University of Maribor

Faculty of Electrical Engineering
and Computer Science

FERI

Semi-supervised text classification using topic models

Miha Pavlinek

15th Workshop on "Software Engineering
Education and Reverse Engineering"
Bohinj, Slovenia August 23rd – 30th 2015

Topic modeling

- Topic modeling is a process for finding semantically related clusters of words in text corpora – topics
  - Latent Semantic Analysis - LSA
  - Probabilistic LSA - pLSA
  - Latent Dirichlet Allocation (LDA)
    - Most efective **generative** statistical model, which combines semantically related concepts.

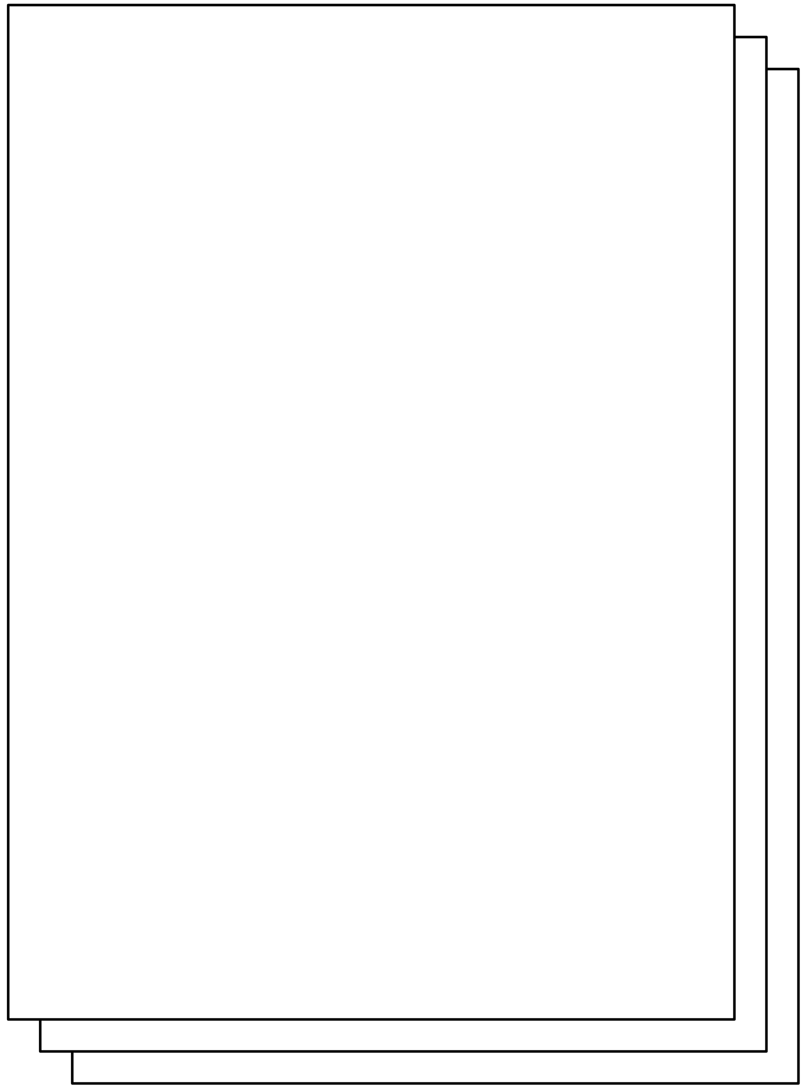
Intuition behind LDA

Documents

For many years, films about football were a bit of a joke. Having John Huston behind the camera meant that 1981's Escape to Victory - with Pelé up front, Ipswich's Russell Osman at the back and Sylvester Stallone in goal - was one of the best of them. But, as Huston's son Danny admitted to me recently, his father had never watched a game in his life and didn't even know how many players a football team should have on each side.
Despite advances in special effects, the spontaneity and skill of a real game is still extremely hard to stage, yet films with football at their core have become increasingly respectable of late. In successive years, the Cannes Film Festival, under the auspices of its Olympic Lyonnais-supporting director Thierry Frémaux, has welcomed films about Zinedine Zidane, Diego Maradona and, only last month, Eric Cantona.
Douglas Gordon and Philippe Parreno's Zidane - a tightly focused study of the former French captain playing a match for Real Madrid - remains, for me, the best work about actually playing the game. Director Emir Kusturica's documentary Maradona is more about the iconography of the Argentinian star.
In Looking for Eric, Loach, working with regular photographer Barry Ackroyd, stays true to his usual realist vision of working Britain, yet also manages to make one of his least typical films in that for much of its first hour, it plays as a comedy. Steve Evets gives a gruff but tender performance as Eric Bishop, the struggling postie whose life has gradually fallen apart. Living with two stoned stepsons, he can't summon the courage to reconnect with his ex-wife, Lily.

