Deep Learning and Language Adaptation

Serge Sharoff

Centre for Translation Studies
University of Leeds

1 June 2017

Outline

1. **Similarity between languages**
   - Variety of languages
   - Limitations of resources
2. **Task: MT quality estimation**
   - Language Adaptation via autoencoders
   - Experimental results
3. **Task: Detection of cognates**
   - Definition of cognates
   - Cross-lingual word embeddings
4. **Tasks: POS and NER tagging**
   - Relexicalisation via cognates
   - Neural Networks for Named-Entity Recognition
   - NER shared task at BSNLP'17

Variety of languages

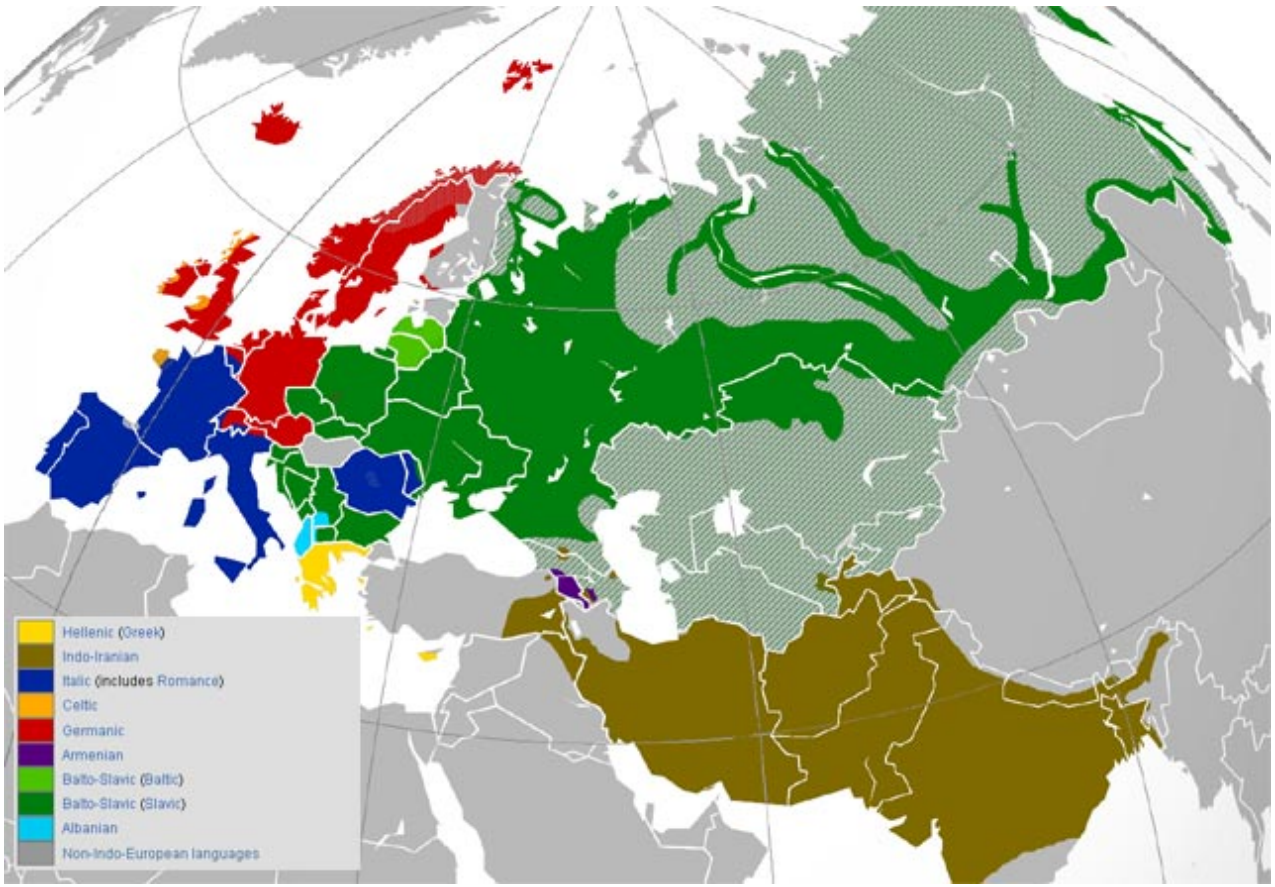
100 languages needed to cover 85% world's population

99-100. Balochi and Konkani, ≈8M native speakers each

Variety of languages

100 languages needed to cover 85% world's population

99-100. Balochi and Konkani, ≈8M native speakers each



Universal grammar claims

Francis Bacon, (c1250)

Universal grammar claims

Francis Bacon, (c1250)

Grammatica una et eadem est secundum substanciam in omnibus linguis, licet accidentaliter varietur.

Universal grammar claims

Francis Bacon, (c1250)

Grammatica una et eadem est secundum substanciam in omnibus linguis, licet accidentaliter varietur.

Grammar is one and the same in its substance in all languages, even if it accidentally varies.

Universal grammar claims

Francis Bacon, (c1250)

Grammatica una et eadem est secundum substanciam in omnibus linguis, licet accidentaliter varietur.

Grammar is one and the same in its substance in all languages, even if it accidentally varies.

Joakim Nivre, (c2015)

Universal grammar claims

Francis Bacon, (c1250)

Grammatica una et eadem est secundum substanciam in omnibus linguis, licet accidentaliter varietur.

Grammar is one and the same in its substance in all languages, even if it accidentally varies.

Joakim Nivre, (c2015)

Grammar is the same in its substance in all languages, even if the annotation accidentally varies.

Universal grammar claims

Francis Bacon, (c1250)

Grammatica una et eadem est secundum substanciam in omnibus linguis, licet accidentaliter varietur.

Grammar is one and the same in its substance in all languages, even if it accidentally varies.

Joakim Nivre, (c2015)

Grammar is the same in its substance in all languages, even if the annotation accidentally varies.

→ UD with shared annotation for 47 languages

Universal grammar claims

Francis Bacon, (c1250)

Grammatica una et eadem est secundum substanciam in omnibus linguis, licet accidentaliter varietur.

Grammar is one and the same in its substance in all languages, even if it accidentally varies.

Joakim Nivre, (c2015)

Grammar is the same in its substance in all languages, even if the annotation accidentally varies.

→ UD with shared annotation for 47 languages

- Balochi and Konkani not covered yet

Universal grammar claims

Francis Bacon, (c1250)

Grammatica una et eadem est secundum substanciam in omnibus linguis, licet accidentaliter varietur.

Grammar is one and the same in its substance in all languages, even if it accidentally varies.

Joakim Nivre, (c2015)

Grammar is the same in its substance in all languages, even if the annotation accidentally varies.

→ UD with shared annotation for 47 languages

- Balochi and Konkani not covered yet

BUT Farsi and Hindi are

Languages	UD	Wiki	MT
Germanic			
Danish	89K	52M	
Dutch	286K	223M	
German	274K	783M	578K
Norwegian	245K	89M	
Swedish	131K	127M	
Romance			
Catalan	442K	181M	
French	367K	667M	432K
Italian	266K	433M	329K
Portuguese	454K	222M	321K
Romanian	109K	63M	
Spanish	853K	530M	265K
Slavonic			
Bulgarian	124K	55M	
Czech	1671K	110M	183K
Polish	70K	227M	213K
Russian	928K	420M	266K
Slovenian	136K	321M	
Ukrainian	10K	161M	

Languages	UD	Wiki	MT
Germanic			
Danish	89K	52M	
Dutch	286K	223M	
German	274K	783M	578K
Norwegian	245K	89M	
Swedish	131K	127M	
Romance			
Catalan	442K	181M	
French	367K	667M	432K
Italian	266K	433M	329K
Portuguese	454K	222M	321K
Romanian	109K	63M	
Spanish	853K	530M	265K
Slavonic			
Bulgarian	124K	55M	
Czech	1671K	110M	183K
Polish	70K	227M	213K
Russian	928K	420M	266K
Slovenian	136K	321M	
Ukrainian	10K	161M	

