Learning word classes for semisupervised learning of NLP tasks

Grzegorz Chrupała

Saarland University

textkernel

The pain of feature engineering

Features

POS Spelling Morphology
WordNet VerbNet Gazeteers
Wikipedia Freebase ...

Learning features

- Distributed representations
- Dimensionality reduction techniques
- ✓ Word classes

Word classes

go come fit try hang read say take see blow
bricks bits food things medicine cream
the your that this a my his some

Berlin Bangkok Tokyo Warsaw
Sarkozy Merkel Obama Berlusconi
Mr Ms President Dr

- Groups of words sharing syntax/semantics
- Useful for generalization and abstraction

Word classes as features

Have been successfully used in

- Named Entity recognition
- Syntactic parsing
- Sentence retrieval

Main points

- Brown classes: good
- But can we do better?
- Soft word classes with LDA
  - Efficiency
  - Effectiveness as features in NLP tasks
- Beyond word classes

Brown clustering

- Brown et al propose their algorithm in 1992
- Agglomerative, hard clustering algorithm
- Mininizes MI between adjacent classes
- Still most commonly used word class type

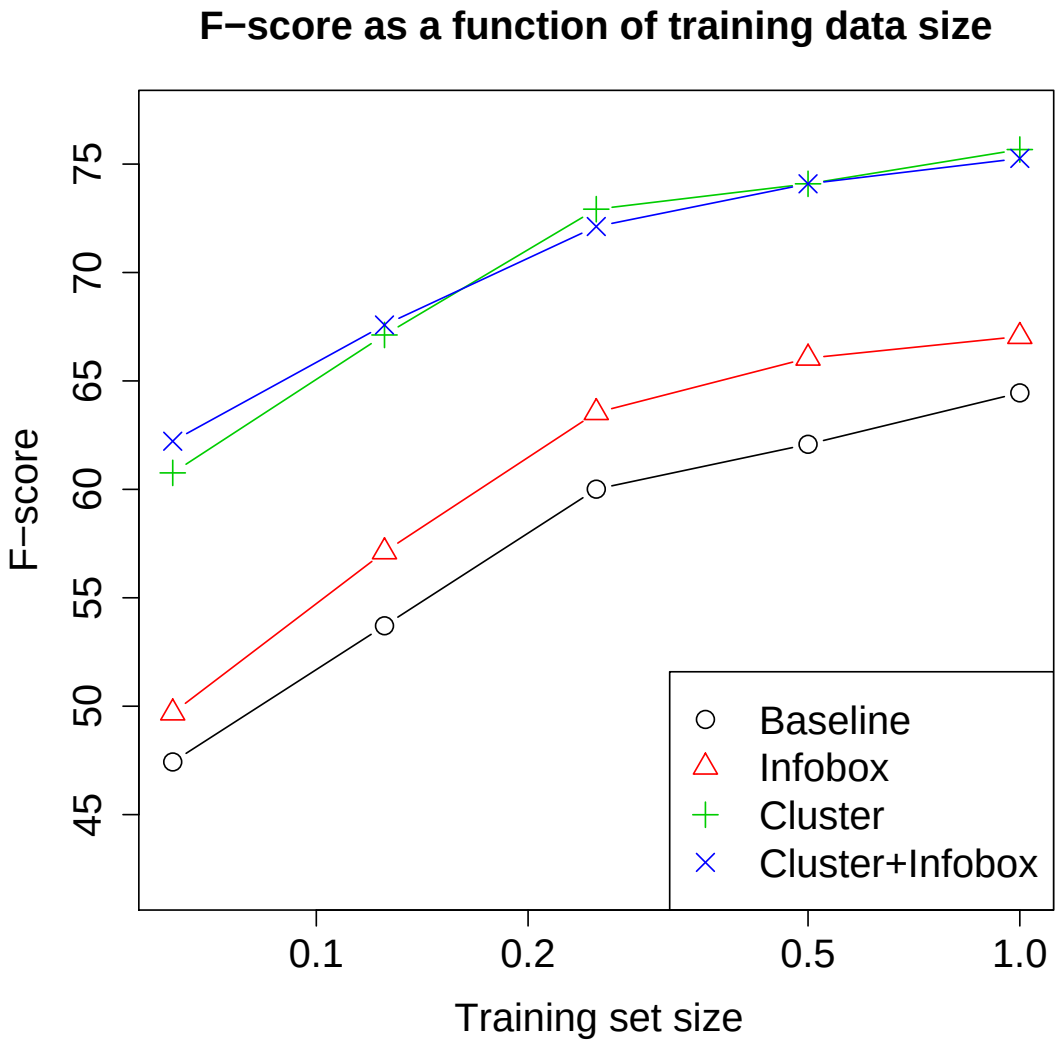
German NER with little effort

- 2009: no usable NE labelers for German
- ML + feature learning to the rescue
  - Sequence perceptron for supervised learning
  - CoNLL 2003 labeled data
  - Simple features: word, lemma, POS, context
  - ✓ Brown class IDs

Brown 500 from 32M words

Groß Rau Müller Zimmermann Frei Becker Möllemann Schmidt
Düsseldorf Berlin München Köln, Stuttgart Hannover Hamburg
nahmen macht zeigt gleichen bringt biete machte sorgt enthält

Brown and Wikipedia features



Brown's weaknesses

1. Time complexity:

$$O(K^2V)$$

Brown's weaknesses

1. Time complexity:

$$O(K^2V)$$

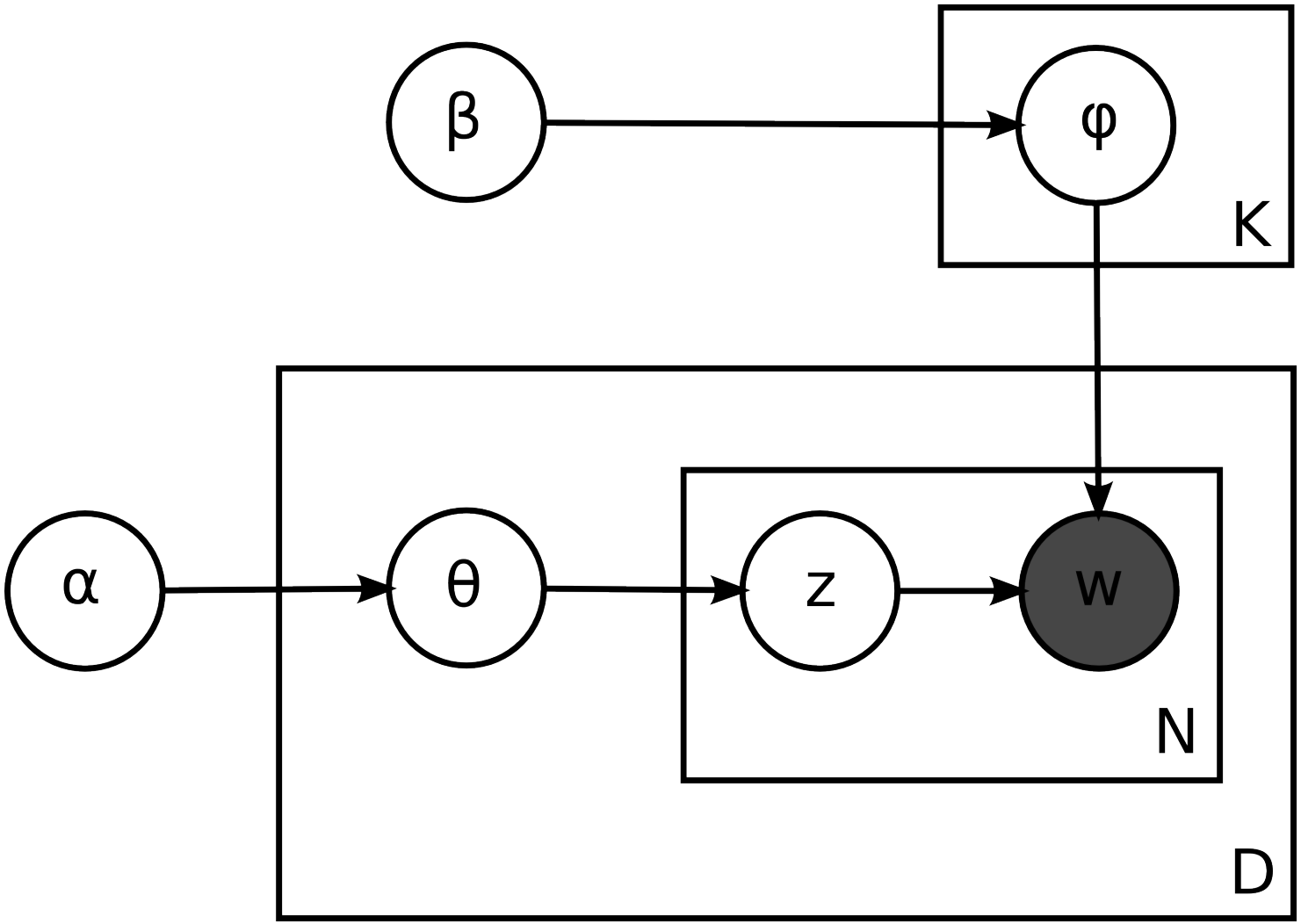
2. Hard clustering
   - Each word form assigned to only one class
   - Need separate classes for:
     - ★ first name
     - ★ last name
     - ★ first name OR last name
     - ★ last name OR city