Intuition behind LDA

Documents



Intuition behind LDA

Topics

Documents

Topic 1	
football	0,04
goal	0,03
game	0,02
play	0,02
team	0,01
…	

Topic 2	
movie	0,06
actor	0,03
play	0,01
film	0,01
camera	0,01
…	

Topic 3	
father	0,02
son	0,02
life	0,02
wife	0,01
parent	0,01
…	

…

Intuition behind LDA

Topics

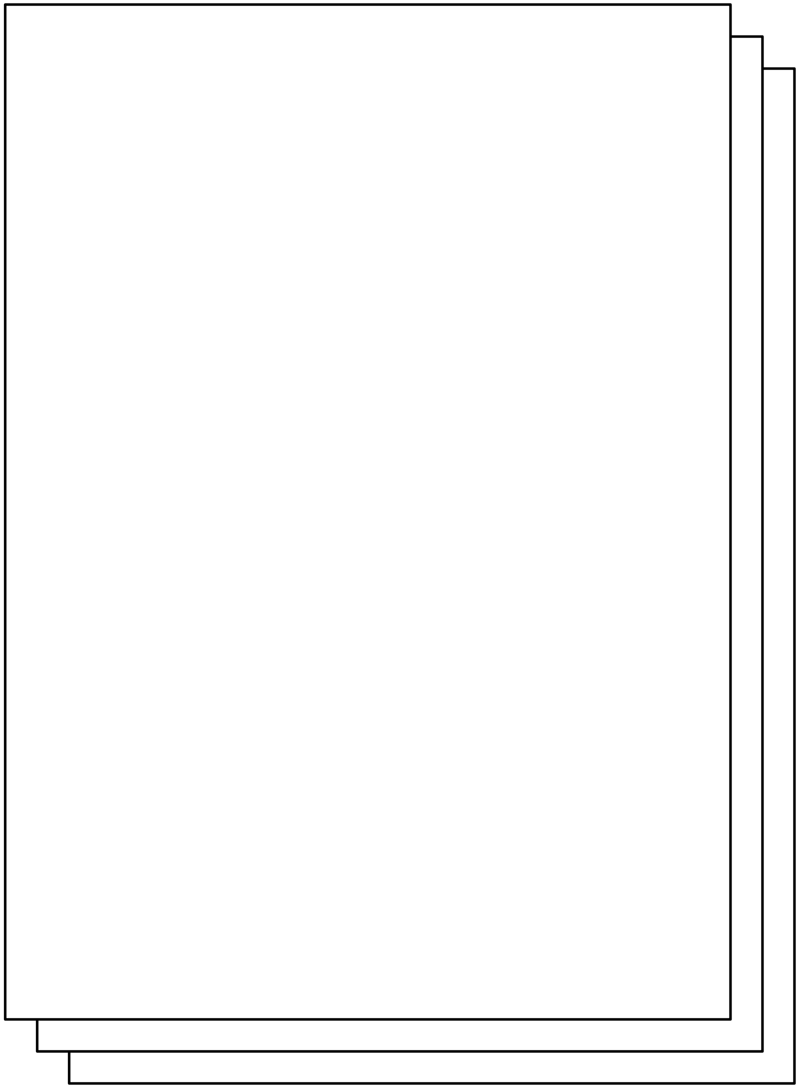
Topic 1	
football	0,04
goal	0,03
game	0,02
play	0,02
team	0,01
…	

Topic 2	
movie	0,06
actor	0,03
play	0,01
film	0,01
camera	0,01
…	

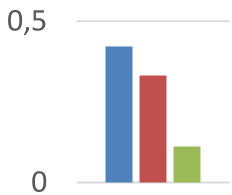
Topic 3	
father	0,02
son	0,02
life	0,02
wife	0,01
parent	0,01
…	