Languages	UD	Wiki	MT
Germanic			
Danish	89K	52M	
Dutch	286K	223M	
German	274K	783M	578K
Norwegian	245K	89M	
Swedish	131K	127M	
Romance			
Catalan	442K	181M	
French	367K	667M	432K
Italian	266K	433M	329K
Portuguese	454K	222M	321K
Romanian	109K	63M	
Spanish	853K	530M	265K
Slavonic			
Bulgarian	124K	55M	
Czech	1671K	110M	183K
Polish	70K	227M	213K
Russian	928K	420M	266K
Slovenian	136K	321M	
Ukrainian	10K	161M	

- Large number of rare events:
  $p(break) = 0.00018308$
  $p(waves+break) = 0.00000005$

Languages	UD	Wiki	MT
Germanic			
Danish	89K	52M	
Dutch	286K	223M	
German	274K	783M	578K
Norwegian	245K	89M	
Swedish	131K	127M	
Romance			
Catalan	442K	181M	
French	367K	667M	432K
Italian	266K	433M	329K
Portuguese	454K	222M	321K
Romanian	109K	63M	
Spanish	853K	530M	265K
Slavonic			
Bulgarian	124K	55M	
Czech	1671K	110M	183K
Polish	70K	227M	213K
Russian	928K	420M	266K
Slovenian	136K	321M	
Ukrainian	10K	161M	

- Large number of rare events:
  $p(break) = 0.00018308$
  $p(waves{+}break) = 0.00000005$

  waves break $\rightarrow$ Fr: *vagues*
  fracassent, se cassent,
  se brisent, déferlent...
  the area where the waves break
  $\rightarrow$ *la zone de déferlement*

Languages	UD	Wiki	MT
Germanic			
Danish	89K	52M	
Dutch	286K	223M	
German	274K	783M	578K
Norwegian	245K	89M	
Swedish	131K	127M	
Romance			
Catalan	442K	181M	
French	367K	667M	432K
Italian	266K	433M	329K
Portuguese	454K	222M	321K
Romanian	109K	63M	
Spanish	853K	530M	265K
Slavonic			
Bulgarian	124K	55M	
Czech	1671K	110M	183K
Polish	70K	227M	213K
Russian	928K	420M	266K
Slovenian	136K	321M	
Ukrainian	10K	161M	

- Large number of rare events:
  $p(break) = 0.00018308$
  $p(waves{+}break) = 0.00000005$

  waves break $\rightarrow$ Fr: *vagues*
  fracassent, se cassent,
  se brisent, déferlent...
  the area where the waves break
  $\rightarrow$ *la zone de déferlement*

- Tagsets are sparse:
  685 uk vs 710 ru

Languages	UD	Wiki	MT
Germanic			
Danish	89K	52M	
Dutch	286K	223M	
German	274K	783M	578K
Norwegian	245K	89M	
Swedish	131K	127M	
Romance			
Catalan	442K	181M	
French	367K	667M	432K
Italian	266K	433M	329K
Portuguese	454K	222M	321K
Romanian	109K	63M	
Spanish	853K	530M	265K
Slavonic			
Bulgarian	124K	55M	
Czech	1671K	110M	183K
Polish	70K	227M	213K
Russian	928K	420M	266K
Slovenian	136K	321M	
Ukrainian	10K	161M	

- Large number of rare events:
  $p(break) = 0.00018308$
  $p(waves{+}break) = 0.00000005$

  waves break → Fr: *vagues*
  fracassent, se cassent,
  se brisent, déferlent...
  the area where the waves break
  → *la zone de déferlement*

- Tagsets are sparse:
  685 uk vs 710 ru

  Still 45 single examples of feature compbinations in ru (Syntagrus):
  колотыми
  V,Asp=Imp,Case=Ins,Num=Plur,
  Tense=Past, Voice=Pass

My story about related languages

- Automatic generation of instructions (Bateman, et al 2000)
  Rule-based grammar for Bulgarian, Czech and Russian
  Используйте команду Multiline, чтобы соединить две точки.
  'Use$_{imp,pl}$ the Multiline command to connect two points$_{gen,sg}$'

My story about related languages

- Automatic generation of instructions (Bateman, et al 2000)
  Rule-based grammar for Bulgarian, Czech and Russian
  Используйте команду Multiline, чтобы соединить две точки.
  'Use$_{imp,pl}$ the Multiline command to connect two points$_{gen,sg}$'
- Language resources for reading and translation skills
  Romanian via French (Ciobanu, et al 2006);
  Ukrainian via Russian (Kurella, et al 2008)

My story about related languages

- Automatic generation of instructions (Bateman, et al 2000)
  Rule-based grammar for Bulgarian, Czech and Russian
  Используйте команду Multiline, чтобы соединить две точки.
  'Use$_{imp,pl}$ the Multiline command to connect two points$_{gen,sg}$'
- Language resources for reading and translation skills
  Romanian via French (Ciobanu, et al 2006);
  Ukrainian via Russian (Kurella, et al 2008)
- MT via related pivot languages (Babych, et al, 2007)
  uk→ru→de/en is far better than uk→de/en

My story about related languages

- Automatic generation of instructions (Bateman, et al 2000)
  Rule-based grammar for Bulgarian, Czech and Russian
  Используйте команду Multiline, чтобы соединить две точки.
  'Use$_{imp,pl}$ the Multiline command to connect two points$_{gen,sg}$'
- Language resources for reading and translation skills
  Romanian via French (Ciobanu, et al 2006);
  Ukrainian via Russian (Kurella, et al 2008)
- MT via related pivot languages (Babych, et al, 2007)
  uk→ru→de/en is far better than uk→de/en
- Chris Brew on POS tagging for related languages
  Catalan via Spanish; Russian via Czech (Feldman, et al, 2006)

My story about related languages

- Automatic generation of instructions (Bateman, et al 2000)
  Rule-based grammar for Bulgarian, Czech and Russian
  Используйте команду Multiline, чтобы соединить две точки.
  'Use$_{imp,pl}$ the Multiline command to connect two points$_{gen,sg}$'
- Language resources for reading and translation skills
  Romanian via French (Ciobanu, et al 2006);
  Ukrainian via Russian (Kurella, et al 2008)
- MT via related pivot languages (Babych, et al, 2007)
  uk→ru→de/en is far better than uk→de/en
- Chris Brew on POS tagging for related languages
  Catalan via Spanish; Russian via Czech (Feldman, et al, 2006)
- POS taggers for Kannada via Telugu (Reddy, Sharoff, 2011);
  and for Ukrainian via Russian (Babych, Sharoff, 2016)

My story about related languages

- Automatic generation of instructions (Bateman, et al 2000)
  Rule-based grammar for Bulgarian, Czech and Russian
  Используйте команду Multiline, чтобы соединить две точки.
  'Use$_{imp,pl}$ the Multiline command to connect two points$_{gen,sg}$'
- Language resources for reading and translation skills
  Romanian via French (Ciobanu, et al 2006);
  Ukrainian via Russian (Kurella, et al 2008)
- MT via related pivot languages (Babych, et al, 2007)
  uk→ru→de/en is far better than uk→de/en
- Chris Brew on POS tagging for related languages
  Catalan via Spanish; Russian via Czech (Feldman, et al, 2006)
- POS taggers for Kannada via Telugu (Reddy, Sharoff, 2011);
  and for Ukrainian via Russian (Babych, Sharoff, 2016)

KEY: Common **representation** for related languages

Outline

1. Similarity between languages
   - Variety of languages
   - Limitations of resources
2. Task: MT quality estimation
   - Language Adaptation via autoencoders
   - Experimental results
3. Task: Detection of cognates
   - Definition of cognates
   - Cross-lingual word embeddings
4. Tasks: POS and NER tagging
   - Relexicalisation via cognates
   - Neural Networks for Named-Entity Recognition
   - NER shared task at BSNLP'17

Translation Quality Estimation (QE)

- Quality Estimation measures confidence in MT output without references, e.g., by predicting Post-Editing (Specia, et al 2013)

Translation Quality Estimation (QE)

- Quality Estimation measures confidence in MT output without references, e.g., by predicting Post-Editing (Specia, et al 2013)

Complexity Indicators Features related to difficulty in translating Source Text (ST), e.g., ST segment length, its language model and phrase table size.