Word class induction with LDA addresses both issues

LDA for topic modeling

- For each topic z draw ϕ_z from a Dirichlet
- For each document d
  - Draw a topic distribution θ_d from a Dirichlet
  - Repeat until generated all the words in d
    - ★ Draw a topic z from θ_d
    - ★ Draw a word w from the ϕ_z

LDA



Topic vs word classes

Topics	$\rightarrow$	Word classes
Documents	$\rightarrow$	Word types
Words	$\rightarrow$	Context features

Krzysztof

argues$_L$ argues$_R$ director$_L$ director$_L$ edits$_R$ said$_R$
Bledkowski$_R$ Kieslowski$_R$ Kieslowski$_R$
Rutkowski$_R$ Sikorski$_R$ and$_L$

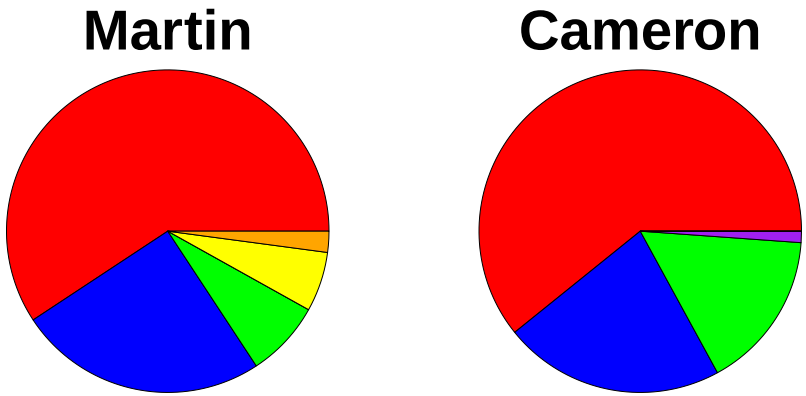
Generative process

- For each class z draw ϕ_z from a Dirichlet
- For each word type d
  - Draw a class distribution θ_d from a Dirichlet
  - Repeat
    - ★ Draw a word class z from θ_d
    - ★ Draw a context feature w from the ϕ_z

Induced distributions

- θ_d: class distribution given word type
- ϕ_z: feature distribution given class

Soft clustering



chief Gingrich Martin Newt Van Scott Roberts

Mr. Ms. John Robert President Dr. David

Street General Texas Fidelity State California

Context

Newt, Speaker	•	executive, operating
says, Chairman	•	Clinton, Dole, J.
Wall, West, East	•	County, AG, Journal