…

Documents



Topic proportions



Intuition behind LDA

Topics

Topic 1

football	0,04
goal	0,03
game	0,02
play	0,02
team	0,01
…	

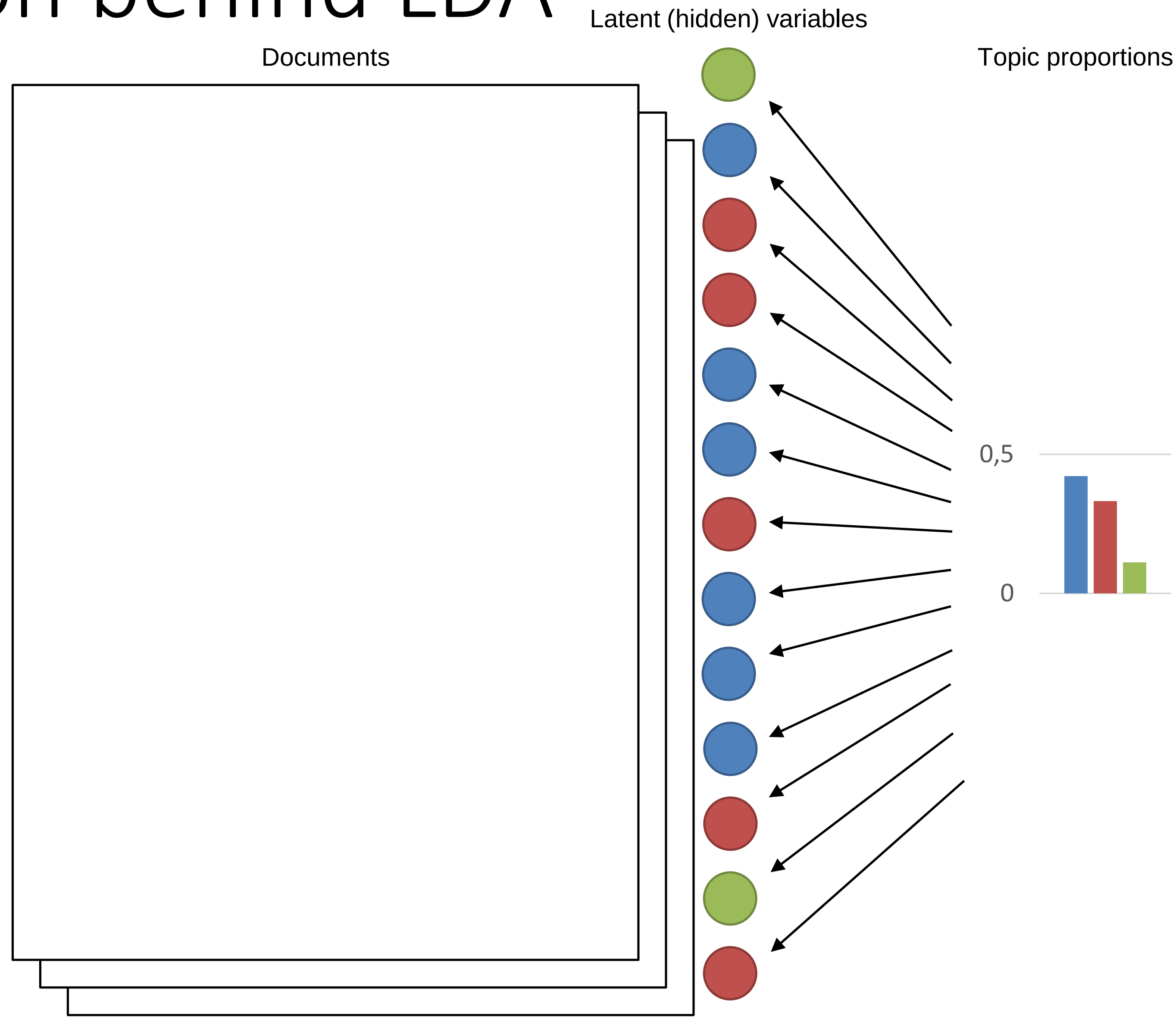
Topic 2

movie	0,06
actor	0,03
play	0,01
film	0,01
camera	0,01
…	

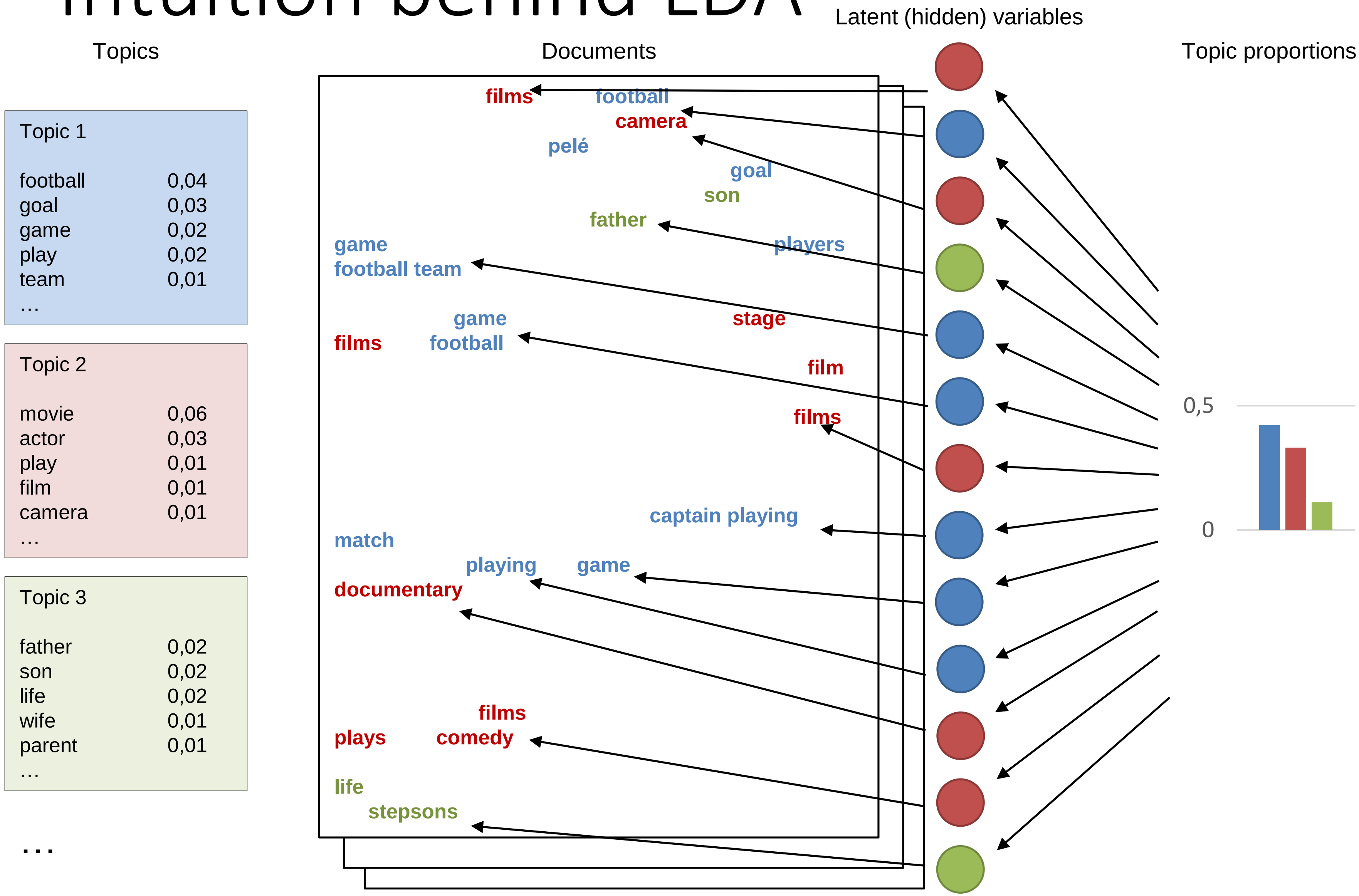
Topic 3

father	0,02
son	0,02
life	0,02
wife	0,01
parent	0,01
…	

…



Intuition behind LDA



Intuition behind LDA

Topics

Topic 1	
football	0,04
goal	0,03
game	0,02
play	0,02
team	0,01
…	

Topic 2	
movie	0,06
actor	0,03
play	0,01
film	0,01
camera	0,01
…	

Topic 3	
father	0,02
son	0,02
life	0,02
wife	0,01
parent	0,01
…	

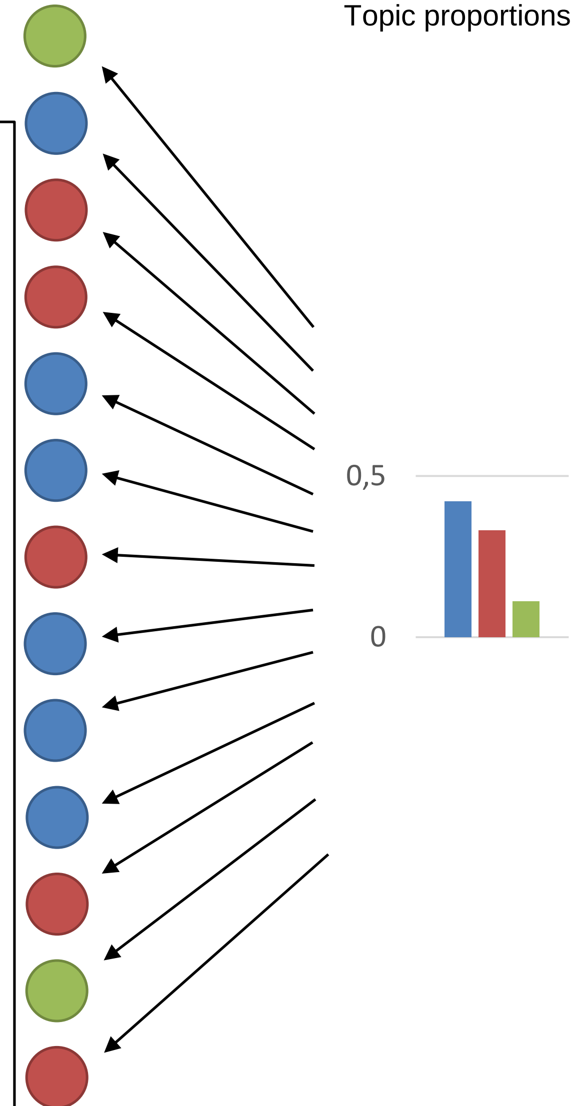
…

Documents

For many years, **films** about **football** were a bit of a joke. Having John Huston behind the **camera** meant that 1981's Escape to Victory - with **Pelé** up front, Ipswich's Russell Osman at the back and Sylvester Stallone in **goal** - was one of the best of them. But, as Huston's **son** Danny admitted to me recently, his **father** had never watched a **game** in his life and didn't even know how many **players** a **football team** should have on each side.