Translation Quality Estimation (QE)

- Quality Estimation measures confidence in MT output without references, e.g., by predicting Post-Editing (Specia, et al 2013)

Complexity Indicators Features related to difficulty in translating Source Text (ST), e.g., ST segment length, its language model and phrase table size.

Fluency Indicators Features related to how fluent MT output is, e.g., its language model.

Translation Quality Estimation (QE)

- Quality Estimation measures confidence in MT output without references, e.g., by predicting Post-Editing (Specia, et al 2013)

Complexity Indicators Features related to difficulty in translating Source Text (ST), e.g., ST segment length, its language model and phrase table size.

Fluency Indicators Features related to how fluent MT output is, e.g., its language model.

Adequacy Indicators Features related to how much meaning is preserved in MT output, e.g., translation model ratios, *semantic similarity via bilingual embeddings*.

Translation Quality Estimation (QE)

- Quality Estimation measures confidence in MT output without references, e.g., by predicting Post-Editing (Specia, et al 2013)

Complexity Indicators Features related to difficulty in translating Source Text (ST), e.g., ST segment length, its language model and phrase table size.

Fluency Indicators Features related to how fluent MT output is, e.g., its language model.

Adequacy Indicators Features related to how much meaning is preserved in MT output, e.g., translation model ratios, *semantic similarity via bilingual embeddings*.

BUT Few languages available for training (Cs, Pl, Ru in Autodesk)

Transferring classifiers (Rios, Sharoff, 2016)

- We know which Russian MT output is good

Transferring classifiers (Rios, Sharoff, 2016)

- We know which Russian MT output is good
- Polish MT output with **similar** features is likely to be good

Transferring classifiers (Rios, Sharoff, 2016)

- We know which Russian MT output is good
- Polish MT output with **similar** features is likely to be good

BUT Polish feature space is different
LM values, phrase table size, translation probabilities

Transferring classifiers (Rios, Sharoff, 2016)

- We know which Russian MT output is good
- Polish MT output with **similar** features is likely to be good

BUT Polish feature space is different
LM values, phrase table size, translation probabilities

Self-Taught Learning (STL) for adapting *feature spaces*

Transferring classifiers (Rios, Sharoff, 2016)

- We know which Russian MT output is good
- Polish MT output with **similar** features is likely to be good

BUT Polish feature space is different
LM values, phrase table size, translation probabilities

Self-Taught Learning (STL) for adapting *feature spaces*

1. Build a function for transforming data using *unlabelled* Russian and Polish data (MT without PE)

Transferring classifiers (Rios, Sharoff, 2016)

- We know which Russian MT output is good
- Polish MT output with **similar** features is likely to be good

BUT Polish feature space is different
LM values, phrase table size, translation probabilities

Self-Taught Learning (STL) for adapting *feature spaces*

1. Build a function for transforming data using *unlabelled* Russian and Polish data (MT without PE)

   Autoencoders: non-linear learnable dimensionality reduction

Transferring classifiers (Rios, Sharoff, 2016)

- We know which Russian MT output is good
- Polish MT output with **similar** features is likely to be good

BUT Polish feature space is different
LM values, phrase table size, translation probabilities

Self-Taught Learning (STL) for adapting *feature spaces*

1. Build a function for transforming data using *unlabelled* Russian and Polish data (MT without PE)

   Autoencoders: non-linear learnable dimensionality reduction

   Autoencoder vs PCA vs MDS, SOM or t-SNE

Transferring classifiers (Rios, Sharoff, 2016)

- We know which Russian MT output is good
- Polish MT output with **similar** features is likely to be good

BUT Polish feature space is different
LM values, phrase table size, translation probabilities

Self-Taught Learning (STL) for adapting *feature spaces*

1. Build a function for transforming data using *unlabelled* Russian and Polish data (MT without PE)

   Autoencoders: non-linear learnable dimensionality reduction

   Autoencoder vs PCA vs MDS, SOM or t-SNE
2. Train a prediction model on transformed Russian data

Transferring classifiers (Rios, Sharoff, 2016)

- We know which Russian MT output is good
- Polish MT output with **similar** features is likely to be good

BUT Polish feature space is different
LM values, phrase table size, translation probabilities

Self-Taught Learning (STL) for adapting *feature spaces*

1. Build a function for transforming data using *unlabelled* Russian and Polish data (MT without PE)

   Autoencoders: non-linear learnable dimensionality reduction

   Autoencoder vs PCA vs MDS, SOM or t-SNE
2. Train a prediction model on transformed Russian data
3. Apply the model to transformed Polish data

Experimental results

- Baseline prediction of PE effort:

Upper baseline (**es**)	MAE	0.14
	RSME	0.18
	Correlation	**0.53**

Experimental results

- Baseline prediction of PE effort:

Upper baseline (**es**)	MAE	0.14
	RSME	0.18
	Correlation	**0.53**

- Our Language Adaptation method:

es	→	pt	it	fr
STL	MAE	0.14	0.16	0.17
	RMSE	0.17	0.21	0.22
	Correlation	**0.52**	**0.40**	**0.30**
Baseline Train: es Test: pt/it/fr	MAE	0.16	0.18	0.18
	RMSE	0.20	0.23	0.23
	Correlation	*0.35*	*0.26*	*0.24*

Power of language adaptation

Upper baseline (ru)	MAE	0.18
	RSME	0.27
	Correlation	**0.47**

Power of language adaptation

Upper baseline (ru)	MAE	0.18
	RSME	0.27
	Correlation	**0.47**

en-ru	→	en-cs	en-pl
STL	MAE	0.19	0.19
	RMSE	0.25	0.25
	Correlation	**0.41**	**0.46**
Baseline Train: ru Test: cs/pl	MAE	0.20	0.21
	RMSE	0.26	0.27
	Correlation	*0.32*	*0.33*

Power of language adaptation

Upper baseline (ru)	MAE	0.18
	RSME	0.27
	Correlation	**0.47**

en-ru	→	en-cs	en-pl
STL	MAE	0.19	0.19
	RMSE	0.25	0.25
	Correlation	**0.41**	**0.46**
Baseline Train: ru Test: cs/pl	MAE	0.20	0.21
	RMSE	0.26	0.27
	Correlation	*0.32*	*0.33*

en-es	→	en-cs	en-pl
STL	MAE	0.22	0.25
	RMSE	0.29	0.32
	Correlation	*0.08*	*0.11*
Baseline Train: es Test: cs/pl	MAE	0.23	0.22
	RSME	0.31	0.29
	Correlation	*0.11*	*0.09*

Outline

1. Similarity between languages
   - Variety of languages
   - Limitations of resources
2. Task: MT quality estimation
   - Language Adaptation via autoencoders
   - Experimental results
3. Task: Detection of cognates
   - Definition of cognates
   - Cross-lingual word embeddings
4. Tasks: POS and NER tagging
   - Relexicalisation via cognates
   - Neural Networks for Named-Entity Recognition
   - NER shared task at BSNLP'17

Definition of cognates

- **Common etymology** *držati, držet, держать, держати*
  or **borrowing** *computer→компьютер, комп'ютер*
  počítač, računalnik

Definition of cognates

- **Common etymology** *držati, držet, держать, держати* or **borrowing** *computer→компьютер, комп'ютер počítač, računalnik*
- **False friends** *вредный* 'harmful' vs *vreden* 'worthy'

