# Efficient induction of probabilistic word classes with LDA

Grzegorz Chrupała

Saarland University

IJCNLP 2011

# Word classes

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

# Brown clustering

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

# Brown's weaknesses

1. Time complexity:

$$O(K^2V)$$

# Brown's weaknesses

1. Time complexity:

$$O(K^2V)$$

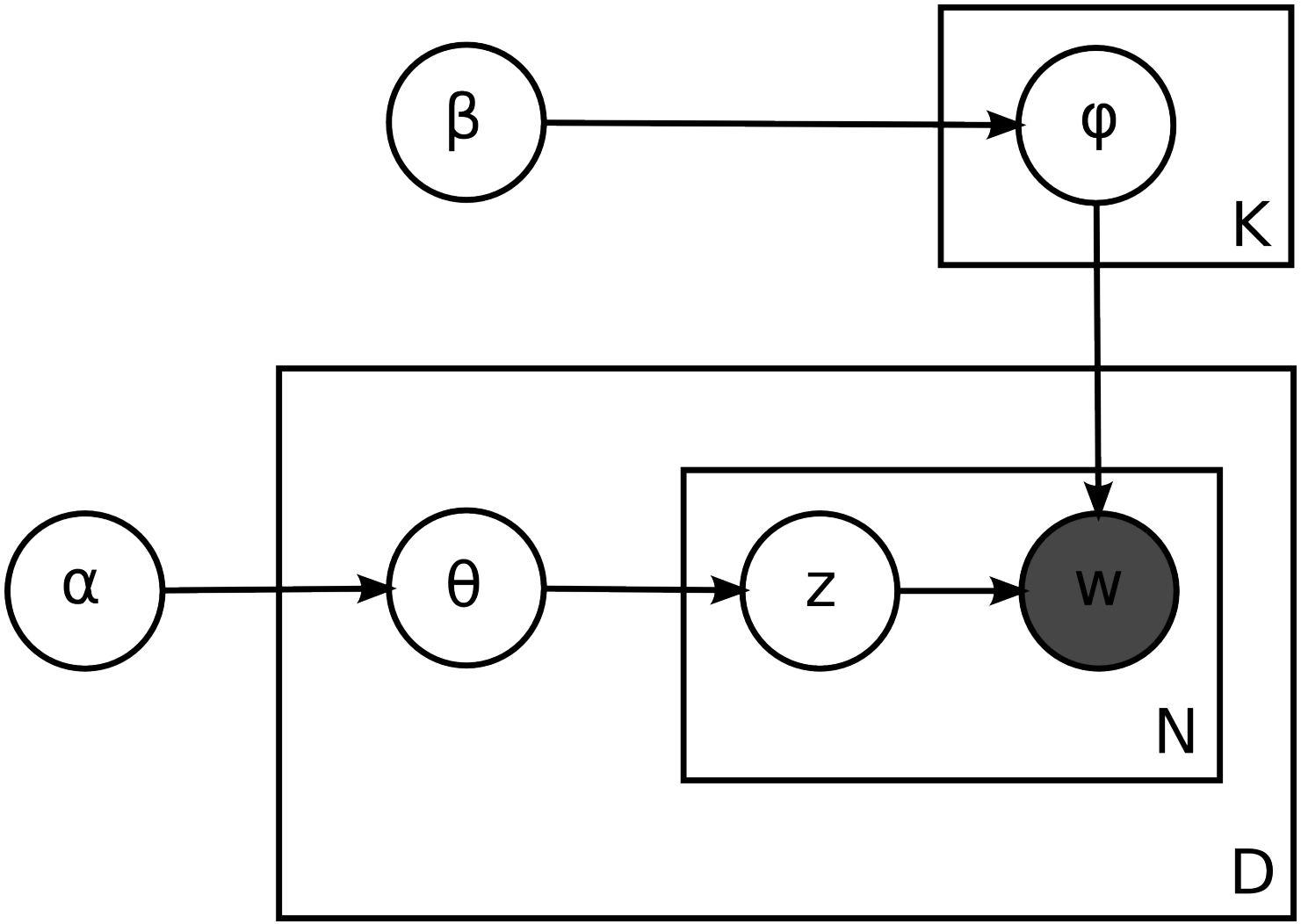
2. Hard clustering
   - Each word form assigned to only one class
   - Need separate classes for:
     - ★ first name
     - ★ last name
     - ★ first name OR last name
     - ★ last name OR city

