"Silence is the language of God,
all else is poor translation."

-Rumi, Sufi Poet & Philosopher (1207-1273)

Isn't NLP all about rules?! | Aditya Joshi | 24.06.2017

Isn't NLP* all about a bunch of rules?!

Aditya Joshi
Cuddle.ai @ Fractal Analytics
Aditya.Joshi@fractalanalytics.com
24.06.2017

* Natural Language Processing

Learning in Humans: The story of my nephew

Learning in Humans: The story of my nephew

Source of 10-rupee note image: Wikimedia commons

Learning in Humans: The story of my nephew

When children learn languages,
is it rule-based or statistical?

Source of 10-rupee note image: Wikimedia commons

Natural Language Processing (NLP)

- Branch of Artificial Intelligence

Natural Language Processing (NLP)

- Branch of Artificial Intelligence
- Automatic approaches to solve problems pertaining to text

Natural Language Processing (NLP)

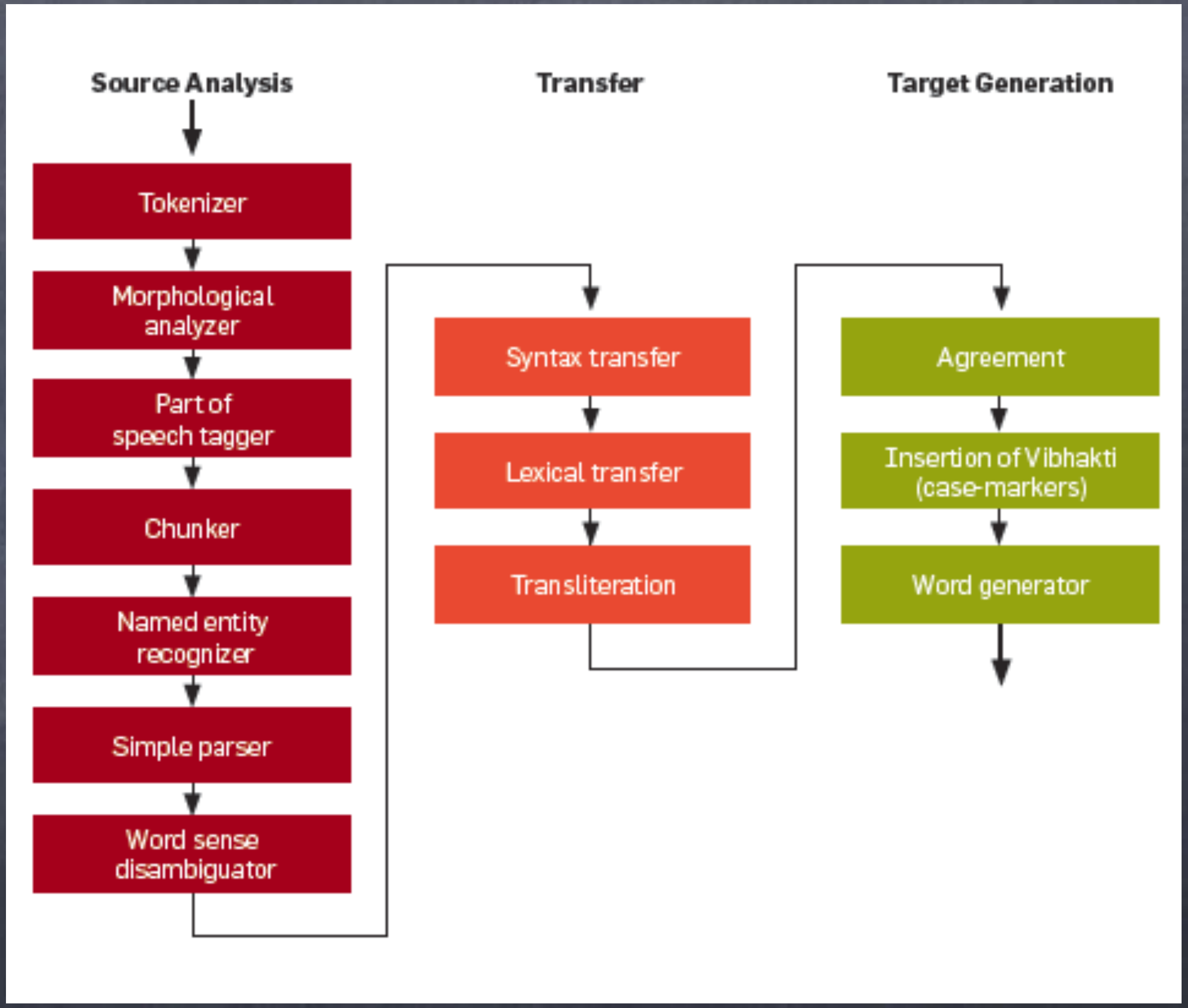
- Branch of Artificial Intelligence
- Automatic approaches to solve problems pertaining to text
- Evolving motivation, evolving use-cases
- Overlapping terms: Computational Linguistics, Text Analytics

Traditional view of NLP

- Layered approach
  - Words - Phrases - (Subject-Verb-Object) - Meaning - Implicatures
- Expert systems

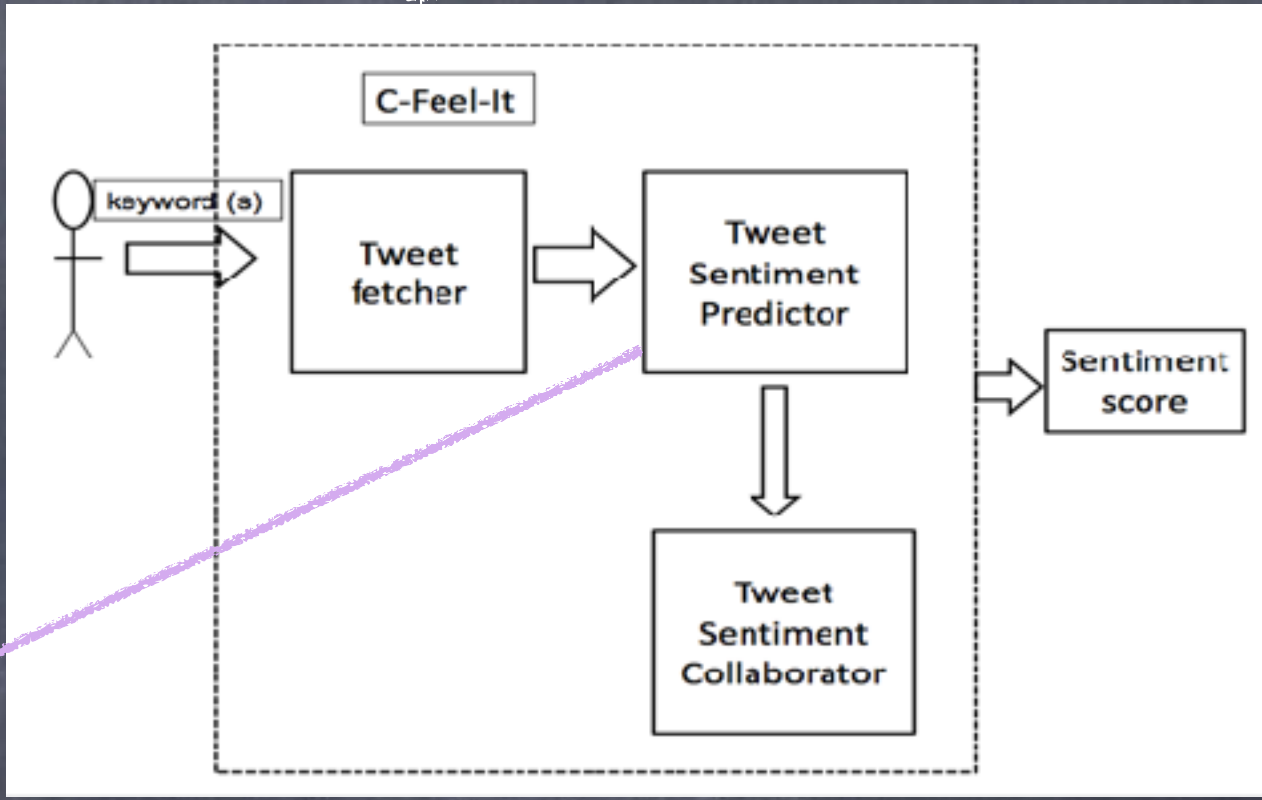
NLP via rules: Traditional automatic translation

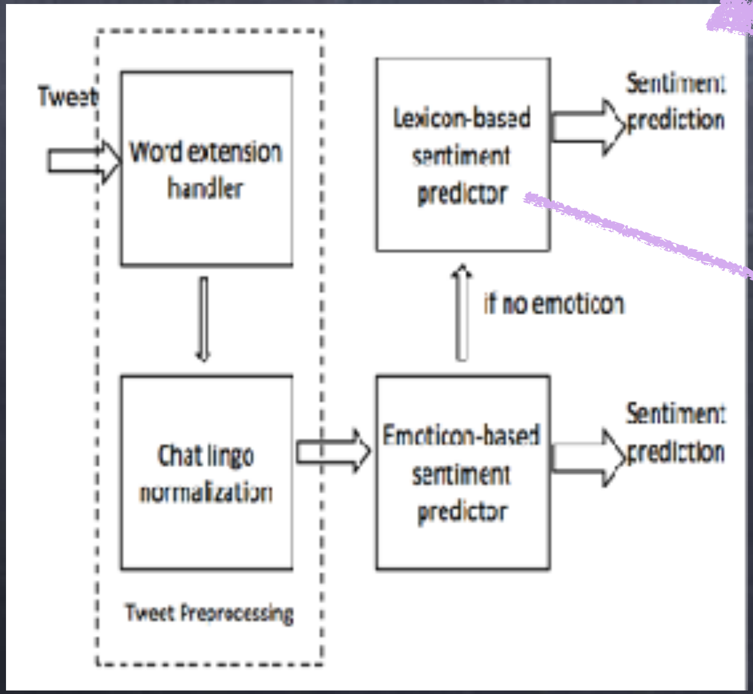
- Layers across source and target side
- Basis of SysTran (1979)