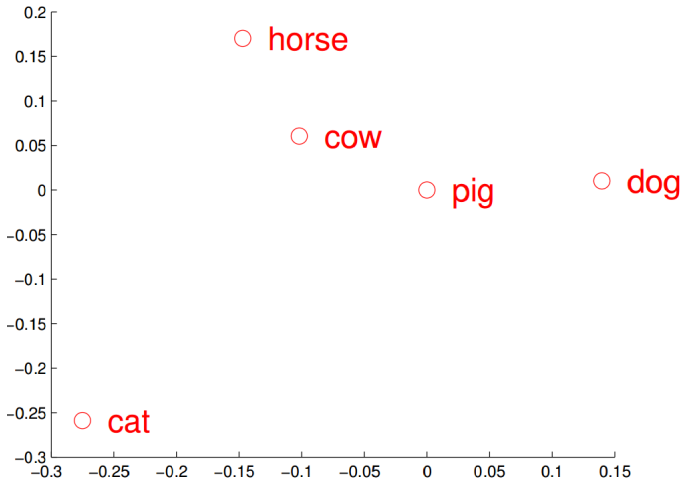
Definition of cognates

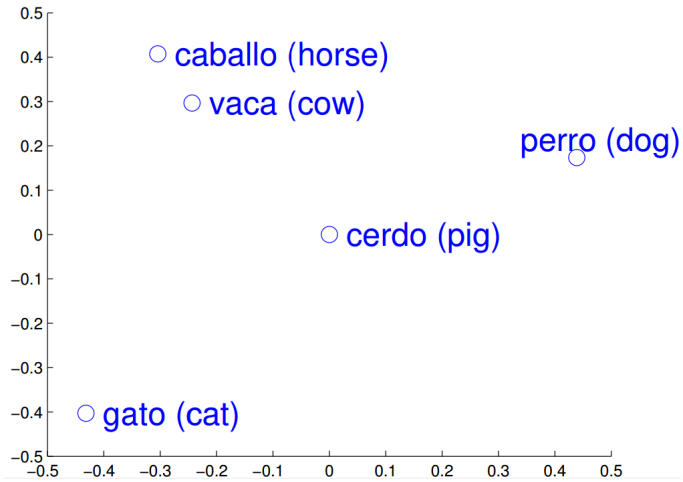
- **Common etymology** *držati, držet, держать, держати* or **borrowing** *computer→компьютер, комп'ютер počítač, računalnik*
- **False friends** *вредный* 'harmful' vs *vreden* 'worthy'
- **Partial friends** *жена* (in Russian 'wife') vs *žena* (in Slovenian: 'wife' OR 'woman')

Definition of cognates

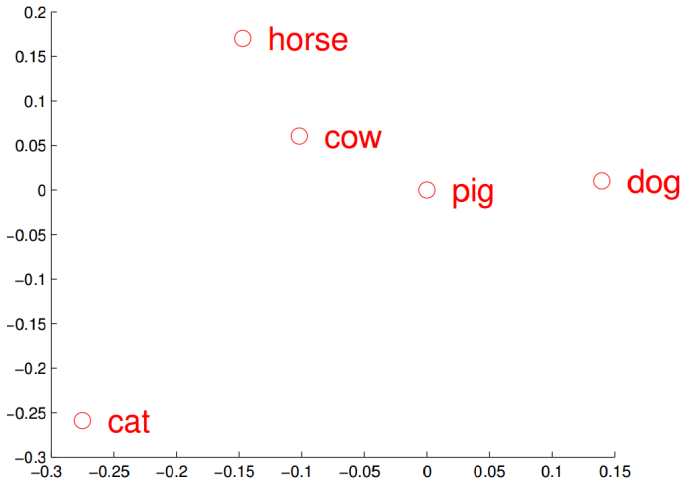
- **Common etymology** *držati, držet, держать, держати* or **borrowing** *computer→компьютер, комп'ютер počítač, računalnik*
- **False friends** *вредный* 'harmful' vs *vreden* 'worthy'
- **Partial friends** *жена* (in Russian 'wife') vs *žena* (in Slovenian: 'wife' OR 'woman')
- **Frequency difference** *debuxo* vs *dibujo* 'a drawing' (Portuguese and Spanish), ranks 100,000 vs 2,000 approx (*desenho* is more common in Portuguese)

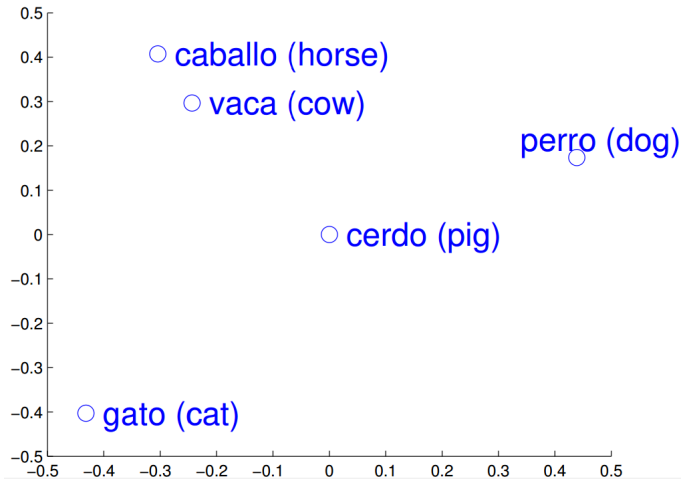
Cross-lingual word embeddings (Mikolov, 2013)





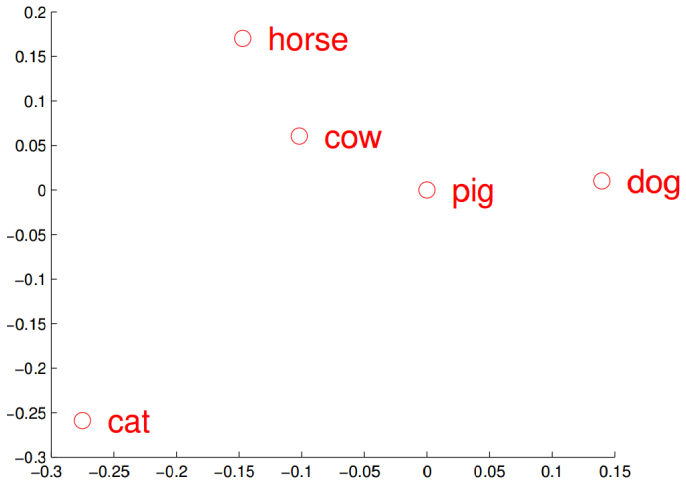
Cross-lingual word embeddings (Mikolov, 2013)

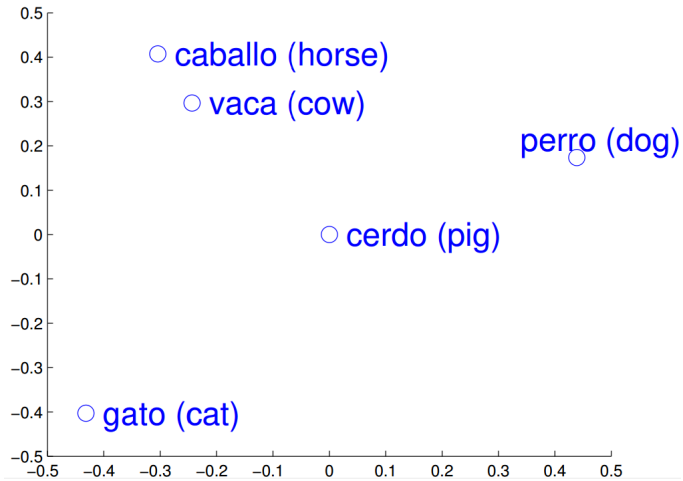




- Monolingual corpora for word embeddings (Wikipedias)