# Word class induction with LDA addresses both issues

# LDA for topic modeling

- For each topic $z$ draw $\phi_z$ from a Dirichlet
- For each document $d$
  - Draw a topic distribution $\theta_d$ from a Dirichlet
  - Repeat until generated all the words in $d$
    - ★ Draw a topic $z$ from $\theta_d$
    - ★ Draw a word $w$ from the $\phi_z$

# 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$_L$

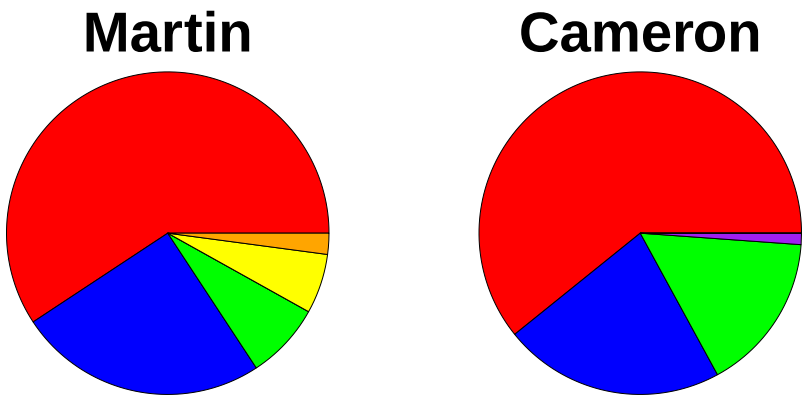
# Generative process

- For each class $z$ draw $\phi_z$ from a Dirichlet
- For each word type $d$
  - Draw a class distribution $\theta_d$ from a Dirichlet
  - Repeat
    - ★ Draw a word class $z$ from $\theta_d$
    - ★ Draw a context feature $w$ from the $\phi_z$

# Induced distributions

- $\theta_d$: class distribution given word type
- $\phi_z$: 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 |