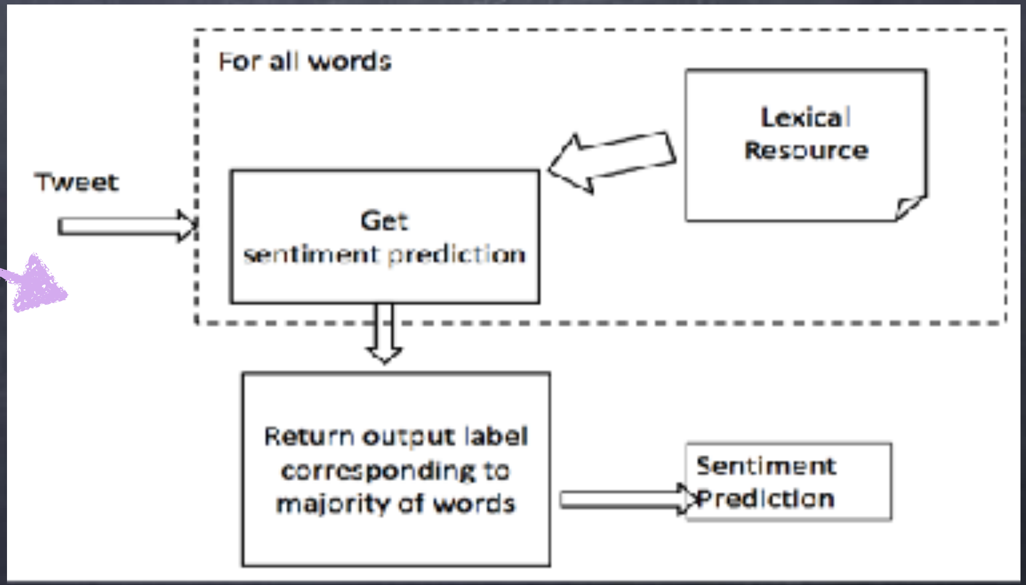


NLP via rules: Traditional sentiment analysis

- Lookup from a sentiment dictionary
- Rules for syntactic structures
- Joshi et al (2011)







Limitations of Rule-based NLP

- Rules are specific and do well in the ‘positive’ cases
- They may not always generalise
- A well-known problem in NLP

Limitation of Rule-based systems: High Precision-Low Recall

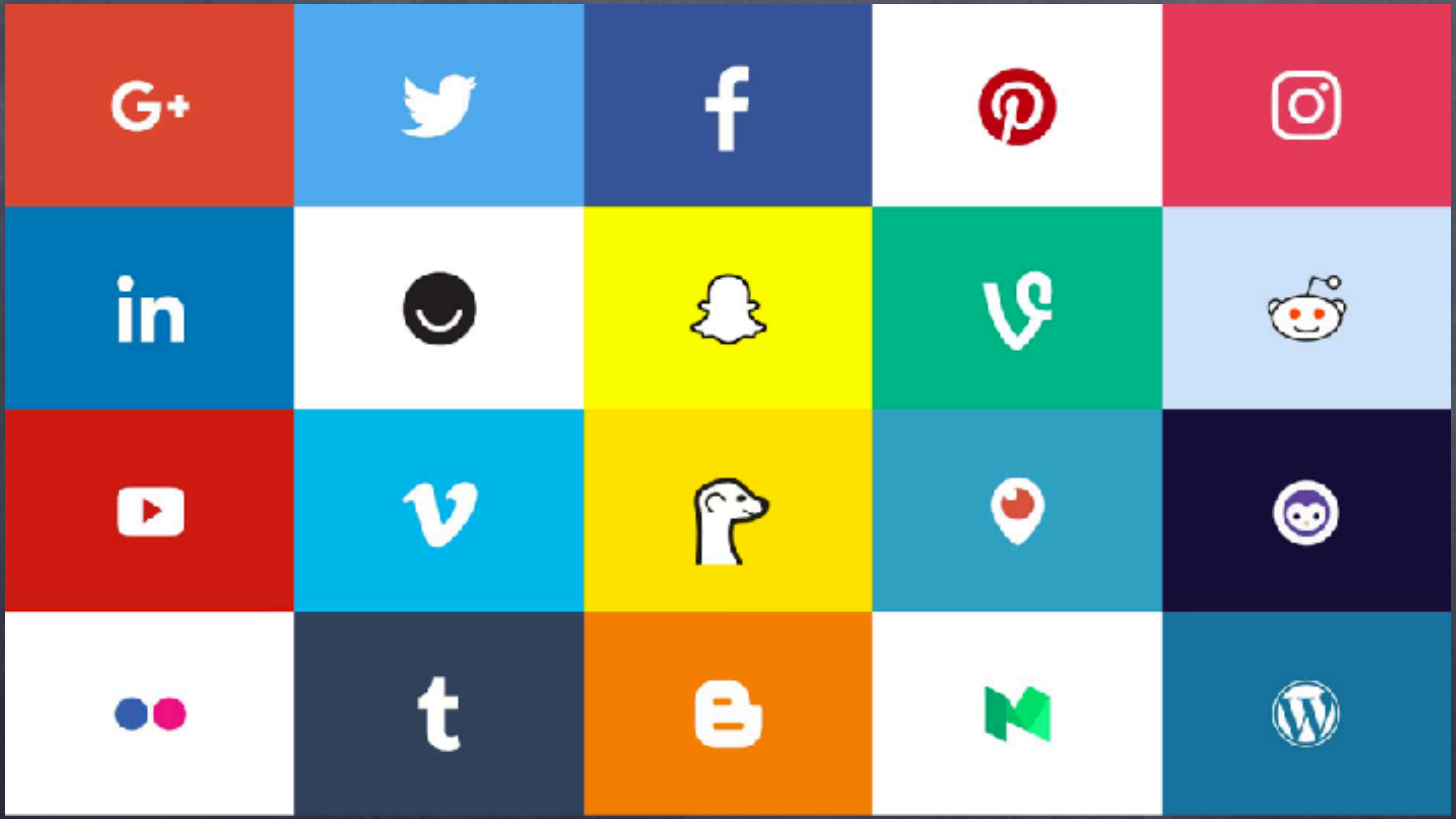
- Rules are specific and do well in the 'positive' cases
- They may not always generalise
- A well-known problem in NLP

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**

Rule-based

Statistical

Enter: Web-scale Textual Data



Source: dustntv

Statistical NLP

- Use of statistical/machine learning techniques for NLP
- The key idea is to model NLP problems as machine learning tasks:
  - Part-of-speech tagging: Sequence labelling
  - Sentiment analysis: Classification, etc.

Manning, Christopher D., and Hinrich Schütze.
Foundations of statistical natural language processing.
Cambridge: MIT press, 1999.

- In this talk, I share my experience with three statistical NLP projects

Outline

Research area of NLP	Specific Task	Statistical Algorithm
Machine Translation		
Opinion Mining		
Sarcasm detection		

Outline

Research area of NLP	Specific Task	Statistical Algorithm
Machine Translation	Discovering aligned words	Expectation-Maximization (EM) Algorithm
Opinion Mining		
Sarcasm detection		

Outline

Research area of NLP	Specific Task	Statistical Algorithm
Machine Translation	Discovering aligned words	Expectation-Maximization (EM) Algorithm
Opinion Mining	Discovering political issues and orientations	Latent Dirichlet Allocation (LDA) Model
Sarcasm detection		

Outline

Research area of NLP	Specific Task	Statistical Algorithm
Machine Translation	Discovering aligned words	Expectation-Maximization (EM) Algorithm
Opinion Mining	Discovering political issues and orientations	Latent Dirichlet Allocation (LDA) Model
Sarcasm detection	Predicting sarcasm in dialogue	Hidden Markov Model (HMM)

Outline

Research area of NLP	Specific Task	Statistical Algorithm
Machine Translation	Discovering aligned words	Expectation-Maximization (EM) Algorithm
Opinion Mining	Discovering political issues and orientations	Latent Dirichlet Allocation (LDA) Model
Sarcasm detection	Predicting sarcasm in dialogue	Hidden Markov Model (HMM)

Part 1 of 3:

Machine Translation

Joint work with Pushpak Bhattacharyya, IIT-B

Algorithm, code and results appear in the textbook 'Machine Translation', CRC Press

Source: https://d1yi8yosmemxim.cloudfront.net/field/image/saying_hello_in_different_languages.jpg

Machine Translation (MT)

- Machine Translation deals with translating a text from a source language to a target language

  Input: I eat rice (EN)
  Output: Io mangio il riso (IT)

- Applications and use cases
- Training data: Parallel sentences

  I eat pasta : Io mangio la pasta
  You eat rice: Tu mangi il riso

- The notion of alignment

Image is a movie still taken from the web

Introducing: Alignment

- Can you guess the translation?

Introducing: Alignment

- Can you guess the translation?

the boy dances : pomesu pimo
the man dances: pomesu grollo
the man jumps: junesu grollo

the boy jumps : ?

Introducing: Alignment

- Can you guess the translation?

the boy dances : pomesu pimo
the man dances: pomesu grollo
the man jumps: junesu grollo

the boy jumps : ?

Source	Target
the man	grollo
the boy	pimo
dances	pomesu
jumps	junesu

Introducing: Alignment

- Can you guess the translation?

the boy dances : pomesu pimo
the man dances: pomesu grollo
the man jumps: junesu grollo

the boy jumps : ?