Despite advances in special effects, the spontaneity and skill of a real **game** is still extremely hard to **stage**, yet **films** with **football** at their core have become increasingly respectable of late. In successive years, the Cannes **Film** Festival, under the auspices of its Olympic Lyonnais-supporting director Thierry Frémaux, has welcomed **films** about Zinedine Zidane, Diego Maradona and, only last month, Eric Cantona.
Douglas Gordon and Philippe Parreno's Zidane - a tightly focused study of the former French **captain playing** a **match** for Real Madrid - remains, for me, the best work about actually **playing** the **game**. Director Emir Kusturica's **documentary** Maradona is more about the iconography of the Argentinian star.
In Looking for Eric, Loach, working with regular photographer Barry Ackroyd, stays true to his usual realist vision of working Britain, yet also manages to make one of his least typical **films** in that for much of its first hour, it **plays** as a **comedy**. Steve Evets gives a gruff but tender performance as Eric Bishop, the struggling postie whose **life** has gradually fallen apart. Living with two stoned **stepsons**, he can't summon the courage to reconnect with his ex-**wife**, Lily.

Latent (hidden) variables

Topic proportions



Intuition behind LDA

Latent (hidden) variables

Topics

Topic 1

Topic 2

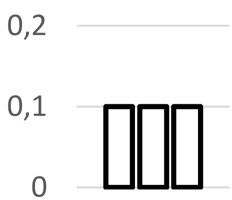
Topic 3

...