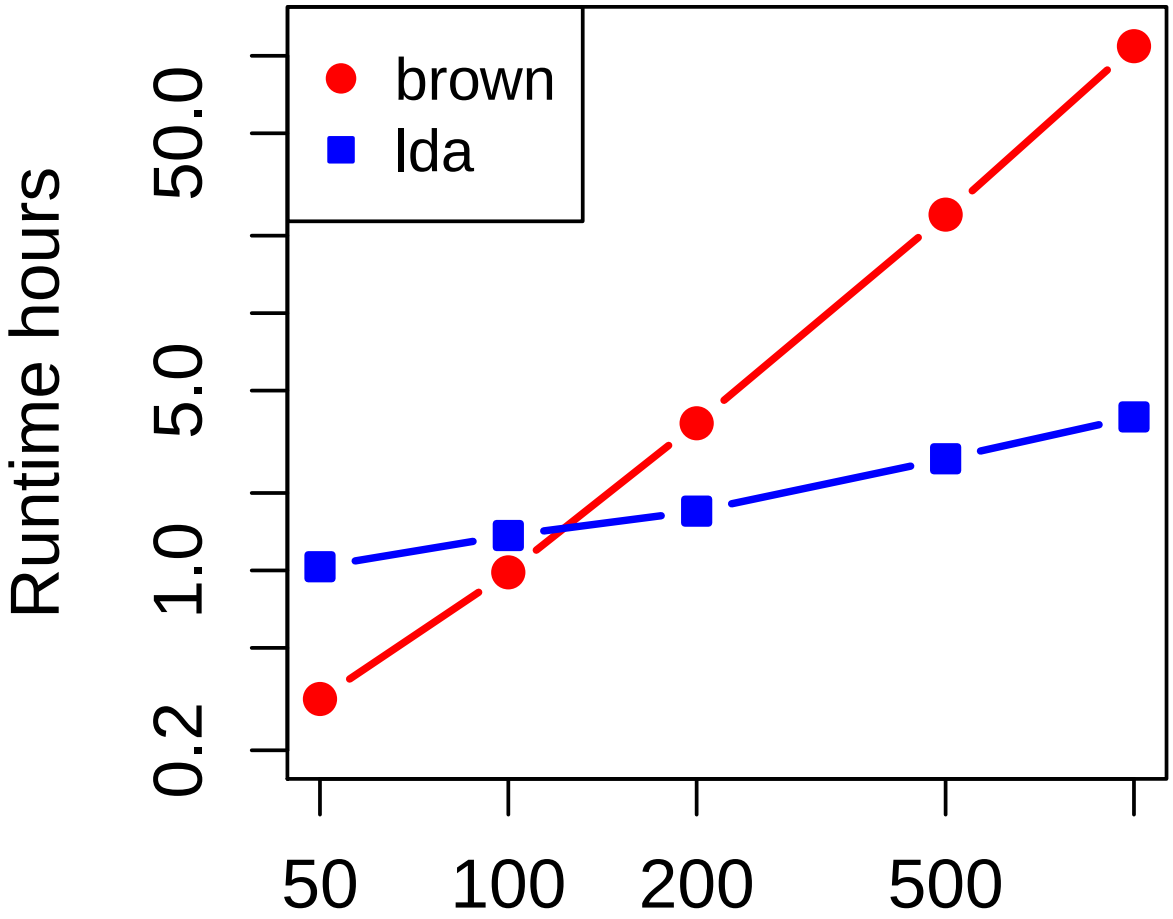
# Efficiency

- 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



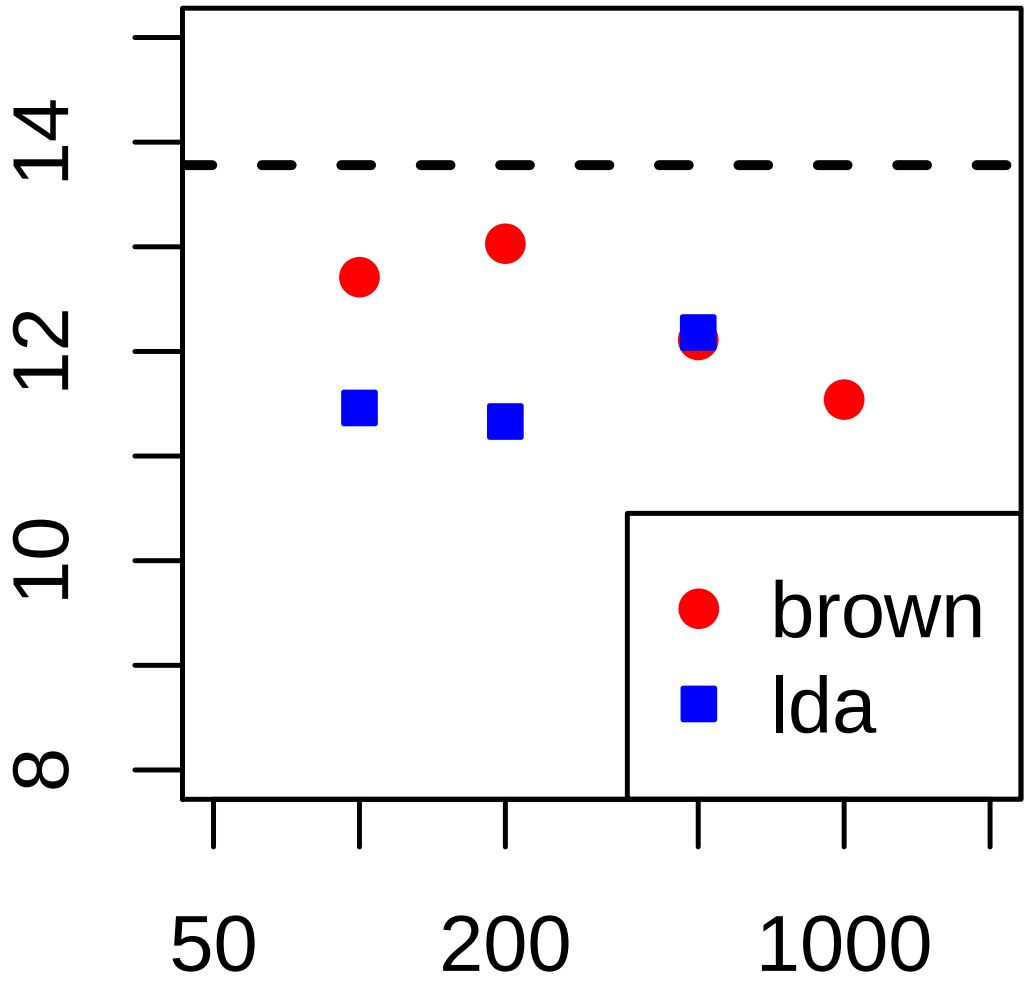
# Semi-supervised learning performance

- 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

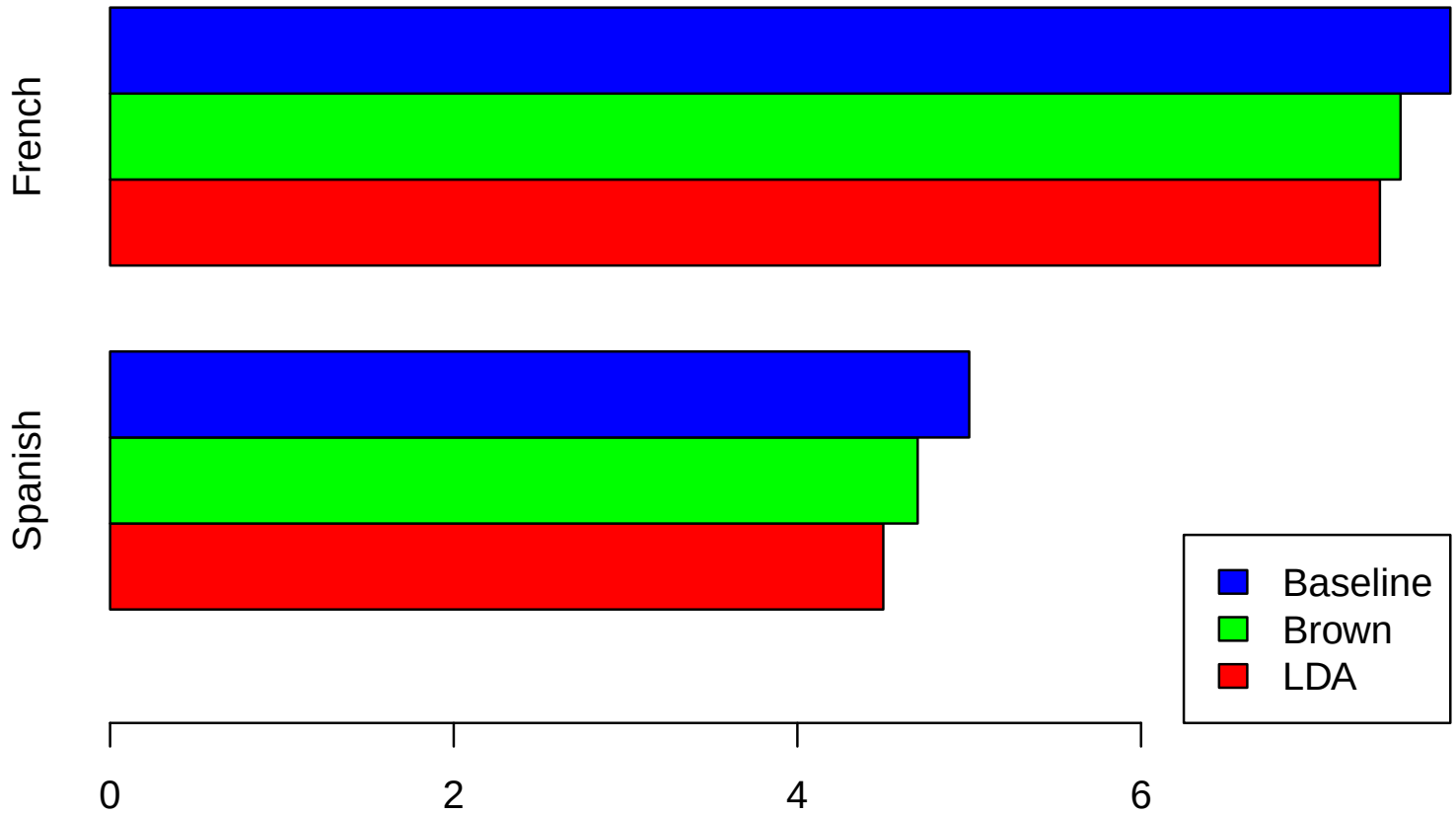