Source	Target
the man	grollo
the boy	pimo
dances	pomesu
jumps	junesu

- Two components: (a) learning the mapping, (b) learning the order
- Alignment is the mapping between words and phrases across the languages
- Complex as the size of parallel text grows

Expectation-Maximization (EM) Algorithm

- Iterative method to find maximum likelihood estimates of parameters
- Iterates between E-step and M-step

Expectation-Maximization (EM) Algorithm

- Iterative method to find maximum likelihood estimates of parameters
- Iterates between E-step and M-step

Initialize parameters

Expectation-Maximization (EM) Algorithm

- Iterative method to find maximum likelihood estimates of parameters
- Iterates between E-step and M-step

Initialize parameters

E-Step

M-Step

Expectation-Maximization (EM) Algorithm

- Iterative method to find maximum likelihood estimates of parameters
- Iterates between E-step and M-step

Initialize parameters

E-Step

Expected log-likelihood given the current estimates of the parameters

M-Step

Update the parameters as the value to maximise the log-likelihood

EM for Alignment discovery in MT

- Source sentence: e_1, e_2... e_m
- Target sentence: f_1, f_2 .... f_n
- Goal: Learn Pr(f_j|e_i) for every pair of source and target language words
- Initialization: Uniform probabilities (or better?)

E-Step

Expected log-likelihood given the current estimates of the parameters

M-Step

Update the parameters as the value to maximise the log-likelihood

EM for Alignment discovery in MT

- Source sentence: e_1, e_2… e_m
- Target sentence: f_1, f_2 …. f_n
- Goal: Learn Pr(f_j|e_i) for every pair of source and target language words
- Initialization: Uniform probabilities (or better?)

E-Step

Expected log-likelihood given the current estimates of the parameters

$$C(e_i \leftrightarrow f_j; e \leftrightarrow f) = \frac{\Pr(f_j \mid e_i)}{\sum_x \Pr(X \mid e_i)}$$

M-Step

Update the parameters as the value to maximise the log-likelihood

EM for Alignment discovery in MT

- Source sentence: e_1, e_2... e_m
- Target sentence: f_1, f_2 .... f_n
- Goal: Learn Pr(f_j|e_i) for every pair of source and target language words
- Initialization: Uniform probabilities (or better?)

E-Step

Expected log-likelihood given the current estimates of the parameters

$$C(e_i \leftrightarrow f_j; e \leftrightarrow f) = \frac{\Pr(f_j \mid e_i)}{\sum_X \Pr(X \mid e_i)}$$

M-Step

Update the parameters as the value to maximise the log-likelihood

$$\Pr(f_j \mid e_i) = \frac{\sum_s C(e_i \leftrightarrow f_j; e^s \leftrightarrow f^s)}{\sum_s \sum_X C(e_i \leftrightarrow X; e^s \leftrightarrow f^s)}$$

EM for Alignment discovery in MT: Example 1

- three rabbits: trois lapins
- rabbits of grenoble: lapins de Grenoble

EM for Alignment discovery in MT: Example 1

- three rabbits: trois lapins
- rabbits of grenoble: lapins de Grenoble

E-Step

Expected log-likelihood given the current estimates of the parameters

M-Step

Update the parameters as the value to maximise the log-likelihood

EM for Alignment discovery in MT: Example 1

- three rabbits: trois lapins
- rabbits of grenoble: lapins de Grenoble

<u>E-Step</u>

Expected log-likelihood given the current estimates of the parameters

<u>M-Step</u>

Update the parameters as the value to maximise the log-likelihood

I	trois	lapins	de	Grenoble
1	0.25	0.25	0.25	0.25
2	0.176471	0.588235	0.117647	0.117647
3	0.14391	0.697713	0.079189	0.079189
4	0.111096	0.788899	0.050002	0.050002
5	0.080374	0.85985	0.029888	0.029888
6	0.054612	0.911439	0.016975	0.016975
7	0.035049	0.946536	0.009207	0.009207
8	0.021411	0.968983	0.004803	0.004803
9	0.01256	0.982584	0.002428	0.002428
10	0.007137	0.990464	0.001199	0.001199
11	0.003962	0.994874	0.000582	0.000582
12	0.002163	0.997279	0.000279	0.000279
13	0.001167	0.998567	0.000133	0.000133

EM for Alignment discovery in MT: Example 2

- three white rabbits : trois lapins blancs
- three white swans : trois cygnes blancs

EM for Alignment discovery in MT: Example 2

- three white rabbits : trois lapins blancs
- three white swans : trois cygnes blancs

E-Step

Expected log-likelihood given the current estimates of the parameters

M-Step

Update the parameters as the value to maximise the log-likelihood

EM for Alignment discovery in MT: Example 2

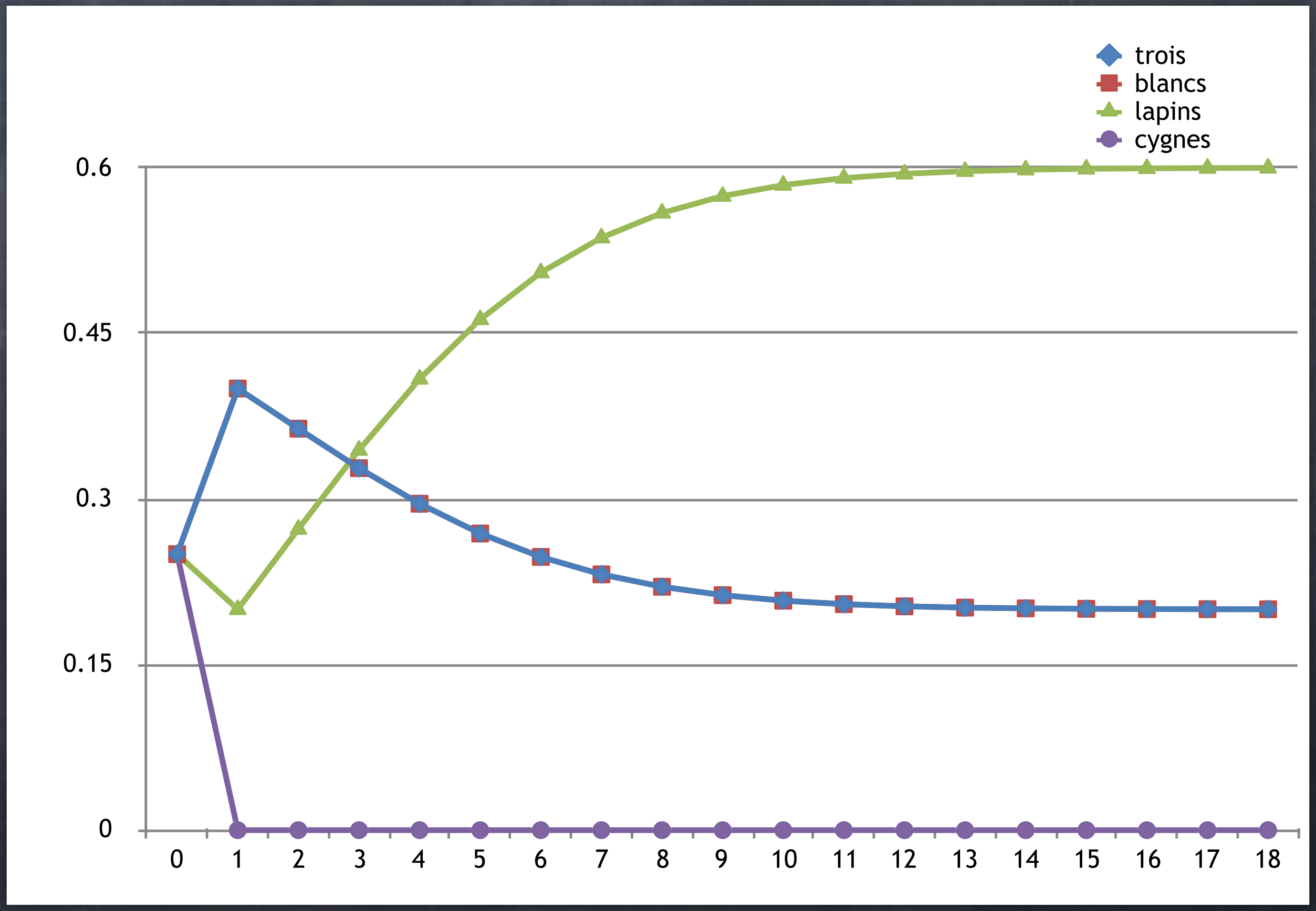
- three white rabbits : trois lapins blancs
- three white swans : trois cygnes blancs

E-Step

Expected log-likelihood given the current estimates of the parameters

M-Step

Update the parameters as the value to maximise the log-likelihood



EM for Alignment discovery in MT: Limitations

- One word aligns to one word
- Not always true
  - He went to Paris -> È andato a Parigi
  - Ram went to Paris -> Ram andò a Parigi
- Advanced concept: Notion of 'fertility'

Example Credit: Soudip Roy Chowdhury, Cuddle.ai

Outline

Research area of NLP	Specific Task	Statistical Algorithm
Machine Translation	Discovering aligned words	Expectation-Maximization (EM) Algorithm
Opinion Mining	Discovering political issues and orientations	Latent Dirichlet Allocation (LDA) Model
Sarcasm detection	Predicting sarcasm in dialogue	Hidden Markov Model (HMM)

Part 2 of 3:

Opinion Mining

Joint work with Pushpak Bhattacharyya and Mark Carman

Aditya Joshi, Pushpak Bhattacharyya, Mark Carman, 'Political Issue Extraction Model: A Novel Hierarchical Topic Model That Uses Tweets By Political And Non-Political Authors', WASSA at NAACL 2016, San Diego, USA, June 2016.