Graded similarity

1.8M CHILDES

train	car
give	bring
shoes	clothes
book	hole
monkey	rabbit

100M BNC

can	will
June	March
man	woman
black	white
business	language

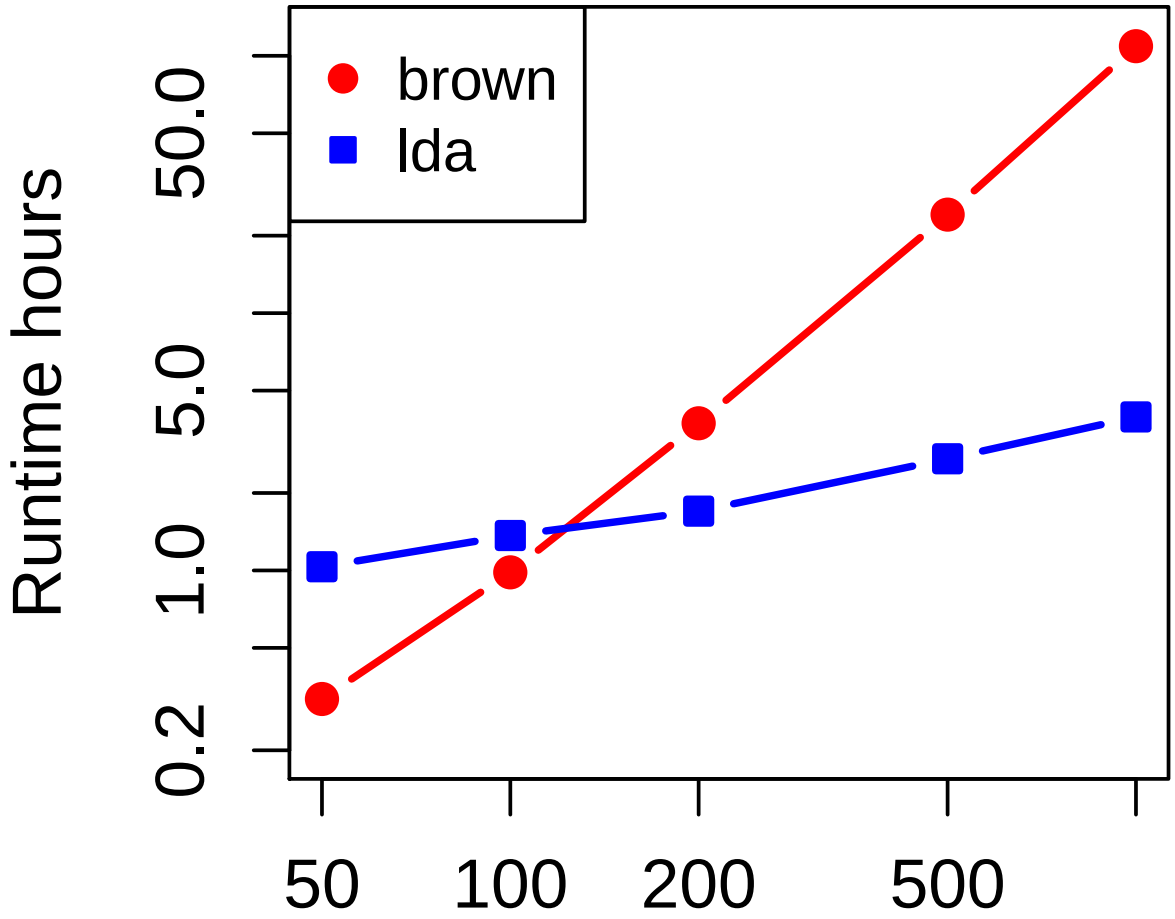
Efficiency

- Brown: $O(K^2V)$
- LDA: $O(KN)$
- Scaling feature counts by $\frac{1}{m}$ reduces LDA runtime m times

Testing efficiency in practice

- 60M words of North American News Text
- LDA, Brown: 100, 200, 500, 1000 classes
- LDA counts scaled by $\frac{1}{3}$