Cross-lingual word embeddings (Mikolov, 2013)

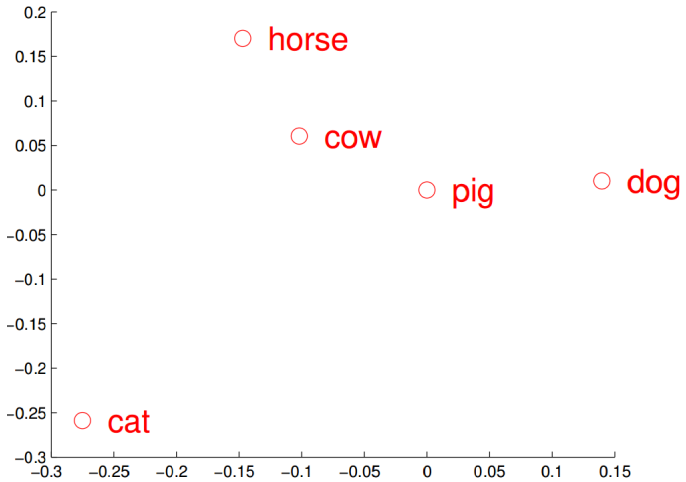


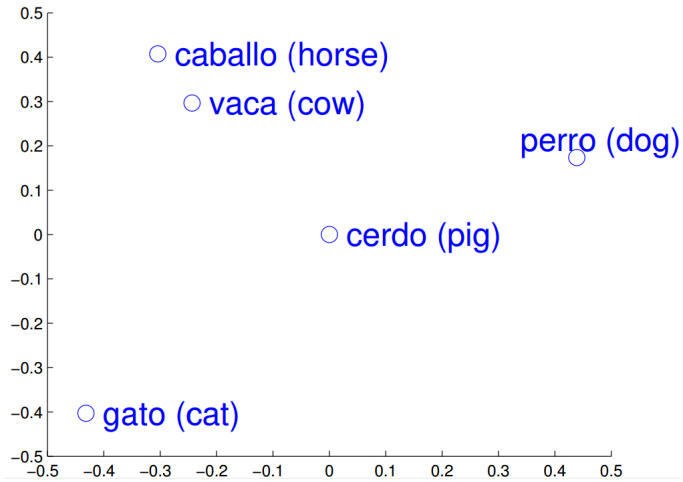


- Monolingual corpora for word embeddings (Wikipedias)
- Linear transformation or MLP for mapping embeddings

$$\min_{\mathbf{W}} \sum ||\mathbf{W}e_i - f_i||^2$$

Cross-lingual word embeddings (Mikolov, 2013)





- Monolingual corpora for word embeddings (Wikipedias)
- Linear transformation or MLP for mapping embeddings

$$\min_{\mathbf{W}} \sum ||\mathbf{W}e_i - f_i||^2$$

- Small (1-2kW) bilingual dictionaries from iWiki links:
  - (sv) Slaget om Filippinen
  - (nl) Lijst van Poolse schrijvers
  - (pl) Z życia marionetek
  - (pl) Wskaźnik jakości życia
  - (de) Schlacht um die Philippinen
  - (de) Liste polnischer Schriftsteller
  - (ru) Из жизни марионеток
  - (ru) Индекс качества жизни

Levenshtein distances

- Baseline Levenshtein distance (LD):
  Philippinen $\rightarrow$ *Filippinen* : 1 deletion, 1 substitution
  Schlacht $\rightarrow$ *Slaget* : 3 deletions, 1 substitution

Levenshtein distances

- Baseline Levenshtein distance (LD):
  Philippinen $\rightarrow$ *Filippinen* : 1 deletion, 1 substitution
  Schlacht $\rightarrow$ *Slaget* : 3 deletions, 1 substitution
- Weighted Levenshtein Distance (WLD) for cognates

Sch l a ch t
S l a ge t

Levenshtein distances

- Baseline Levenshtein distance (LD):
  Philippinen $\rightarrow$ *Filippinen* : 1 deletion, 1 substitution
  Schlacht $\rightarrow$ *Slaget* : 3 deletions, 1 substitution
- Weighted Levenshtein Distance (WLD) for cognates

Sch l a ch t
S l a ge t

$$p(sch \rightarrow s) = 0.7; p(e \rightarrow o) = 0.5$$

Levenshtein distances

- Baseline Levenshtein distance (LD):
  Philippinen $\rightarrow$ *Filippinen* : 1 deletion, 1 substitution
  Schlacht $\rightarrow$ *Slaget* : 3 deletions, 1 substitution
- Weighted Levenshtein Distance (WLD) for cognates

Sch l a ch t
S l a ge t

$$p(sch \rightarrow s) = 0.7; p(e \rightarrow o) = 0.5$$

$$WLD = \frac{\sum_{(e,f) \in al(s_e, s_f)} (1 - p(f|e))}{\max(len(s_e), len(s_f))}$$

Levenshtein distances

- Baseline Levenshtein distance (LD):
  Philippinen $\rightarrow$ *Filippinen* : 1 deletion, 1 substitution
  Schlacht $\rightarrow$ *Slaget* : 3 deletions, 1 substitution
- Weighted Levenshtein Distance (WLD) for cognates

  Sch l a ch t
  S l a ge t

$$p(sch \rightarrow s) = 0.7; p(e \rightarrow o) = 0.5$$

$$WLD = \frac{\sum_{(e,f) \in al(s_e, s_f)} (1 - p(f|e))}{\max(len(s_e), len(s_f))}$$

- Also WLD works across charsets:
  m a r i o n e t e k ż y * c ia
  м а р и о н е т о к ж и з н и

Distance concentration and hubness

- More items at the same distance when d grows

Distance concentration and hubness

- More items at the same distance when d grows
- More hubby items as d grows $N_k(\mathbf{x})$ number of times $\mathbf{x}$ is among k nearest neighbours

Distance concentration and hubness

- More items at the same distance when d grows
- More hubby items as d grows $N_k(\mathbf{x})$ number of times $\mathbf{x}$ is among k nearest neighbours

#	**Standard**	#	**With GC**
77	bennett	10	arranged
75	curtis	10	corresponded
59	featuring	10	exists
58	laurie	10	learnt
56	james	10	milanese
56	miller	10	represents
55	elliot	10	traceable
54	gavin	9	association
51	convinced	9	chosen
47	keith	9	coincides
46	barker	9	consist
46	titled	9	forwarded
45	persuaded	9	grade

Evaluation of cognate detection for en-it

Vectors from (Dinu, et al. 2014)	
TM as in Mikolov et al. (2013b)	0.349
CCA as in Faruqui and Dyer (2014)	0.378
Orth as in Artetxe et al. (2016)	0.393
GC as in Dinu et al. (2014)	0.377

Evaluation of cognate detection for en-it

Vectors from (Dinu, et al. 2014)	
TM as in Mikolov et al. (2013b)	0.349
CCA as in Faruqui and Dyer (2014)	0.378
Orth as in Artetxe et al. (2016)	0.393
GC as in Dinu et al. (2014)	0.377

Evaluation of cognate detection for en-it

Vectors from (Dinu, et al. 2014)	
TM as in Mikolov et al. (2013b)	0.349
CCA as in Faruqui and Dyer (2014)	0.378
Orth as in Artetxe et al. (2016)	0.393
GC as in Dinu et al. (2014)	0.377
GC+LD	0.501
GC+WLD	**0.531**

Evaluation of cognate detection for en-it

Vectors from (Dinu, et al. 2014)	
TM as in Mikolov et al. (2013b)	0.349
CCA as in Faruqui and Dyer (2014)	0.378
Orth as in Artetxe et al. (2016)	0.393
GC as in Dinu et al. (2014)	0.377
GC+LD	0.501
GC+WLD	**0.531**
Vectors from (Bojanowski, et al. 2016)	
FT+Orth	0.529
FT+Orth+GC	0.477
FT+Orth+GC+WLD	**0.616**