Source: Wikimedia Commons

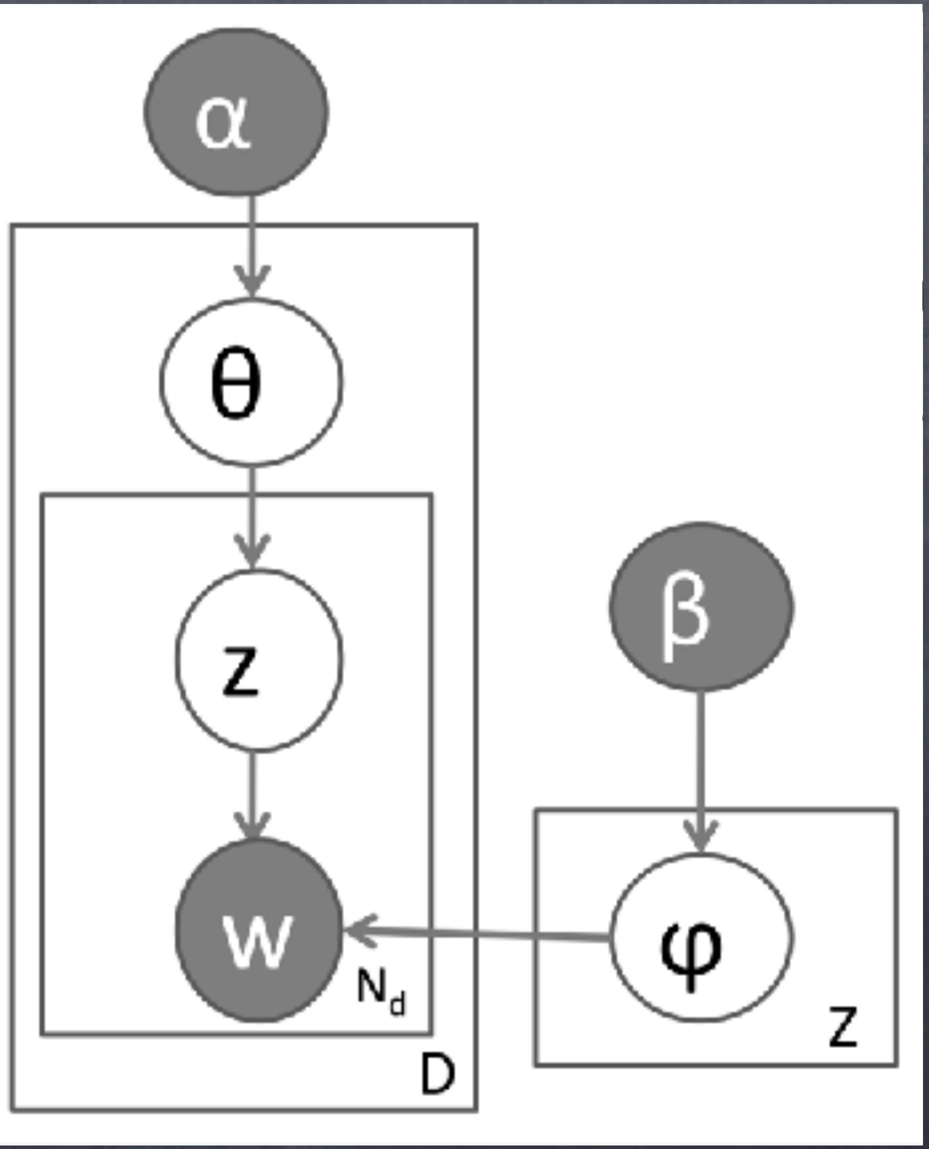
Opinion Mining

- Predicting opinion polarity in text
- Fine-grained opinion mining: Along aspects and issues
- Example: Political issue, opinion and orientation discovery

Source: Wikimedia Commons

Latent Dirichlet Allocation (LDA) Model

- Helps to discover themes underlying an unsupervised corpus
- Generate word clusters based on co-occurrence of words
- Several extensions for specific tasks have been reported