Runtimes



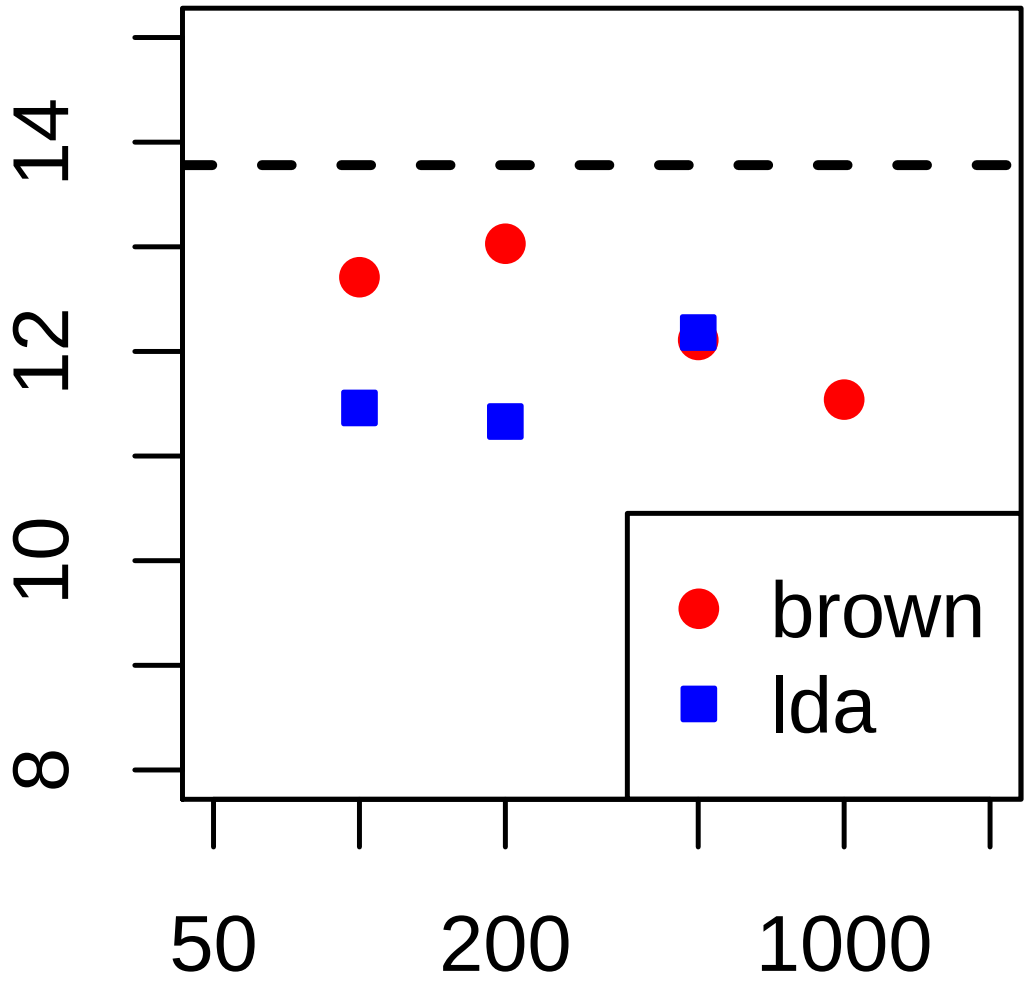
Semi-supervised learning performance

- Use word classes as features
- Brown
  - different levels of hierarchy
- LDA
  - class distributions and context information
- Explore several class granularities

Fine-grained NER on BBN

ANIMAL CARDINAL AGE DATE DURATION
DISEASE BUILDING HIGHWAY-STREET CITY
COUNTRY STATE-PROVINCE LAW CONTINENT
REGION MONEY NATIONALITY POLITICAL
ORDINAL CORPORATION EDUCATIONAL
GOVERNMENT PERCENT PERSON PLANT VEHICLE
WEIGHT CHEMICAL DRUG FOOD TIME

