29 [Hassan & Menzes, 2014]
Parts-of-speech Tagging (word labeling accuracy)	98% [Stanford Tagger]	85% [Gimpel et al. 2011]
Sentiment detection (tweet labeling accuracy)		70-80% [Barbosa and Feng, 2010]

89% for Twitter-specific POS tagger

Developing SMD-specific NLP systems

- SMD specific data creation
- SMD specific features
- Completely new techniques/models

- Systems/techniques specifically built for SMD.

Summary

- NLP is all about
  - building systems that deal with human language input and/or output
  - computational study of human languages
- [ARC] Ambiguity, resource intensity and need for deep context understanding makes NLP one of the hardest engg. Goals
- Language can be broken down into sub-systems of sounds, words, syntax, meaning and the interactions within and between these layers.
- Most modern NLP systems are data-driven or hybrid, though rule-based systems might be useful in some cases

Summary (contd.)

- NLP for social media has several applications, but is hard because of volume, velocity, variety, and departure from standard language
- Language of social media resembles informal speech conversation, even though it is primarily expressed through text.
- Social Media also provides opportunities for NLPers and linguists in the form of large volumes of socially grounded data.
- NLP tools designed for standard text of a language do not work well on social media data.
- NLP for SMD either relies on converting the informal text to standard text (normalization) or building systems that are specifically designed to tackle SMD.

Suggested Readings

For Language on Social Media:

- Michele Zappavigna, *Discourse of Twitter and Social Media: How We Use Language to Create Affiliation on the Web,* 2012
  - Ch 1 (Introduction), Ch2 (Social Media as a corpora), Ch 3 (Language of Microblogging)
- http://www.englishtown.com/blog/has-social-media-changed-the-way-we-speak-and-write-english/

For Intro to NLP:

- https://en.wikipedia.org/wiki/Natural_language_processing
- http://nlpers.blogspot.in/ (Hal Daume's Blog)
- http://languagelog.ldc.upenn.edu/nll/ (Collaborative Blog maintained by Mark Liberman)

References

- Hassan, Hany, and Arul Menezes. "Social Text Normalization using Contextual Graph Random Walks." *ACL* 2013.
- Gimpel, Kevin, et al. "Part-of-speech tagging for twitter: Annotation, features, and experiments." *ACL* 2011.
- Pang, Bo, and Lillian Lee. "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts." *ACL* 2004.
- Barbosa, Luciano, and Junlan Feng. "Robust sentiment detection on twitter from biased and noisy data.“ *COLING* 2010.
- Stanford Tagger: http://nlp.stanford.edu/software/tagger.shtml
- WMT12: http://www.statmt.org/wmt12/
- Moses: http://www.statmt.org/moses/