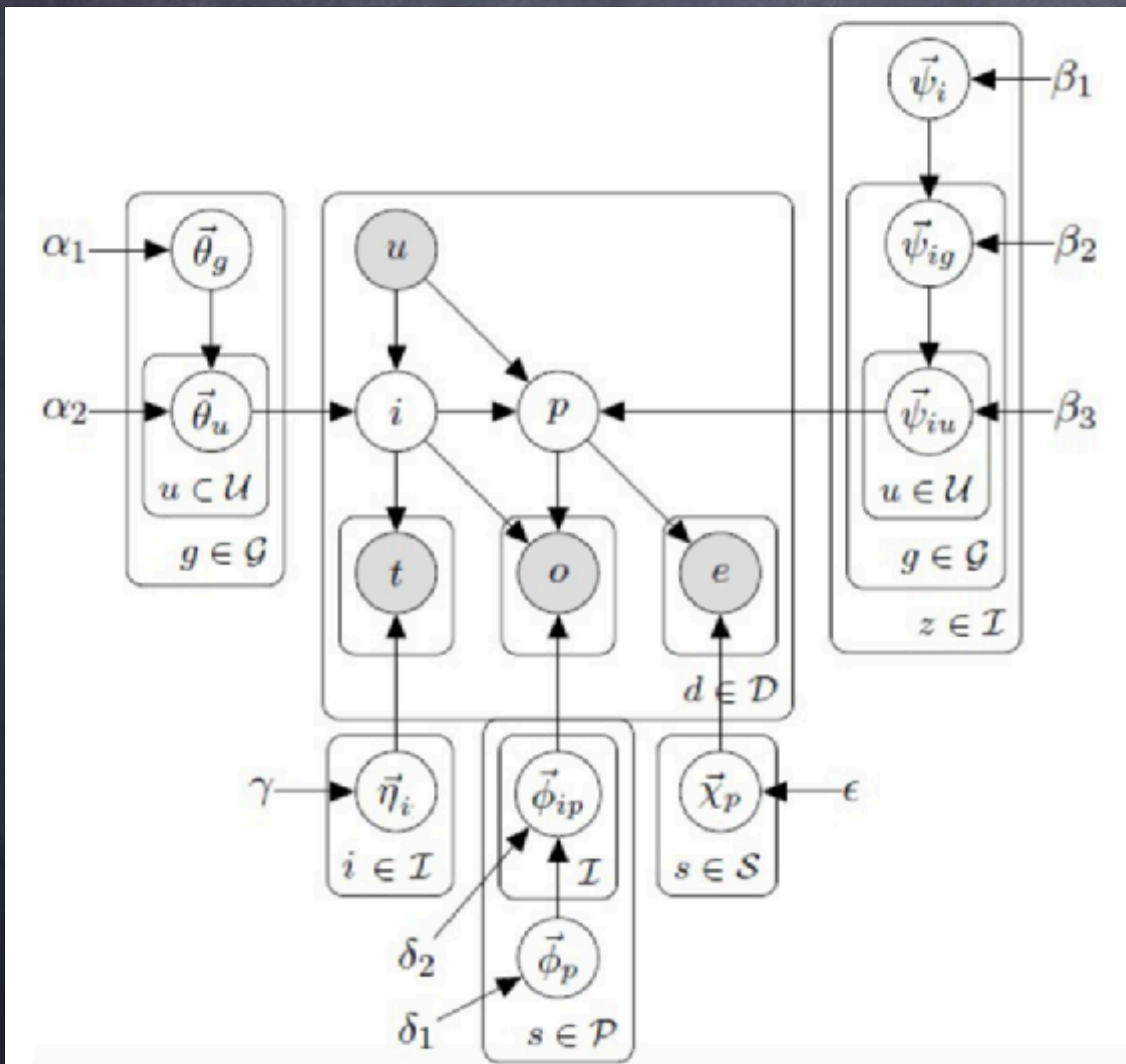


LDA Model

Political Issue & Orientation Mining

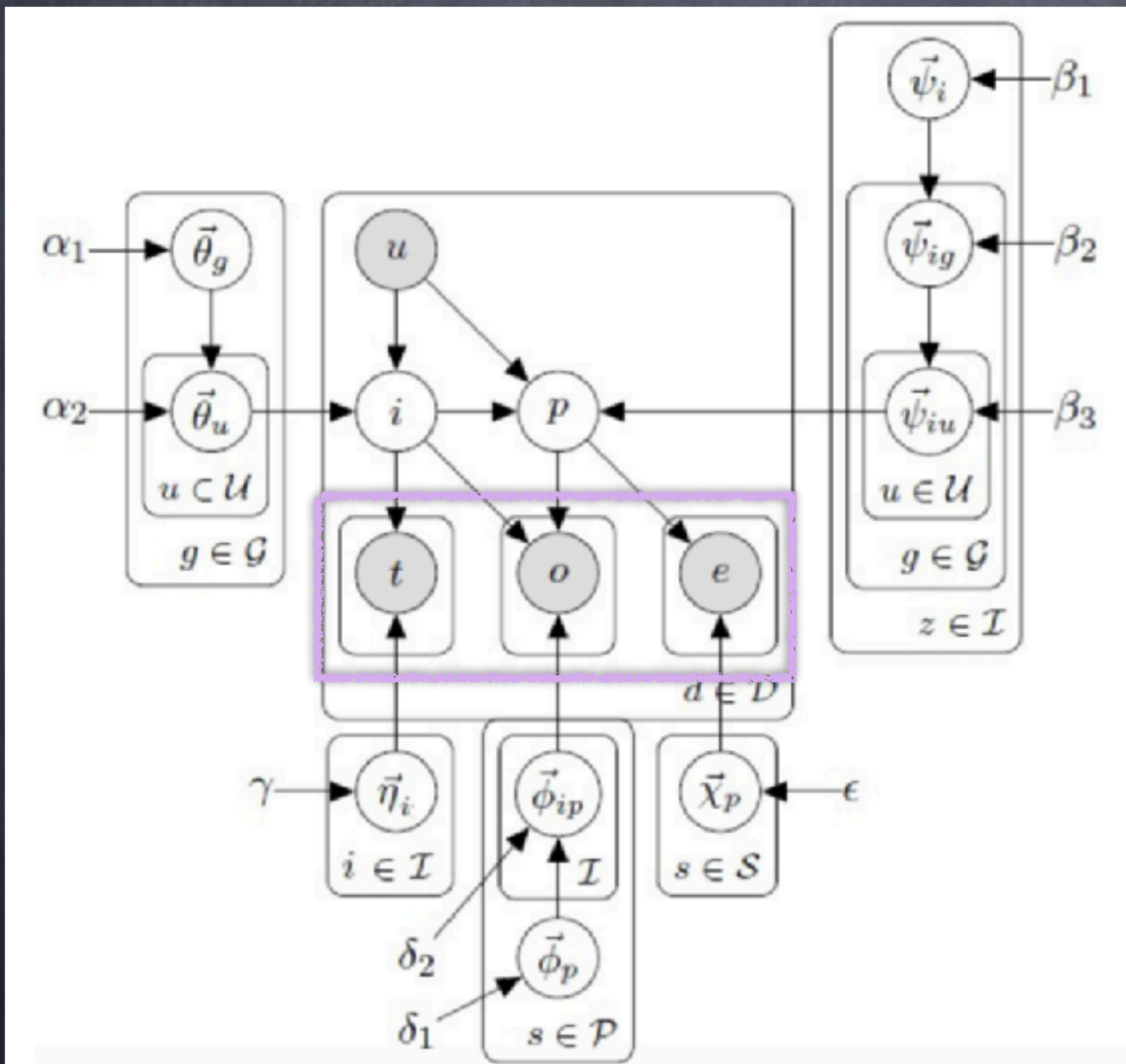
- Goal: An LDA-based model to discover political issues and positions from an unlabelled dataset of tweets
- Premise: Use twitter timelines of both political and nonpolitical authors
- Dataset: 2.4 million tweets from American users
- Input: (a) Twitter timelines of authors, and (b) political affiliation for a subset of authors (namely, Democrats, Republicans, Unknown)
- Output: (a) Clusters of words corresponding to political issues and positions, and (b) Group-wise distributions for political issues and positions

LDA-based Hierarchical Topic Model



Random Variables	
u, g	Author of a tweet and Group of the author
i, p	Issue and position of a tweet
t, o	Issue/Position-word in a tweet
e	Emoticon in a tweet
Distributions	
$\vec{\theta}_{u/g}$	Dist. over issues for author u / group g
$\vec{\psi}_{i,u/g}$	Dist. over positions for issue i and author/group
$\vec{\eta}_i$	Dist. over topic-words for issue i
$\vec{\phi}_{i,p}$	Dist. over opinion-words for issue-position pair
$\vec{\chi}_p$	Dist. over emoticons for position p
Hyper-parameters	
α, β	Concentration par. issue/position dist.
γ	Concentration par. for issue-word dist.
δ	Concentration par. for position-word dist.
ϵ	Concentration par. for emoticon dist.
Counts	
$N_{i,i}^{(t)}$	Frequency of topic-word t in tweets for topic i
$N_{i,u}^{(i)}$	Frequency of tweets on topic i by author u
$V^{(t)}$	Vocabulary size for topic words

LDA-based Hierarchical Topic Model



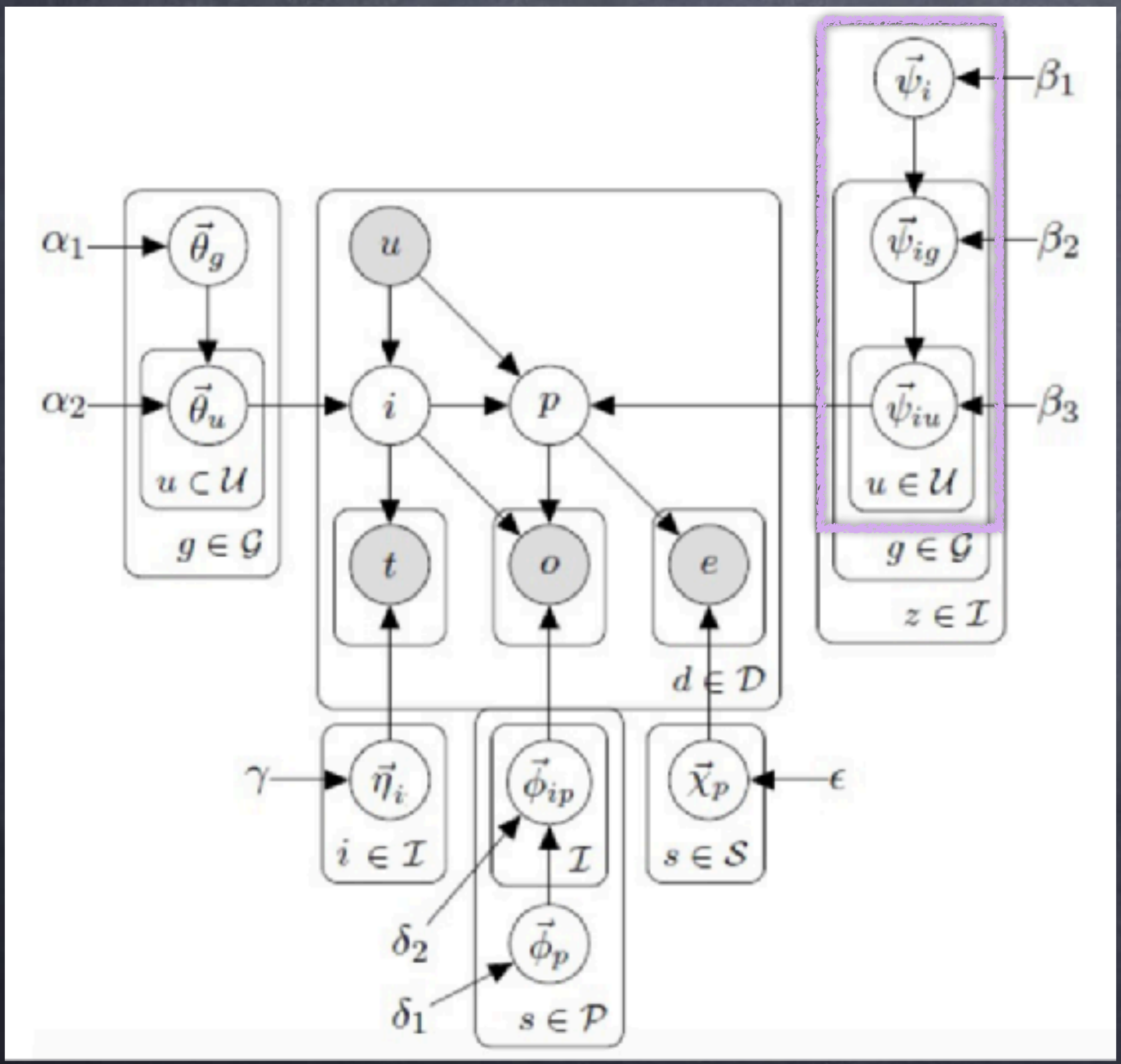
This country needs to seal its borders :)

Issue words: 'country', 'borders'

Position words: 'seal'

Emoticon: :)

LDA-based Hierarchical Topic Model



This country needs to seal its borders :)

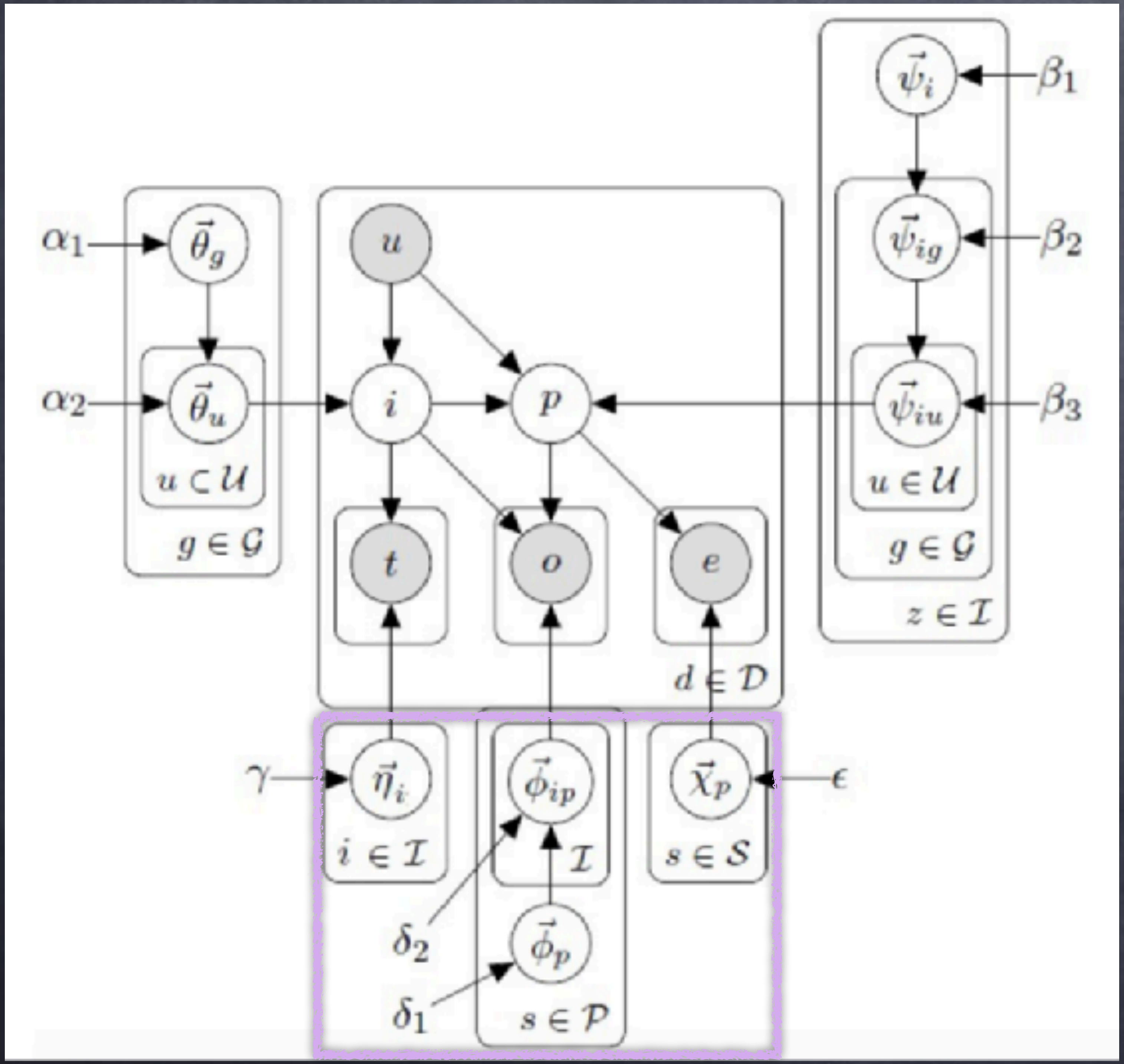
Issue words: 'country', 'borders'