Evaluation of cognate detection for en-it

Vectors from (Dinu, et al. 2014)	
TM as in Mikolov et al. (2013b)	0.349
CCA as in Faruqui and Dyer (2014)	0.378
Orth as in Artetxe et al. (2016)	0.393
GC as in Dinu et al. (2014)	0.377
GC+LD	0.501
GC+WLD	**0.531**
Vectors from (Bojanowski, et al. 2016)	
FT+Orth	0.529
FT+Orth+GC	0.477
FT+Orth+GC+WLD	**0.616**
FT+Orth (c)	0.562
FT+Orth+GC (c)	0.601
FT+Orth+GC+WLD (c)	**0.681**

Dictionaries for Slavonic languages

Best En-It (c) without WLD: 0.601
Best En-It (c) with WLD: 0.681

Dictionaries for Slavonic languages

Best En-It (c) without WLD: 0.601
Best En-It (c) with WLD: 0.681

- Dictionary induction without WLD:

	sl-hr	sl-cs	sl-pl	sl-ru	ru-uk	cs-sk
Prec@1:	0.429	0.611	0.584	0.566	0.929	0.814
Prec@10:	0.688	0.868	0.842	0.818	0.976	0.971

Dictionaries for Slavonic languages

Best En-It (c) without WLD: 0.601
Best En-It (c) with WLD: 0.681

- Dictionary induction without WLD:

	sl-hr	sl-cs	sl-pl	sl-ru	ru-uk	cs-sk
Prec@1:	0.429	0.611	0.584	0.566	0.929	0.814
Prec@10:	0.688	0.868	0.842	0.818	0.976	0.971

- Dictionary induction with WLD:

	sl-hr	sl-cs	sl-pl	sl-ru	ru-uk	cs-sk
Prec@1:	0.840	0.763	0.751	0.662	0.945	0.910
Prec@10:	0.963	0.973	0.977	0.883	0.994	0.996

Morphological structure of cognate words

- There are problems with full forms:

	French	Italian
Sing	*maladie*	*malattia*

Morphological structure of cognate words

- There are problems with full forms:

	French	Italian
Sing	*maladie*	*malattia*

Morphological structure of cognate words

- There are problems with full forms:

	French	Italian
Sing	*malad**ie***	*malattia*
Plur	*maladies*	*malatt**ie***

Morphological structure of cognate words

- There are problems with full forms:

	French	Italian
Sing	*maladie*	*malattia*
Plur	*maladies*	*malattie*

ru-uk cognates for *green*

nominative	зелёный, зелёная	зелений, зелена,
genitive	зелёного,зелёной	зеленого, зелен**ої**
dative	зелён**ому**,зелёной	зеленому, зелен**ій**
instrumental	зелёным, зелёной	зеленим, зелен**ою**
prepositional	зелён**ом**, зелёной	зеленому, зелен**ій**

Morphological structure of cognate words

- There are problems with full forms:

	French	Italian
Sing	*malad**ie***	*malattia*
Plur	*maladies*	*malatt**ie***

ru-uk cognates for *green*

nominative	зелёный, зелёная	зелений, зелена,
genitive	зелёного,зелёной	зеленого, зелен**ої**
dative	зелён**ому**,зелёной	зеленому, зелен**ій**
instrumental	зелёным, зелёной	зеленим, зелен**ою**
prepositional	зелён**ом**, зелёной	зеленому, зелен**ій**

- Stem alterations are important too:
  $Америки_{gen} \rightarrow Америки$

Morphological structure of cognate words

- There are problems with full forms:

	French	Italian
Sing	*maladie*	*malattia*
Plur	*maladies*	*malattie*

ru-uk cognates for *green*

nominative	зелёный, зелёная	зелений, зелена,
genitive	зелёного,зелёной	зеленого, зелен**ої**
dative	зелён**ому**,зелёной	зеленому, зелен**ій**
instrumental	зелёным, зелёной	зеленим, зелен**ою**
prepositional	зелён**ом**, зелёной	зеленому, зелен**ій**

- Stem alterations are important too:
  $Америки_{gen} \rightarrow Америки$
  $Америке_{dat} \rightarrow Америці$

Outline

1. Similarity between languages
   - Variety of languages
   - Limitations of resources
2. Task: MT quality estimation
   - Language Adaptation via autoencoders
   - Experimental results
3. Task: Detection of cognates
   - Definition of cognates
   - Cross-lingual word embeddings
4. Tasks: POS and NER tagging
   - Relexicalisation via cognates
   - Neural Networks for Named-Entity Recognition
   - NER shared task at BSNLP'17

Cross-lingual tagging and parsing

Het	is	heel	anders	gelopen	dan	ik	dacht
PRON	AUX	ADV	ADV	VERB	SCONJ	PRON	VERB
Es	ist	ganz	anders	gelaufen	als	ich	dachte

Cross-lingual tagging and parsing

Het	is	heel	anders	gelopen	dan	ik	dacht
PRON	AUX	ADV	ADV	VERB	SCONJ	PRON	VERB
Es	ist	ganz	anders	gelaufen	als	ich	dachte

- De-lexicalisation (Feldman, et al, 2006; Mcdonald, et al 2011; Reddy, Sharoff, 2011)

Cross-lingual tagging and parsing

Het	is	heel	anders	gelopen	dan	ik	dacht
PRON	AUX	ADV	ADV	VERB	SCONJ	PRON	VERB
Es	ist	ganz	anders	gelaufen	als	ich	dachte

- De-lexicalisation (Feldman, et al, 2006; Mcdonald, et al 2011; Reddy, Sharoff, 2011)

Это	прошло	совсем	иначе	чем	я	думал
PRON	VERB	ADV	ADV	SCONJ	PRON	VERB

Cross-lingual tagging and parsing

Het	is	heel	anders	gelopen	dan	ik	dacht
PRON	AUX	ADV	ADV	VERB	SCONJ	PRON	VERB
Es	ist	ganz	anders	gelaufen	als	ich	dachte

- De-lexicalisation (Feldman, et al, 2006; Mcdonald, et al 2011; Reddy, Sharoff, 2011)

Это	прошло	совсем	иначе	чем	я	думал
PRON	VERB	ADV	ADV	SCONJ	PRON	VERB

它	和	我	想	的	完全	不同
it	and	I	thought	of	completely	differ
PRON	CCONJ	PRON	NOUN	PART	ADV	ADJ

Cross-lingual tagging and parsing

Het	is	heel	anders	gelopen	dan	ik	dacht
PRON	AUX	ADV	ADV	VERB	SCONJ	PRON	VERB
Es	ist	ganz	anders	gelaufen	als	ich	dachte

- De-lexicalisation (Feldman, et al, 2006; Mcdonald, et al 2011; Reddy, Sharoff, 2011)

Это	прошло	совсем	иначе	чем	я	думал
PRON	VERB	ADV	ADV	SCONJ	PRON	VERB

它	和	我	想	的	完全	不同
it	and	I	thought	of	completely	differ
PRON	CCONJ	PRON	NOUN	PART	ADV	ADJ

- Projection from parallel corpora (Tiedemann, 2014): word alignment with pruning

Cross-lingual tagging and parsing

Het	is	heel	anders	gelopen	dan	ik	dacht
PRON	AUX	ADV	ADV	VERB	SCONJ	PRON	VERB
Es	ist	ganz	anders	gelaufen	als	ich	dachte