Documents

For many years, films about football were a bit of a joke. Having John Huston behind the camera meant that 1981's Escape to Victory - with Pelé up front, Ipswich's Russell Osman at the back and Sylvester Stallone in goal - was one of the best of them. But, as Huston's son Danny admitted to me recently, his father had never watched a game in his life and didn't even know how many players a football team should have on each side.
Despite advances in special effects, the spontaneity and skill of a real game is still extremely hard to stage, yet films with football at their core have become increasingly respectable of late. In successive years, the Cannes Film Festival, under the auspices of its Olympic Lyonnais-supporting director Thierry Frémaux, has welcomed films about Zinedine Zidane, Diego Maradona and, only last month, Eric Cantona.
Douglas Gordon and Philippe Parreno's Zidane - a tightly focused study of the former French captain playing a match for Real Madrid - remains, for me, the best work about actually playing the game. Director Emir Kusturica's documentary Maradona is more about the iconography of the Argentinian star.
In Looking for Eric, Loach, working with regular photographer Barry Ackroyd, stays true to his usual realist vision of working Britain, yet also manages to make one of his least typical films in that for much of its first hour, it plays as a comedy. Steve Evets gives a gruff but tender performance as Eric Bishop, the struggling postie whose life has gradually fallen apart. Living with two stoned stepsons, he can't summon the courage to reconnect with his ex-wife, Lily.

Topic proportions



Example of real topics

computer	health	espn	imdb	movie
security	heart	sports	database	film
network	disease	news	title	actor
virus	cholesterol	baseball	movie	reviews
spam	food	scores	celebs	episode
spyware	life	nba	internet	scripts
home	nutrition	stats	management	cinema
anti	conditions	game	spielberg	character
internet	living	basketball	search	dvd
users	medical	college	character	scene
guide	healthy	standings	festival	star
information	risk	team	award	action
email	tips	player	board	news

Semi-Supervised learning

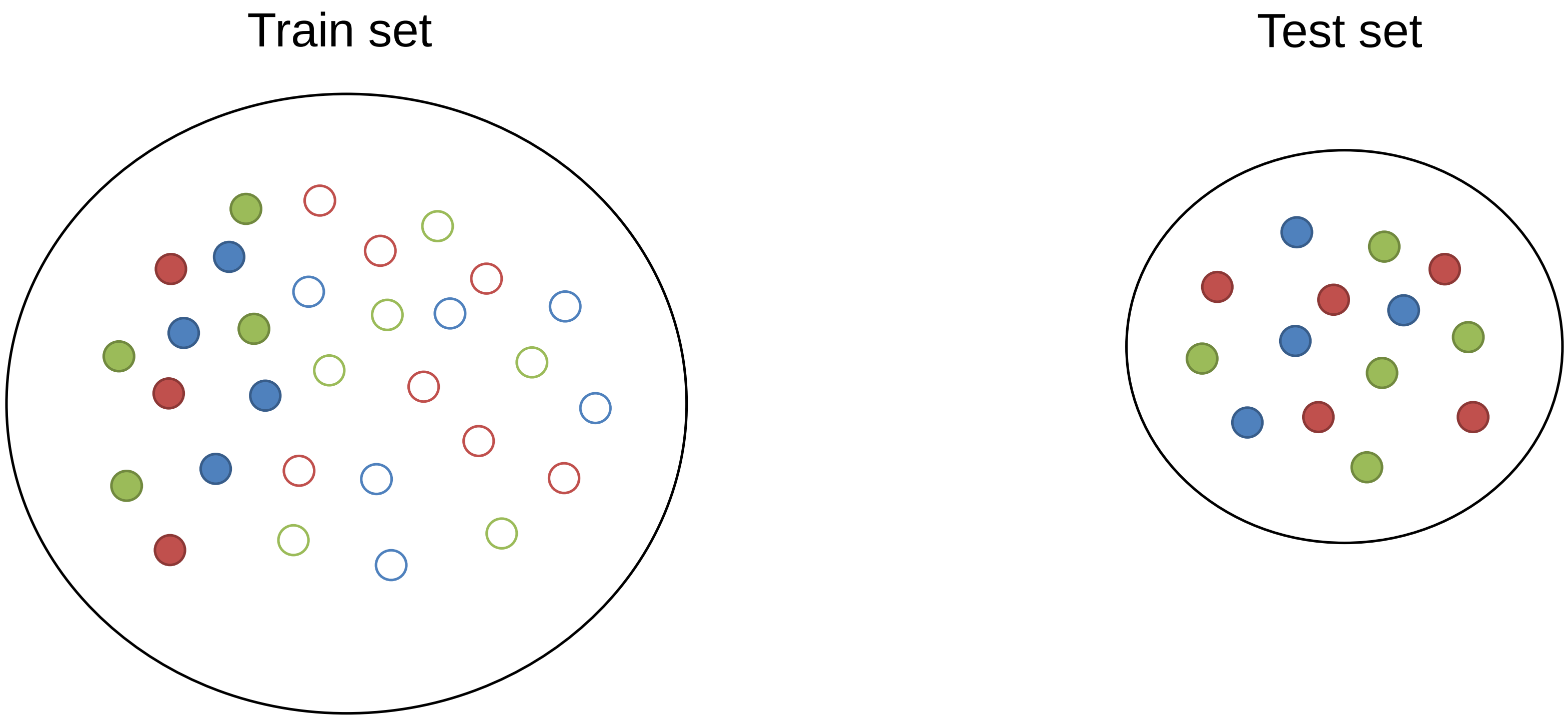
- Between unsupervised and supervised
- Learning with labeled and unlabeled data
  - Labeled instances are difficult and expensive to obtain
  - Unlabeled data may be easy to collect
- How it works?
  - Similar distributions across labeled and unlabeled instances