Position words: 'seal'

Emoticon: :)

User u takes position 'pro' towards issue 'abortion' in 90% of their tweets

LDA-based Hierarchical Topic Model



This country needs to seal its borders :)

Issue words: 'country', 'borders'

Position words: 'seal'

Emoticon: :)

User u takes position 'pro' towards issue 'abortion' in 90% of their tweets

'abortion' is most likely described using 'baby', 'life'; 'pro-abortion' using 'freedom', 'health'

Estimation

- Block-based Gibbs sampling
- I = 35, P = 2
- 32 Republicans, and 46 Democrats

What are the political issues and positions?

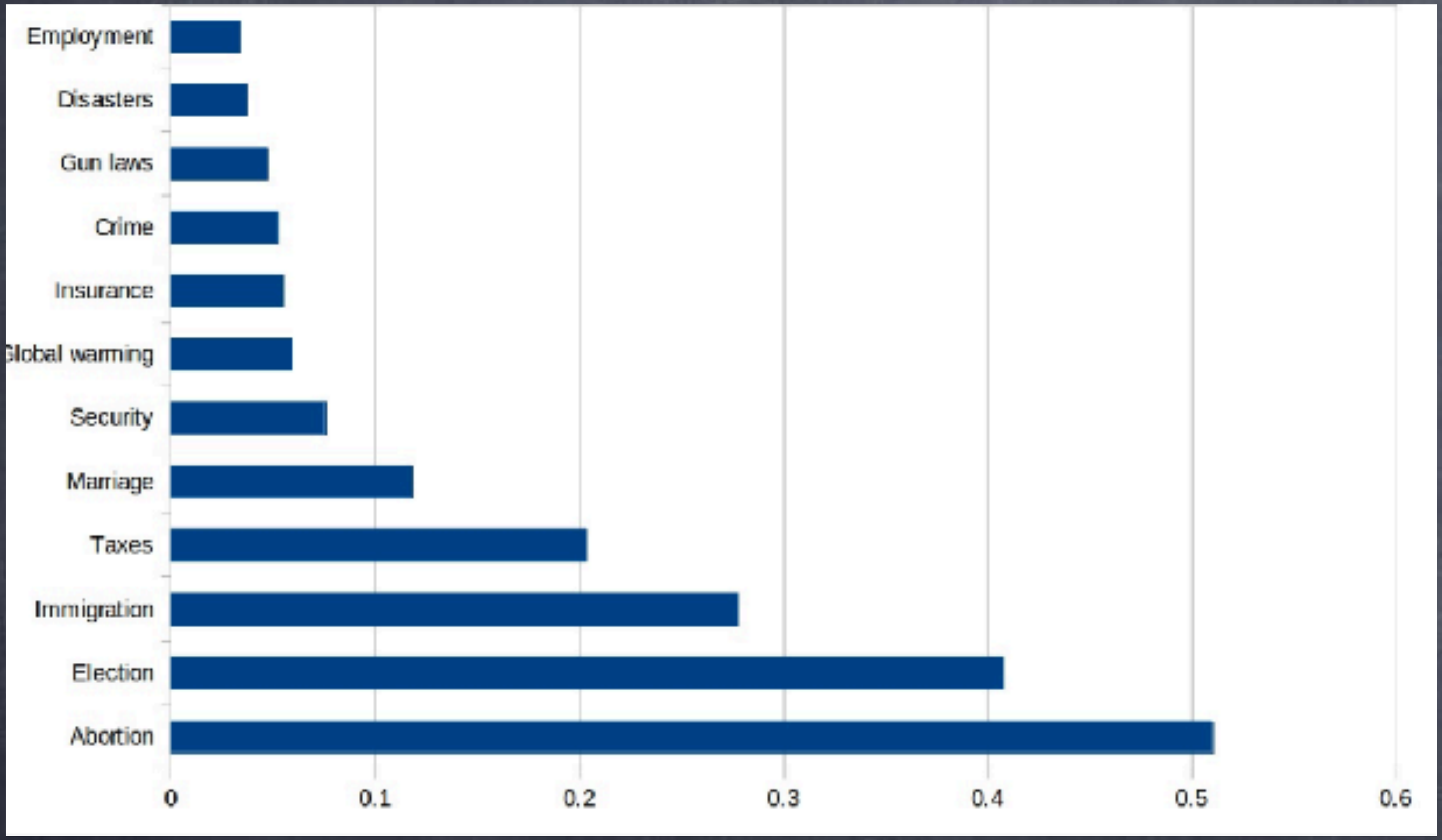
Insurance	Gun laws	Crime
health	people	scene
insurance	gun	police
people	laws	man
care	guns	fire
plan	control	suspect
Abortion	**Security/War**	**Employment**
abortion	attack	workers
baby	video	job
babies	security	wages
freedom	police	jobs
women	forces	people
Immigration	**Economy**	**Climate**
workers	tax	climate
immigration	jobs	people
stories	debt	change
patriot	taxes	warming
politics	spending	years
Marriage	**Election**	**Disasters**
people	bill	acres
freedom	vote	fire
marriage	campaign	weather
rights	state	snow
women	election	storm

What are the political issues and positions?

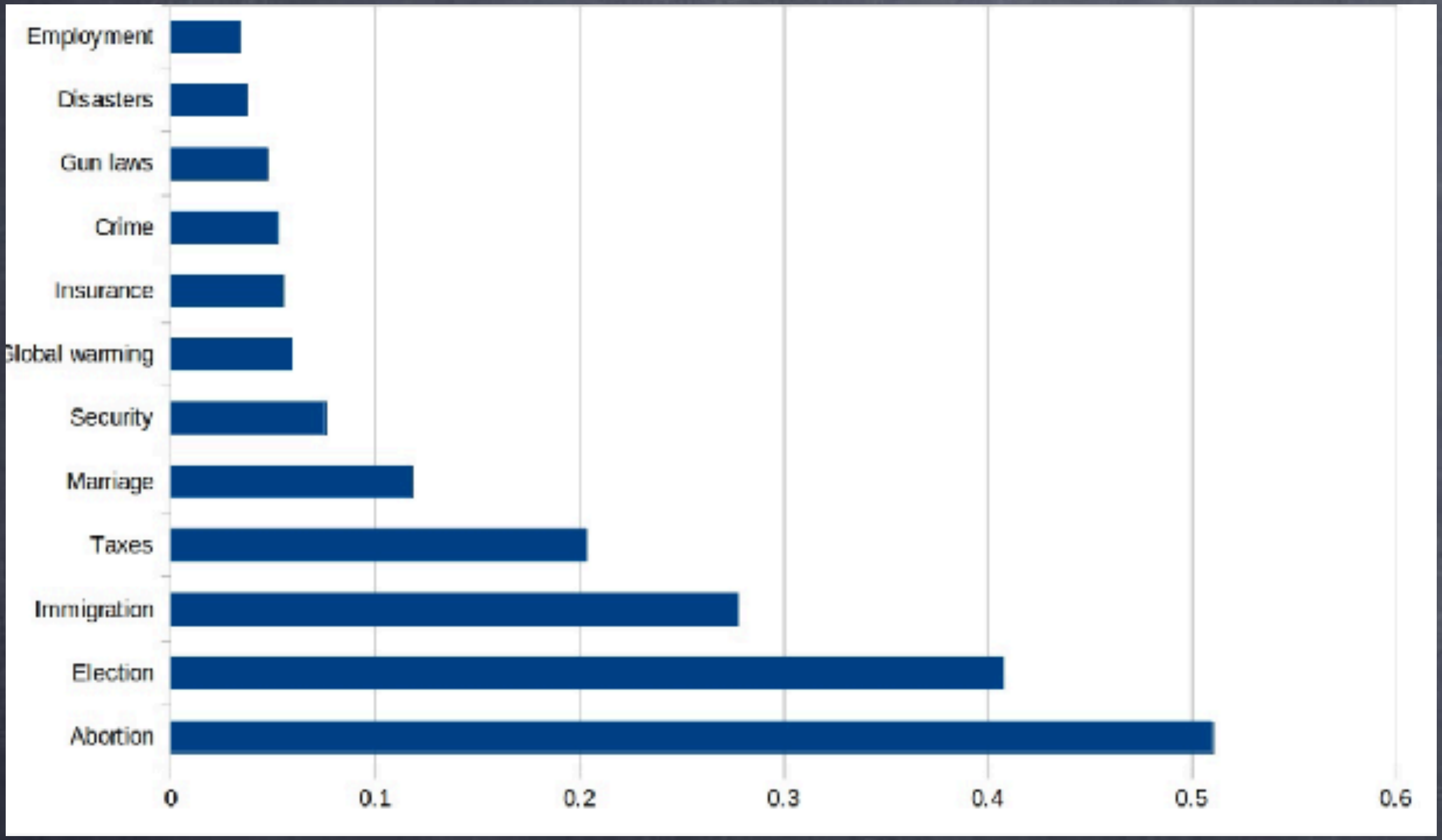
Insurance	**Gun laws**	**Crime**
health	people	scene
insurance	gun	police
people	laws	man
care	guns	fire
plan	control	suspect
Abortion	**Security/War**	**Employment**
abortion	attack	workers
baby	video	job
babies	security	wages
freedom	police	jobs
women	forces	people
Immigration	**Economy**	**Climate**
workers	tax	climate
immigration	jobs	people
stories	debt	change
patriot	taxes	warming
politics	spending	years
Marriage	**Election**	**Disasters**
people	bill	acres
freedom	vote	fire
marriage	campaign	weather
rights	state	snow
women	election	storm