F1 error

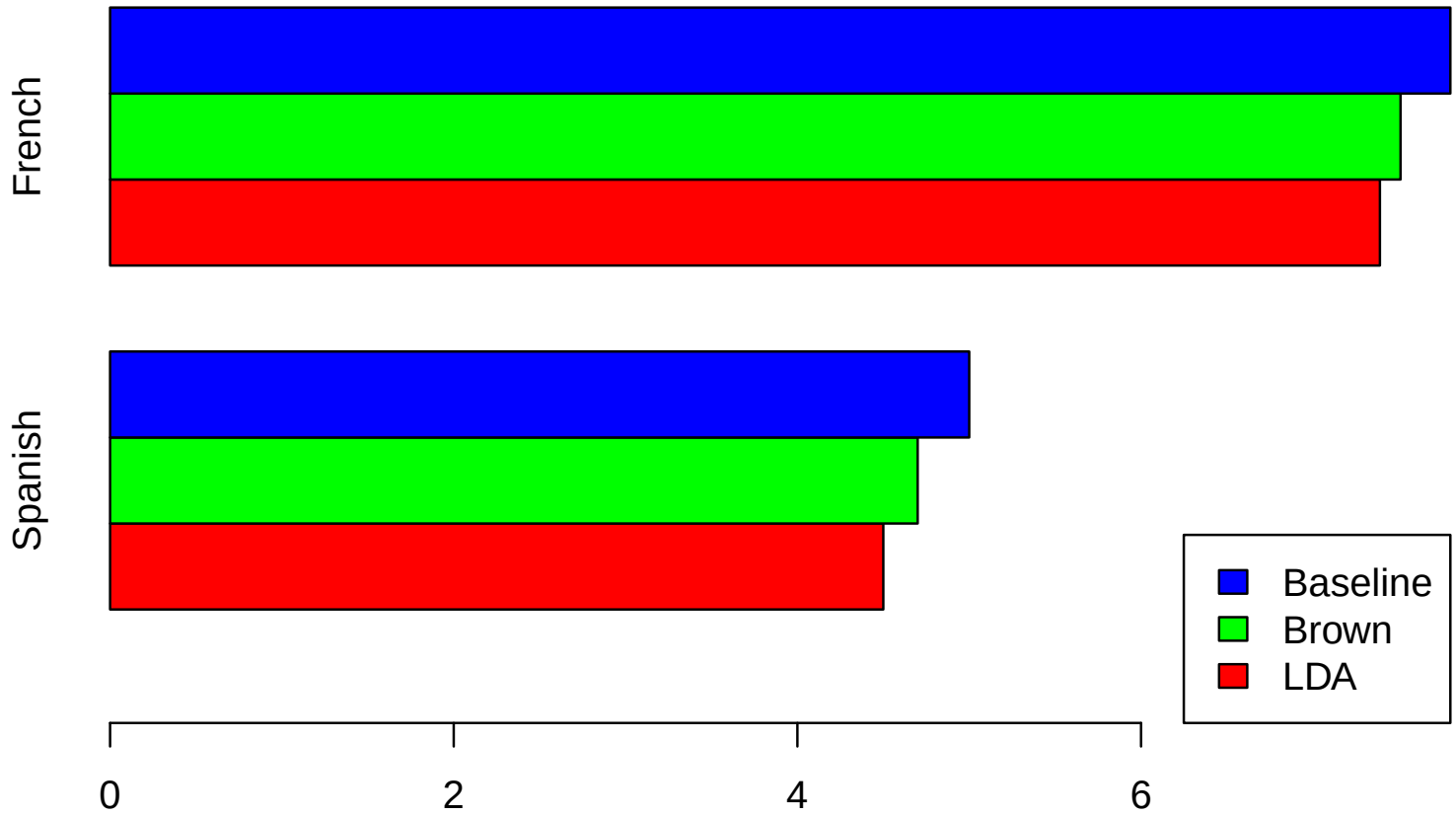


Morphological analysis

Token	Lemma	MSD	Gloss
Pero	pero	cc	but
cuando	cuando	cs	when
era	ser	vsii3s0	he was
niño	niño	ncms000	boy
le	el	pp3csd00	to him
gustaba	gustar	vmii3p0	it pleased

MA results with Morfette

- Brown: 500 classes
- LDA: 50 classes on Spanish, 100 on French



Semantic relation classification

- Task defined at Semeval 2007 and 2010
- *The* **bowl**$_{arg_1}$ *was full of apples,* **pears**$_{arg_2}$ *and oranges*
- CONTENT-CONTAINER(*pears*, *bowl*)