Topic distributions for classes

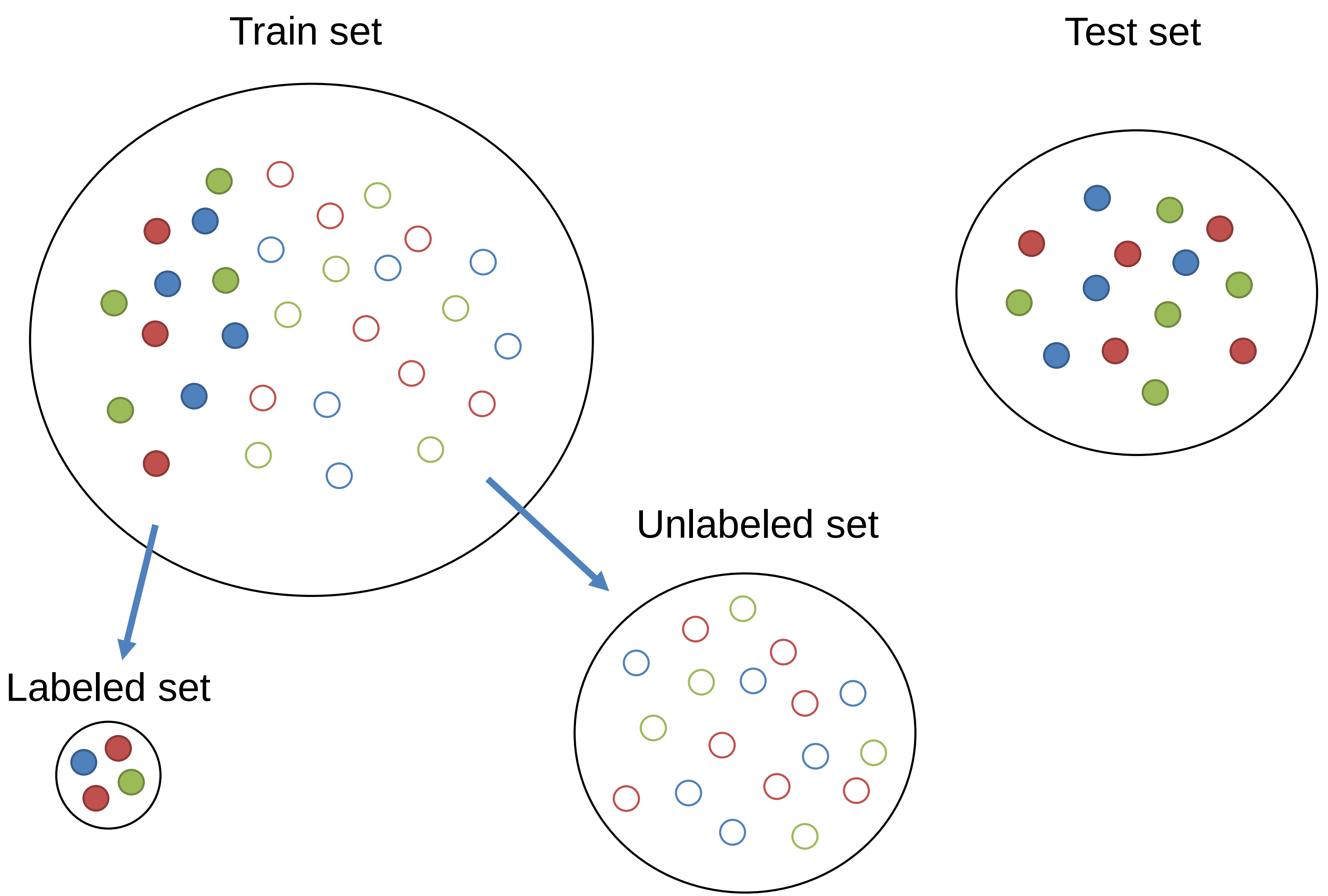
Self-training

- Semi-supervised algorithm
- Learning process uses its own predictions to teach itself
- Repeat:
  1. Train *f* on labeled data set
  2. Use *f* to predict labels for unlabeled data
  3. Unlabeled instances with most confident predictions are added to labeled data set