Abortion		**Security/War**	
Join	Prolife	Killed	Military
Religious	Killed	Syrian	Illegal
Stand	Born	Military	Russian
Support	Unborn	Fast	Targeting
Conservative	Aborted	Furious	Back
Gun laws		**Immigration**	
Illegal	Dont	Join	Top
Free	Free	Support	Enter
Dont	Stop	Back	Check
Vote	Illegal	Stand	Stop
Stop	Give	Proud	Join
Insurance		**Marriage**	
Pay	Check	Back	Gay
Federal	Hear	Don	Religious
Signed	Here	Lost	Political
Paid	Call	Liberal	Free
Uninsured	Hope	Great	**

What are the most controversial issues? Can this be used to predict political affiliation?



What are the most controversial issues? Can this be used to predict political affiliation?



Approach	Accuracy (%)
Baseline: *Gottipati et al. (2013)*	60
Log likelihood-based	68

Outline

Research area of NLP	Specific Task	Statistical Algorithm
Machine Translation	Discovering aligned words	Expectation-Maximization (EM) Algorithm
Opinion Mining	Discovering political issues and orientations	Latent Dirichlet Allocation (LDA) Model
Sarcasm detection	Predicting sarcasm in dialogue	Hidden Markov Model (HMM)

Part 3 of 3:

Sarcasm Detection

Joint work with Vaibhav Tripathi, Pushpak Bhattacharyya and Mark Carman

Aditya Joshi, Vaibhav Tripathi, Pushpak Bhattacharyya and Mark J Carman, 'Harnessing Sequence Labeling for Sarcasm Detection in Dialogue from TV Series Friends', CONLL 2016, Berlin, Germany, August 2016.

Image: Maya Sarabhai, Sarabhai versus Sarabhai
Source: India Today

Sarcasm Detection

- Sarcasm is verbal irony that is intended to express contempt or ridicule
  - 'I love being ignored.'
- Past work in sarcasm detection: Features+Classifiers
- Applications?

Sarcasm Detection

- Sarcasm is verbal irony that is intended to express contempt or ridicule
  - 'I love being ignored.'
- Past work in sarcasm detection: Features+Classifiers
- Applications?

Sarcasm Detection of Dialogue: Motivation

Shubham: You are loving this talk!

Sarcasm Detection of Dialogue: Motivation

Shubham: You are loving this talk!

Shubham's neighbor: (takes notes, with a lot of interest)
Shubham: You are loving this talk!

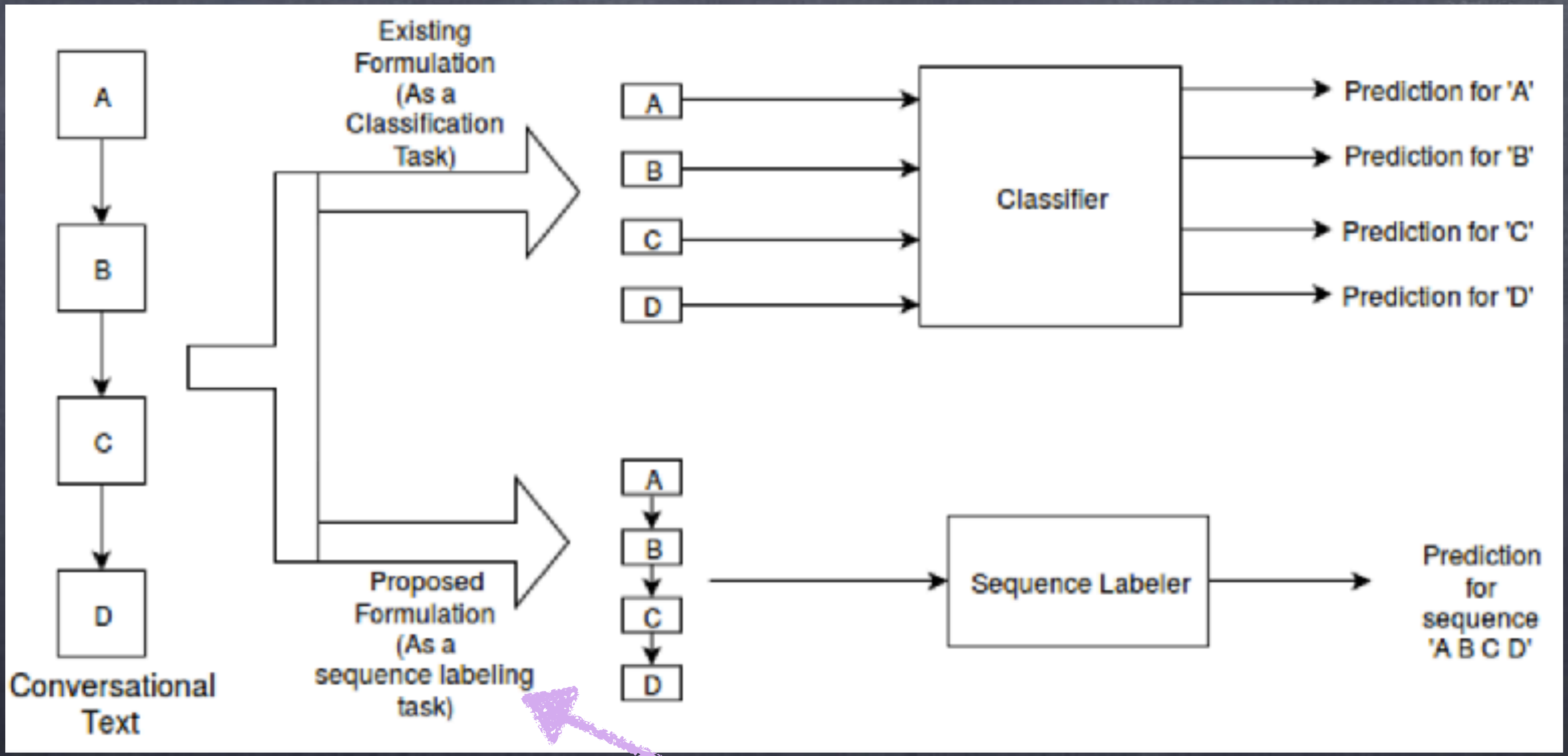
Sarcasm Detection of Dialogue: Motivation

Shubham: You are loving this talk!

Shubham's neighbor: (yawns) (yawns again)
Shubham: You are loving this talk!

Shubham's neighbor: (takes notes, with a lot of interest)
Shubham: You are loving this talk!

Sequence Labeling for Sarcasm Detection of Dialogue



SVM-HMM (a variant of HMM)

Dataset (1/2)

- Transcripts of TV Series 'Friends'
- 913 scenes, 17338 utterances, 1888 sarcastic utterances, average 18.6 utterances per scene

Dataset (2/2)

Character	% sarcastic
Phoebe	9.70
Joey	11.05
Rachel	9.74
Monica	8.87
Chandler	22.24
Ross	8.42

	Surface Positive Sentiment Score	Surface Negative Sentiment Score
Sarcastic	1.55	1.20
Non-sarcastic	0.97	0.75
All	1.03	0.79

	Actions (%)
Sarcastic	28.23
Non-sarcastic	23.95
All	24.43

Features & Results

Feature	Description
Lexical Features	
Spoken words	Unigrams of spoken words
Conversational Context Features	
Actions	Unigrams of action words
Sentiment Score	Positive & Negative score of utterance
Previous Sentiment Score	Positive & Negative score of previous utterance in the sequence
Speaker Context Features	
Speaker	Speaker of this utterance
Speaker-Listener	Pair of speaker of this utterance and speaker of the previous utterance

Algorithm	Precision (%)	Recall (%)	F-Score (%)
Formulation as Classification			
SVM (U)	83.6	48.6	57.2
SVM (O)	**84.4**	76.8	79.8
Naïve Bayes	77.2	33.8	42
Formulation as Sequence Labeling			
SVM-HMM	83.8	**88.2**	**84.2**
SEARN	82.6	83.4	82.8