Relation inventory

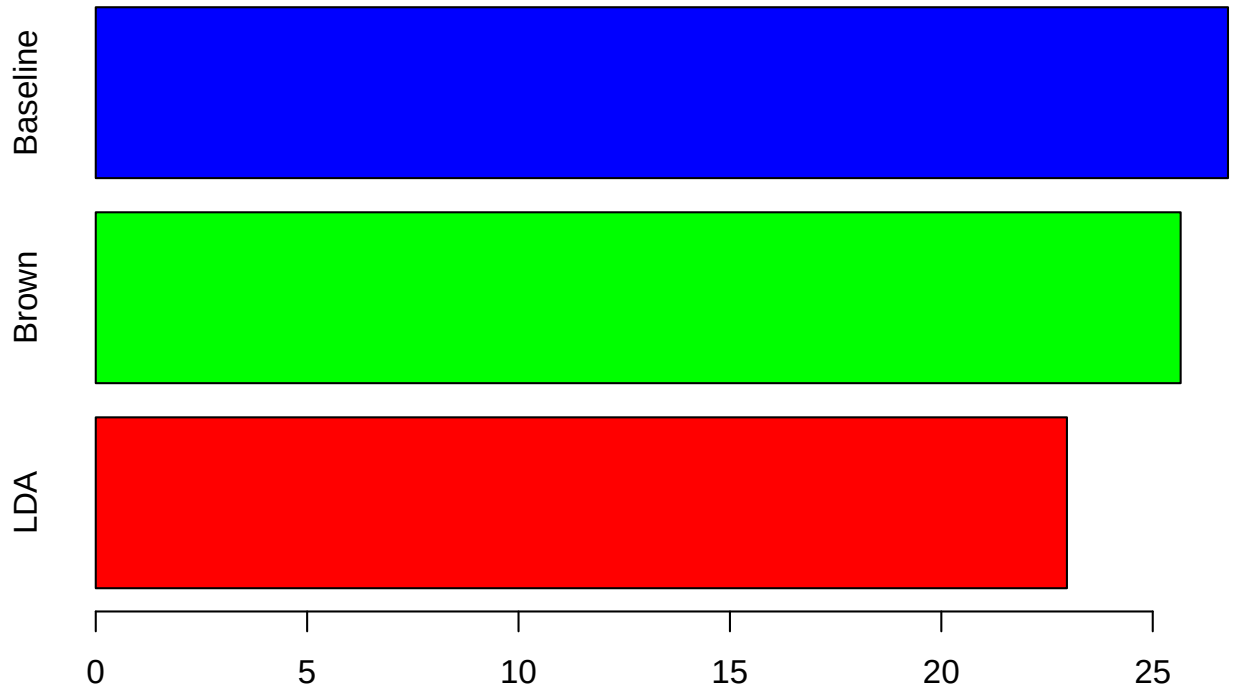
- CAUSE-EFFECT
- INSTRUMENT-AGENCY
- PRODUCT-PRODUCER
- CONTENT-CONTAINER
- ENTITY-ORIGIN
- ENTITY-DESTINATION
- COMPONENT-WHOLE
- MEMBER-COLLECTION
- COMMUNICATION-TOPIC