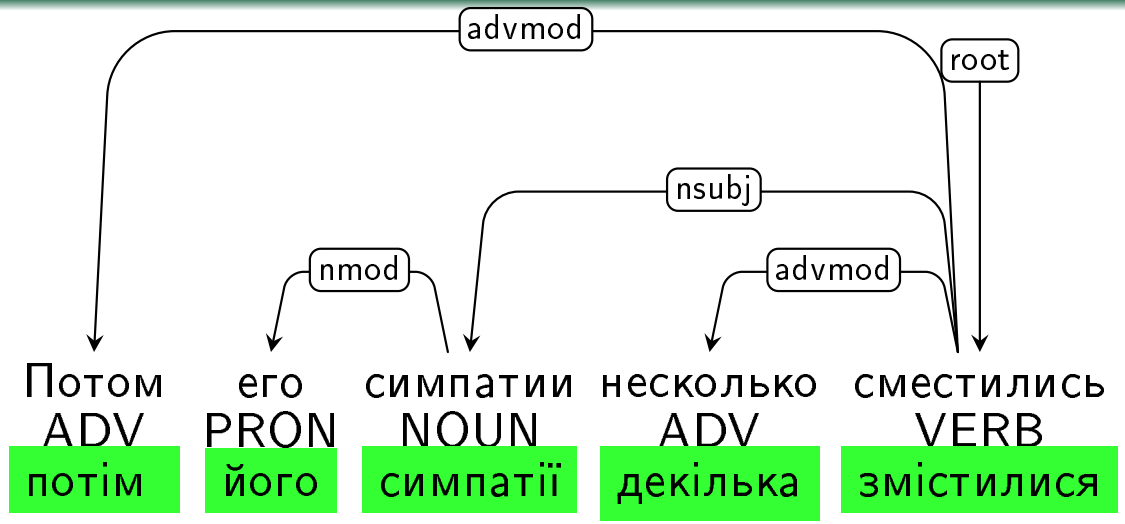
- De-lexicalisation (Feldman, et al, 2006; Mcdonald, et al 2011; Reddy, Sharoff, 2011)

Это	прошло	совсем	иначе	чем	я	думал
PRON	VERB	ADV	ADV	SCONJ	PRON	VERB

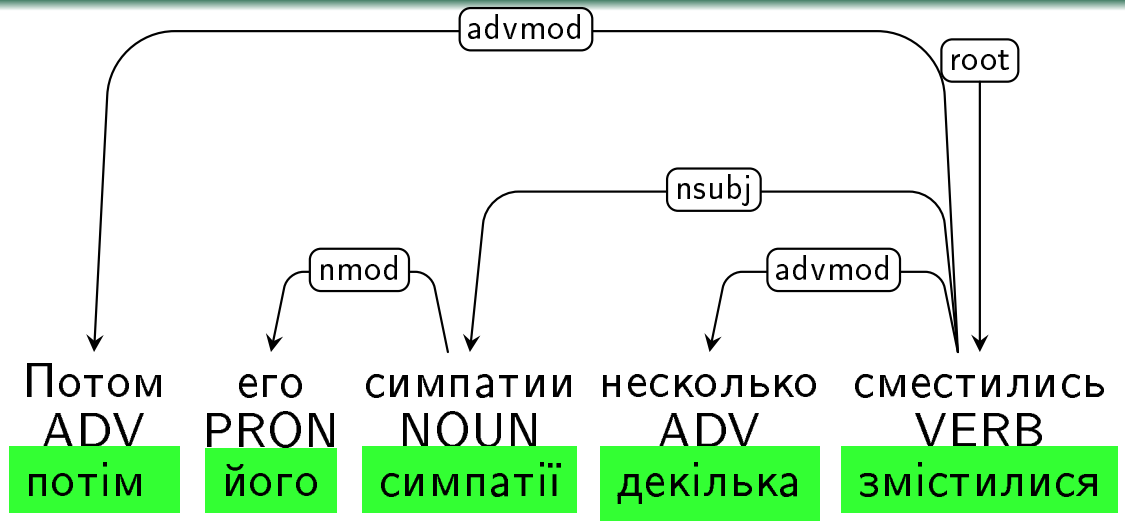
它	和	我	想	的	完全	不同
it	and	I	thought	of	completely	differ
PRON	CCONJ	PRON	NOUN	PART	ADV	ADJ

- Projection from parallel corpora (Tiedemann, 2014): word alignment with pruning
- Machine Translation: replacing parallel corpora with Machine Translation

Relexicalisation via cognates



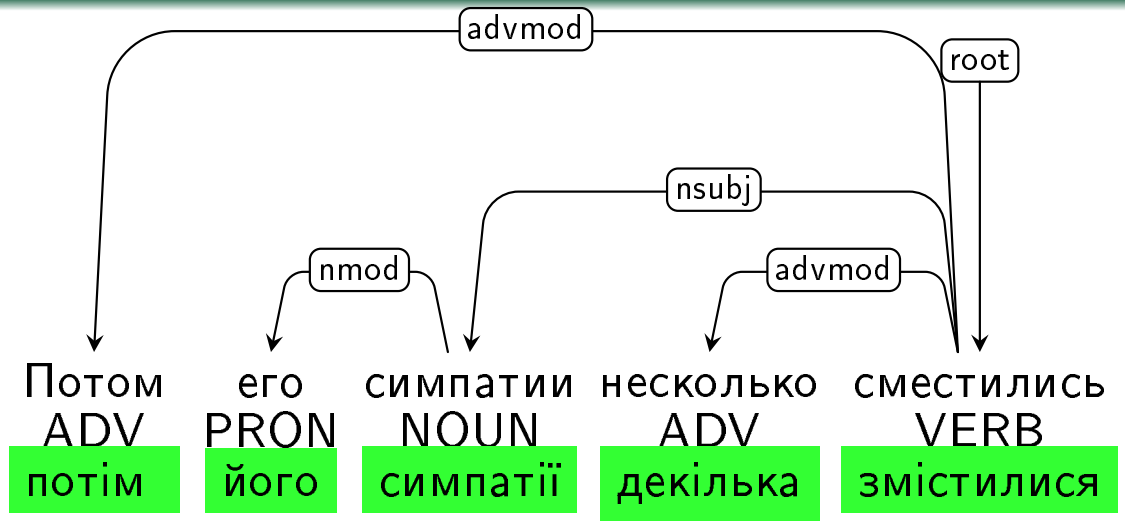
Relexicalisation via cognates



- Relexicalisation problems:

Найкраще	це	робити	в	русі	.
найкраще	это	работать	в	россии	.

Relexicalisation via cognates



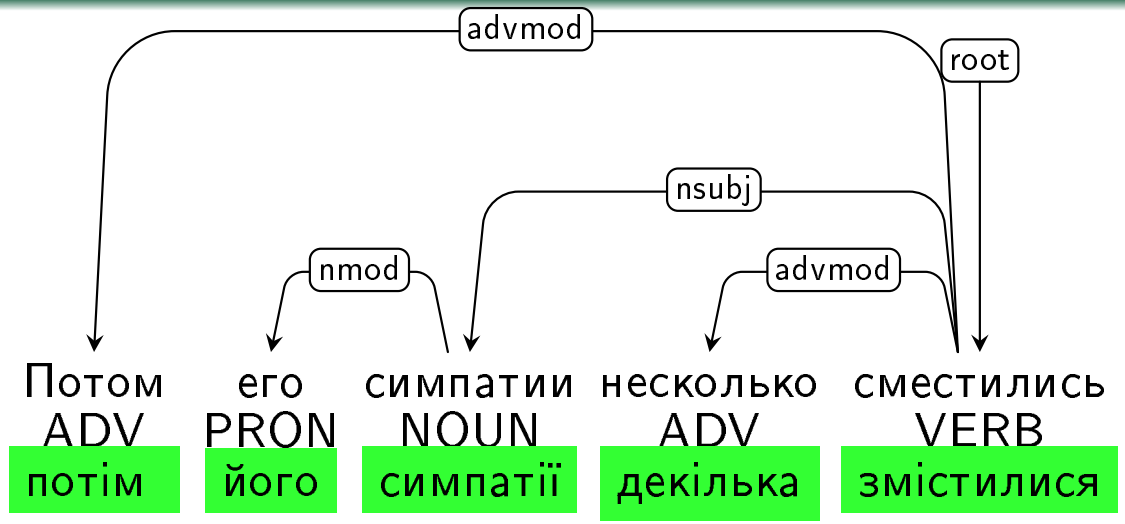
- Relexicalisation problems:

  Найкраще це робити в русі .