# Morphological analysis

| 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

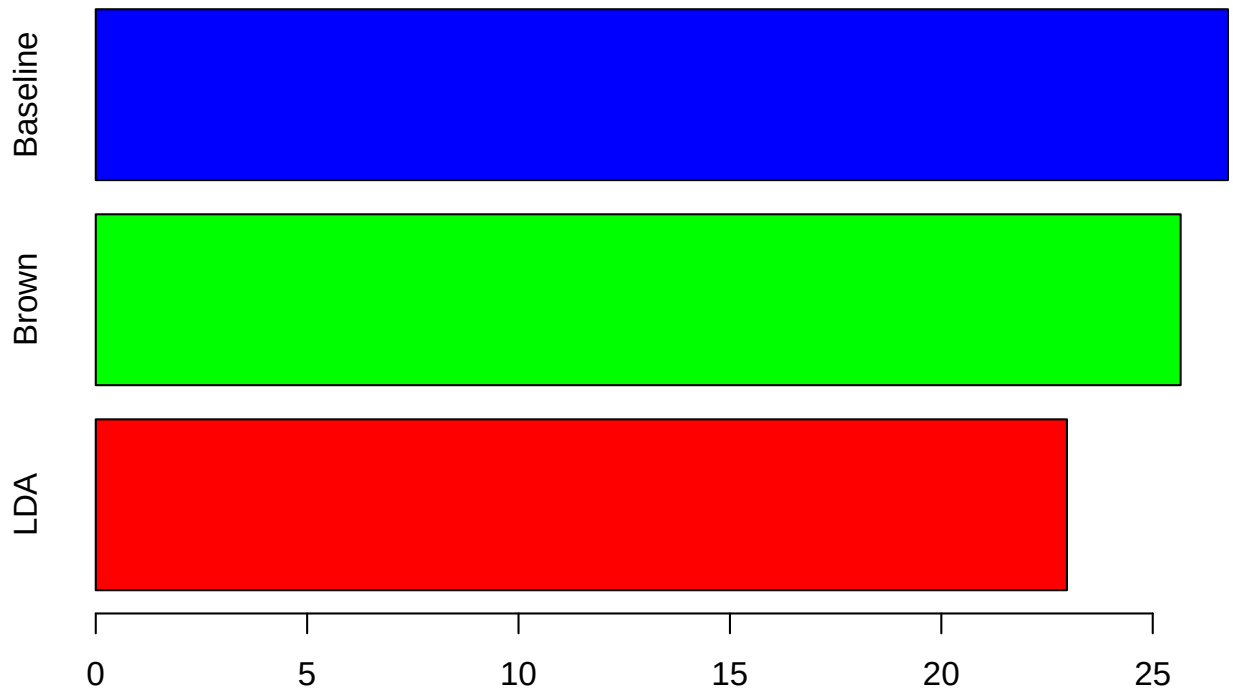
- Task defined at Semeval 2007 and 2010
- *The* **bowl** *was full of apples,* **pears** *and oranges*
- CONTENT-CONTAINER(*pears*, *bowl*)

# Relation inventory

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

# Relation classification results

- 500 Brown classes, 100 LDA classes



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

# To conclude:

- **Efficient** induction of
- **Probabilistic** word classes which
- **Match** or **improve** on hierarchical Brown classes

# Thank you

# Relation classification