Features

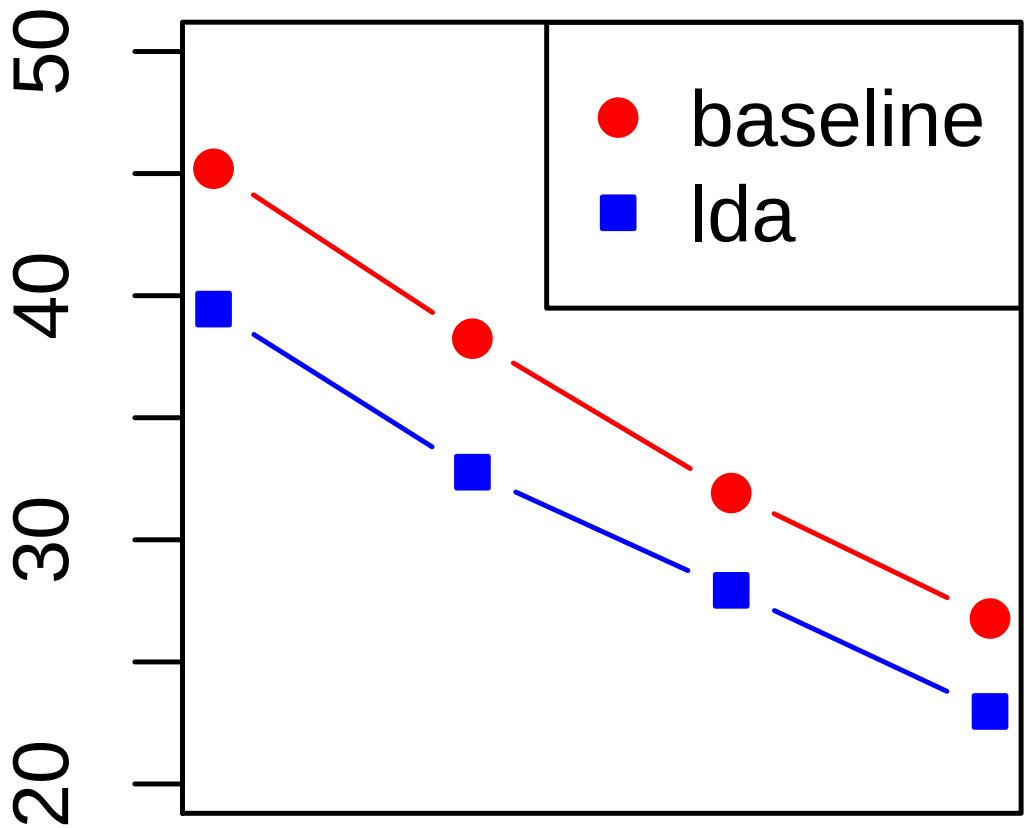
arg_1	The first argument
arg_2	The second argument
between	Tokens between arg_1 and arg_2
before	3 tokens before arg_1
after	3 tokens after arg_2
word classes for each of the above	

Relation classification results

- 500 Brown classes, 100 LDA classes



Relation classification



- LDA RC would rank third in Semeval 2010
- **Without** PropBank, FrameNet, WordNet, NomLex, Text Runner, Cyc...

Beyond word classes

- Chunks: small non-recursive typed fragments:

  $[\text{Mr. Mattei}]_{\text{NP}}$ $[\text{cited}]_{\text{VP}}$ $[\text{family reasons}]_{\text{NP}}$ $[\text{for}]_{\text{PP}}$ $[\text{his resignation}]_{\text{NP}}$
- Data: chunk trigrams centered on VP
- Hierarchical Bayesian model with latent variables

Example relation discovered

Argument 1	**Predicate**	**Argument 2**
court	rejected	plan
department	approve	bill
board	passed	proposal
congress	consider	legislation
judge	authorized	program

To conclude:

- Efficiently **learn**
- **expressive** features from large datasets
- to solve NLP tasks with
- **less annotation** and **less engineering**

Thank you

Online learning of word classes

- Want to learn from evolving data streams
- Online LDA compared by Canini et al 2005 for topic modeling
- Only one, oLDA, strictly online
- oLDA did not work very well for inferring document topic

Word classes with online LDA (CoLaDA)

- d - word type
- w - context feature
- z - class
- Replicate incoming sentence j times
  - For each w_i in the sentence, sample:

$$P(z_i|\mathbf{z}_{i-1}, \mathbf{w}_i, \mathbf{d}_i) \propto \frac{(n_{z,d} + \alpha) \times (n_{z,w} + \beta)}{n_{z,\bullet} + V\beta}$$

and update the counts.

CoLaDA

- oLDA did not work for inferring topics
- Key difference: word types d recur

CoLaDA

- oLDA did not work for inferring topics
- Key difference: word types d recur

  - Classes for common word types will be **frequently resampled**
  - **Without any special arrangements**

Task: Word prediction

- (Soft)-assign classes from context
- Rank words based on predicted class

Reciprocal rank

want␣to | put | them␣on

Task: Word prediction

- (Soft)-assign classes from context
- Rank words based on predicted class

Reciprocal rank

want_to	put y_{123}	them_on

y_{123}
make
take
put
get
sit
eat
let

$rank^{-1} = \frac{1}{3}$

CoLaDA results