Summary

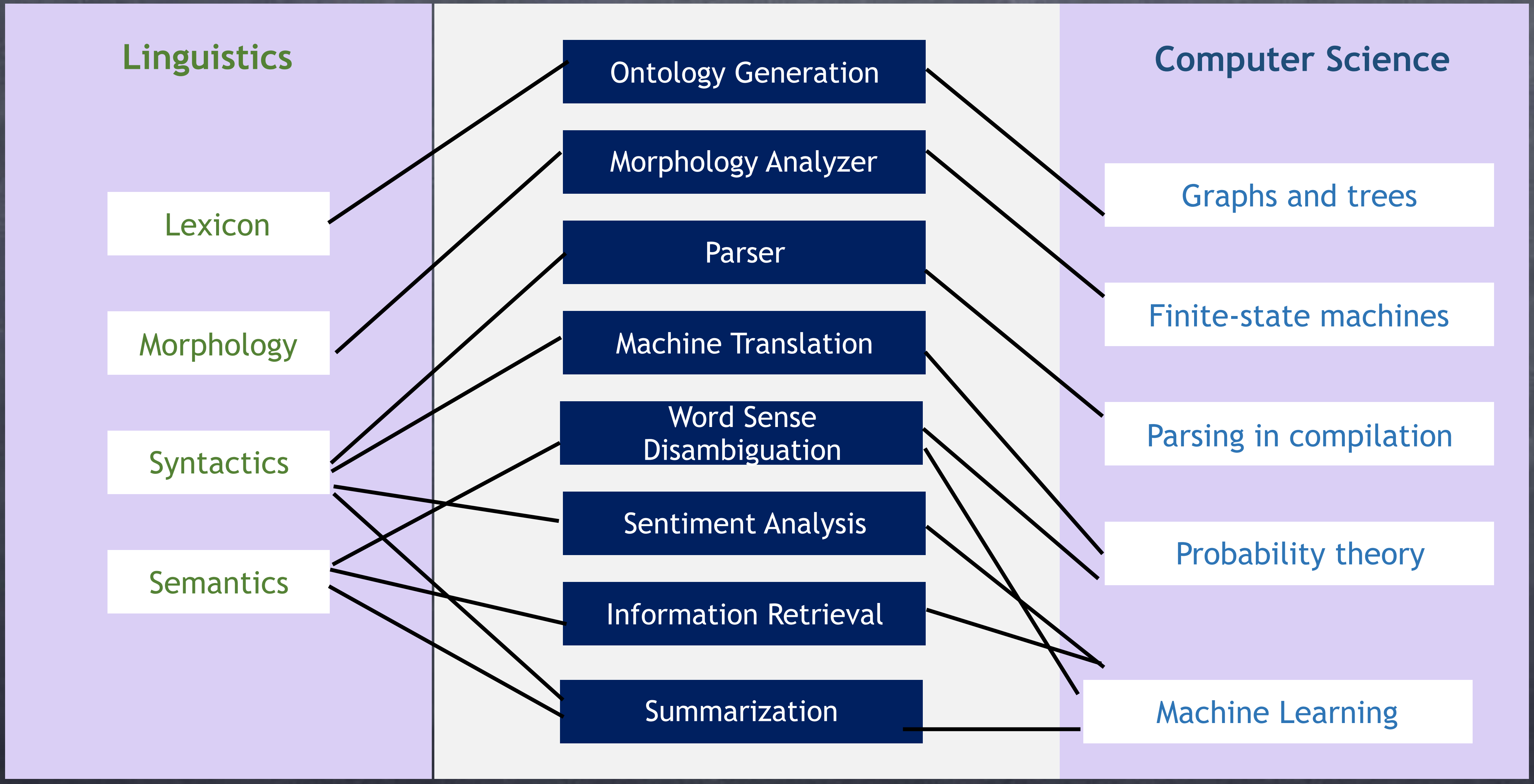
- We looked at three statistical NLP problems: machine translation, opinion mining and sarcasm detection
- Statistical NLP leverages on large scale data and machine learning to automate NLP tasks

NLP: At the confluence of linguistics and computer science

Linguistics		Computer Science
Lexicon		Graphs and trees
Morphology		Finite-state machines
Syntactics		Parsing in compilation
Semantics		Probability theory
		Machine Learning

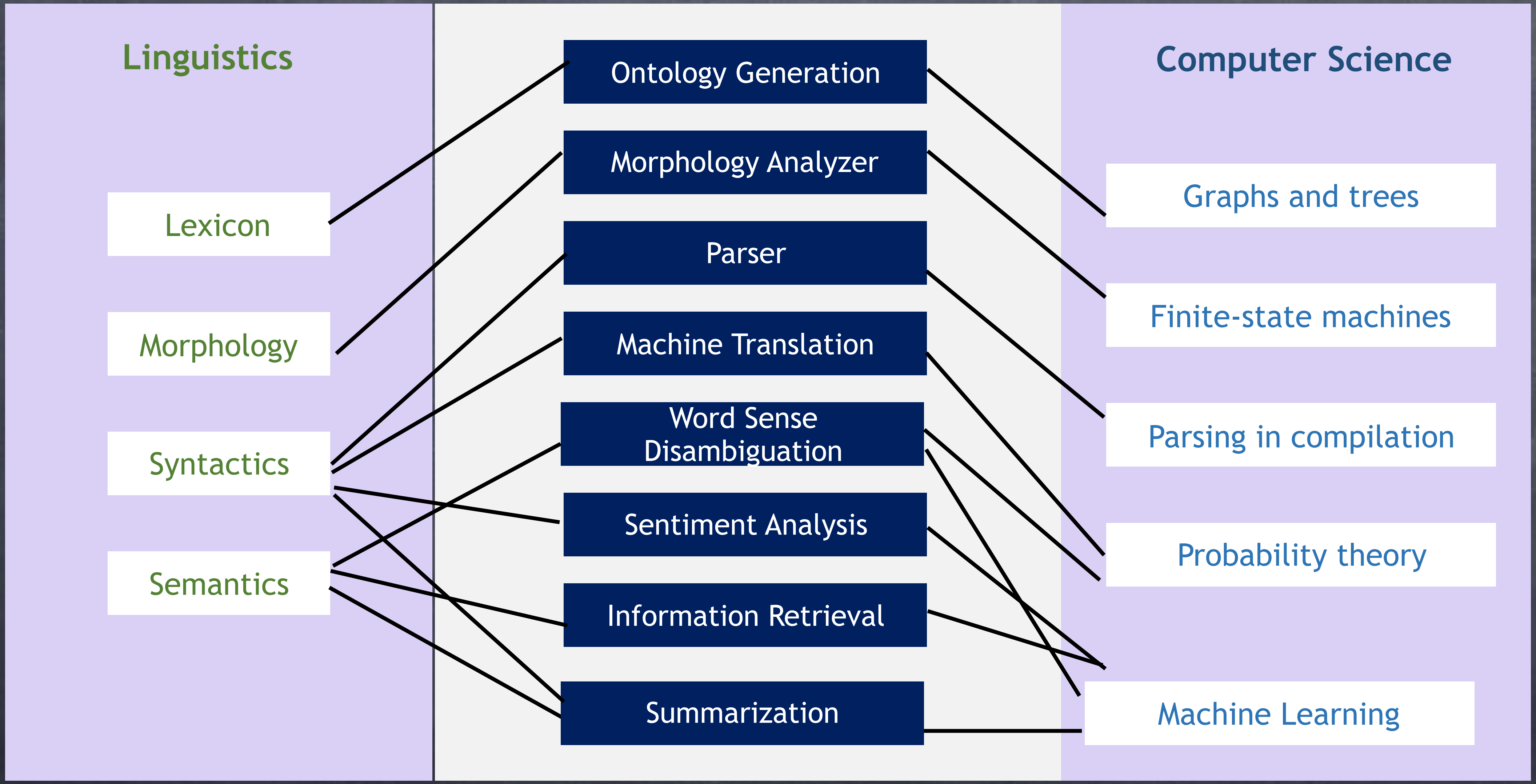
Credit: Rupa Hadaye, Cuddle.ai, Fractal Analytics

NLP: At the confluence of linguistics and computer science



Credit: Rupa Hadaye, Cuddle.ai, Fractal Analytics

NLP: At the confluence of linguistics and computer science



Credit: Rupa Hadaye, Cuddle.ai, Fractal Analytics

Errm.. you didn't really answer the question...

Isn't NLP all about a bunch of rules?!

Credit: Rupa Hadaye, Cuddle.ai, Fractal Analytics

Errm.. you didn't really answer the question...

Isn't NLP all about a bunch of rules?!

Statistical NLP is an alternative.

Credit: Rupa Hadaye, Cuddle.ai, Fractal Analytics

Errm.. you didn't really answer the question...

Isn't NLP all about a bunch of rules?!

Statistical NLP is an alternative.

How do I decide then?

Availability of data

Potential to formulate as a ML task

Credit: Rupa Hadaye, Cuddle.ai, Fractal Analytics

References

- Aditya Joshi, Vaibhav Tripathi, Pushpak Bhattacharyya and Mark J Carman, 'Harnessing Sequence Labeling for Sarcasm Detection in Dialogue from TV Series Friends', CONLL 2016, Berlin, Germany, August 2016.

- Aditya Joshi, Pushpak Bhattacharyya, Mark Carman, 'Political Issue Extraction Model: A Novel Hierarchical Topic Model That Uses Tweets By Political And Non-Political Authors', WASSA at NAACL 2016, San Diego, USA, June 2016.

- Aditya Joshi, Balamurali A.R., Pushpak Bhattacharyya and Rajat Mohanty, C-Feel-It: A Sentiment Analyzer for Micro-blogs, ACL 2011, Oregon, USA, June 2011.

- Van Slype, Georges. "Systran: Evaluation of the 1978 version of the Systran English-French automatic system of the commission of the European communities." The Incorporated Linguist 18.3 (1979): 1-4

- Bhattacharyya, Pushpak. Machine translation. CRC Press, 2015.