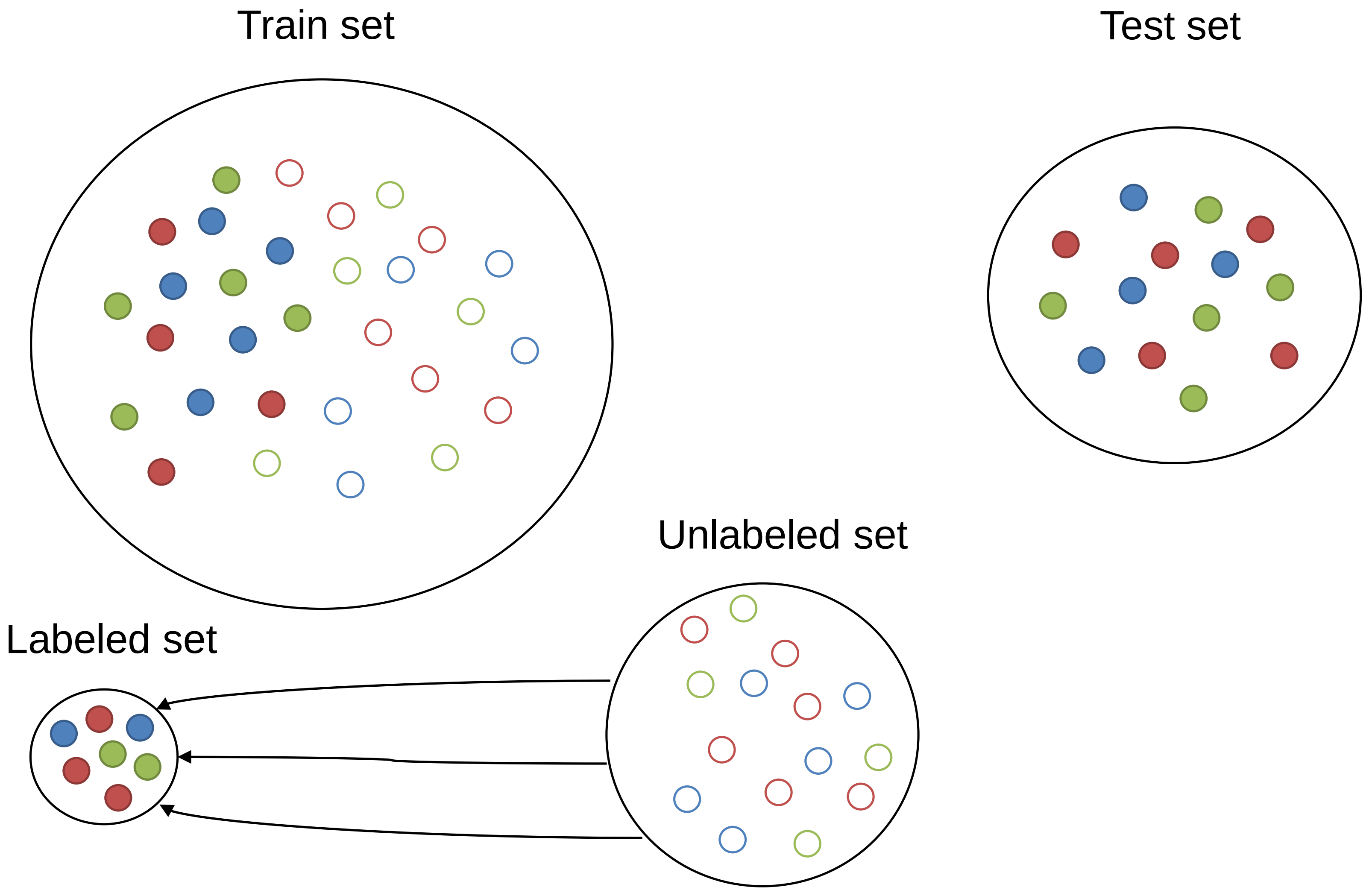
Self-training with topics



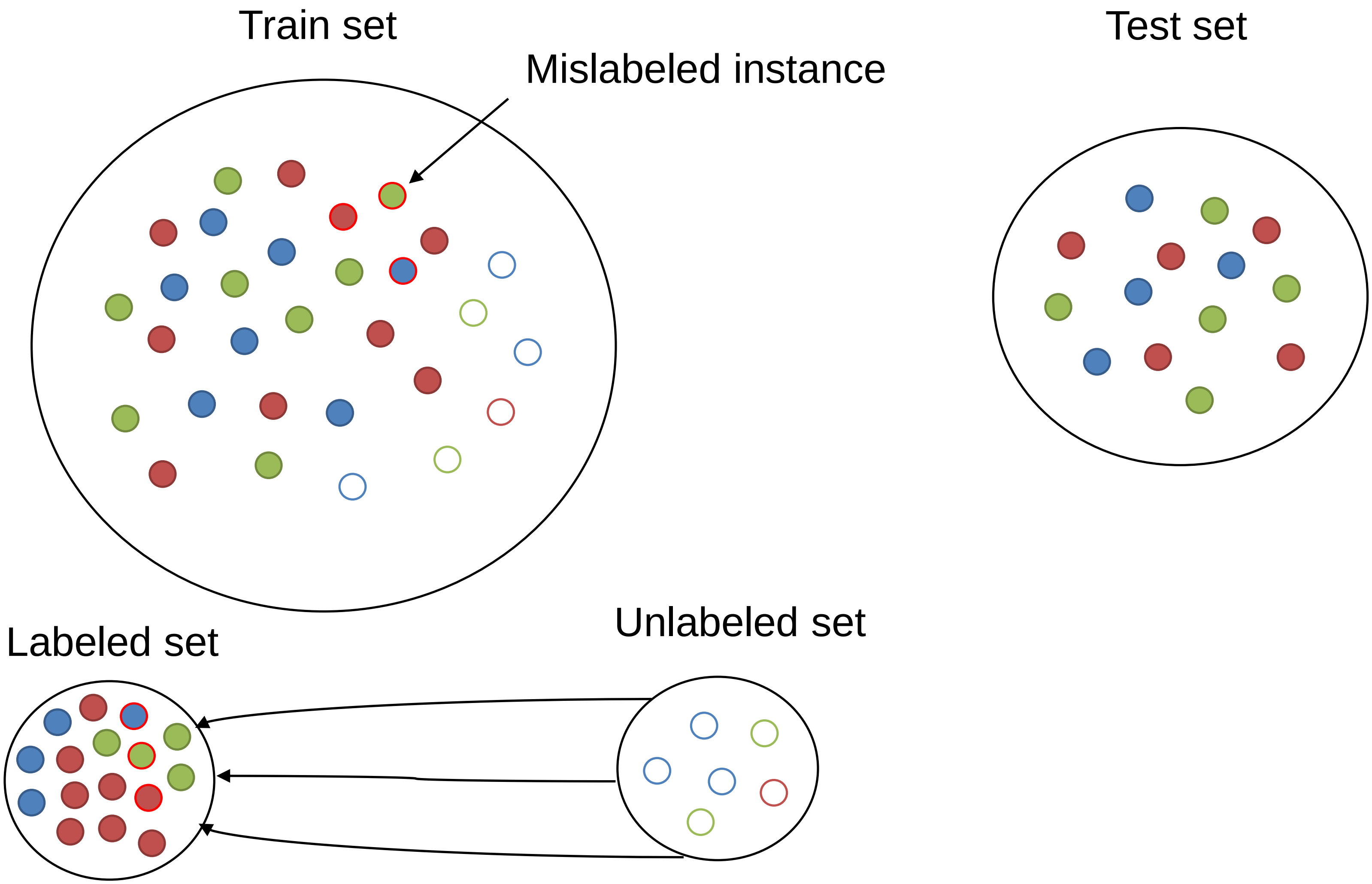
Self-training with topics



Self-training with topics



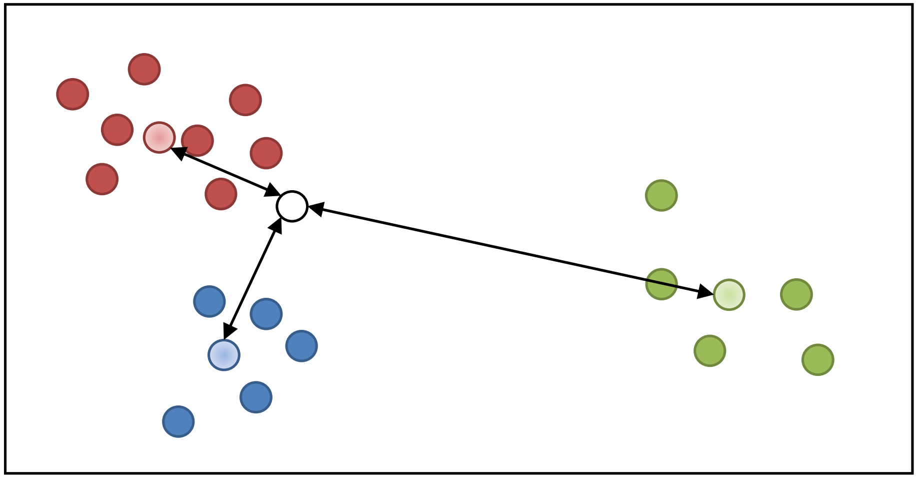
Self-training with topics



Self-training with topics

Centroids

T1	T2	T3	…	Tn
0.34	0.01	0.04	…	0.13
0.01	0.46	0.11	…	0.01
0.14	0.04	0.42	…	0.21



Topic distributions for unlabeled instances

T1	T2	T3	…	Tn
0.07	0.01	0.04	…	0.11
0.01	0.32	0.15	…	0.33
0.82	0.02	0.02	…	0.12
0.37	0.14	0.04	…	0.19
0.17	0.42	0.01	…	0.03
0.01	0.07	0.41	…	0.11
0.11	0.09	0.23	…	0.17
…	…	…	…	…

Proposed model

1. Generate topic model on entire train set
2. Split train set to labeled and unlabeled instances
3. While similar instances exists
   a. Take topic distributions from labeled set and calculate centroids for each class
   b. Measure distance between unlabeled instances and centroids (cosine similarity)
   c. Label most similar instances from unlabeled set with class from closest centroid and move instances to labeled set

Results

RTV SLO (NBMN: 81.8, SVM: 84.95)

Labeled instances	Baseline result		Simple self-training		Our method	
	NBMN	SVM	NBMN	SVM	NBMN	SVM
1%	58.3	36.64	67.59	66.69	71.80	72.83*
5%	73.74	74.8	75.62	76.05*	70.84	72.93
10%	76.5	77.65	76.4	78.16*	72.93	76.44

20 Newsgroups (NBMN: 79.77, SVM: 75.6)

Labeled instances	Baseline result		Simple self-training		Our method	
	NBMN	SVM	NBMN	SVM	NBMN	SVM
1%	32.26	24.88	45.15	39.54	70.59*	61.63
5%	59.09	43.36	65.27	57.06	71.48*	62.49
10%	67.06	56.68	71.3	68.61	72.44*	63.2

Results

Reuters R8 (NBMN: 90.64, SVM: 96.09)

Labeled instances	Baseline result		Simple self-training		Our method	
	NBMN	SVM	NBMN	SVM	NBMN	SVM
1%	84.11*	81.32	75.92	-	80.76	82.32
5%	88.54	90.13*	83.7	-	76.43	76.76
10%	90.96	93.76*	85.56	-	80.76	84.58

Google snippets – short texts (NBMN: 81.36, SVM: 64.61)

Labeled instances	Baseline result		Simple self-training		Our method	
	NBMN	SVM	NBMN	SVM	NBMN	SVM
1%	36.27	29.65	77.89*	-	64,04	55,7
5%	61.09	57.15	80.79*	-	68.55	57.37
10%	71.71	63.16	82.02*	-	75.09	67.99

Conclusions

- Representation with topics yields good results in semi-supervised settings with few labeled instances

- In some cases our approach outperforms other methods

- Future work
  - Improve our model with testing different parameters
  - Implement model on real example
  - Include texts from third corpora (e.g. Wikipedia)

Thank you for your attention!

QUESTIONS? COMMENTS?

(miha.pavlinek@um.si)