  найкраще это работать в россии .

→ 'this is best done in motion' (*рух→руси*)

Relexicalisation via cognates



- Relexicalisation problems:

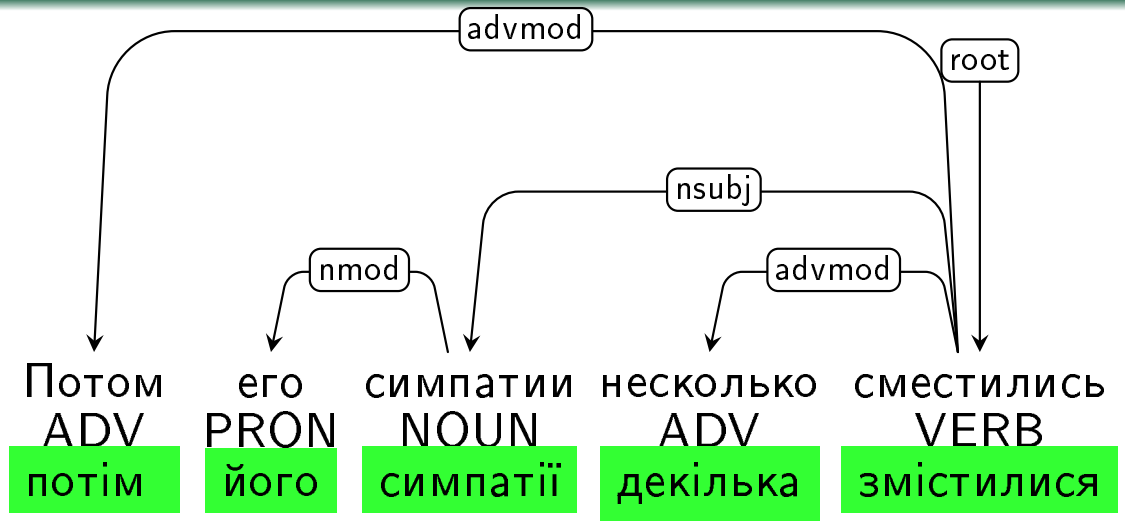
  Найкраще це робити в русі .

  найкраще это работать в россии .

→ 'this is best done in motion' (*рух→руси*)

- Representation of ambiguity in embeddings

Relexicalisation via cognates



- Relexicalisation problems:

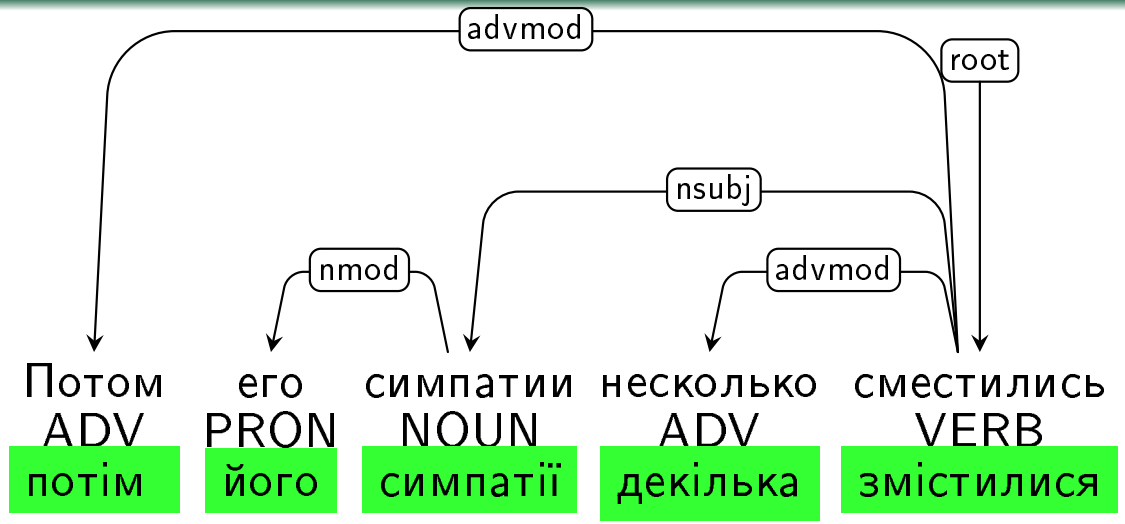
  Найкраще це робити в русі .

  найкраще это работать в россии .

→ 'this is best done in motion' (*рух→руси*)

- Representation of ambiguity in embeddings
- Train: tags from the donor, words from the recipient

Relexicalisation via cognates



- Relexicalisation problems:
  Найкраще це робити в русі .
  найкраще это работать в россии .

→ 'this is best done in motion' (*рух→руси*)

- Representation of ambiguity in embeddings
- Train: tags from the donor, words from the recipient
- Disambiguated RNC with Ukrainian cognates using TnT:
  ru: 0.95 (Sharoff, Nivre, 2011),
  uk: 0.92 (Babych, Sharoff, 2016)

Neural Networks experiments

- Recurrent Neural Language Models (Bengio, et al 2003)

Neural Networks experiments

- Recurrent Neural Language Models (Bengio, et al 2003)

  (Karpathy, 2015) character RNN on Shakespeare:
  PANDARUS:
  Alas, I think he shall be come approached and the day
  When little srain would be attain'd into being never fed,
  And who is but a chain and subjects of his death,
  I should not sleep.

Neural Networks experiments

- Recurrent Neural Language Models (Bengio, et al 2003)

  (Karpathy, 2015) character RNN on Shakespeare:
  PANDARUS:
  Alas, I think he shall be come approached and the day
  When little srain would be attain'd into being never fed,
  And who is but a chain and subjects of his death,
  I should not sleep.
- Neural Machine Translation (Bahdanau, et al 2014)

Neural Networks experiments

- Recurrent Neural Language Models (Bengio, et al 2003)

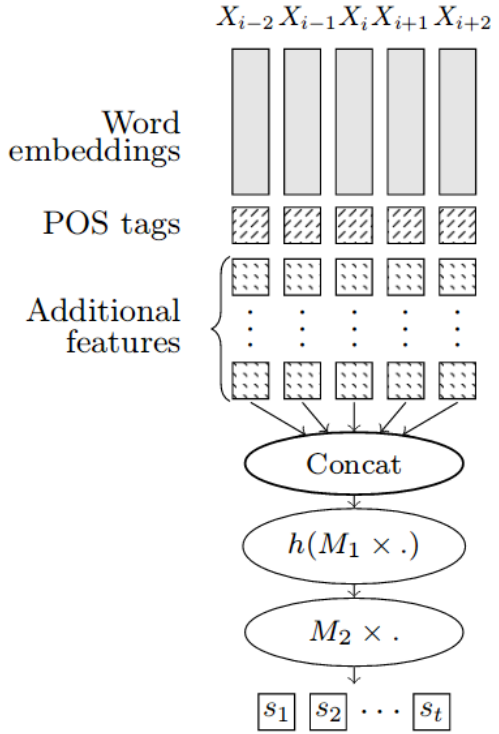
  (Karpathy, 2015) character RNN on Shakespeare:
  PANDARUS:
  Alas, I think he shall be come approached and the day
  When little srain would be attain'd into being never fed,
  And who is but a chain and subjects of his death,
  I should not sleep.
- Neural Machine Translation (Bahdanau, et al 2014)
- Sequence-to-sequence models (Collobert, et al 2011)

NER shared task

- "A glej , v pradavnem budističnem tekstu **Madhyamika/B-misc Karika/I-misc Vrrti/I-misc** piše : » **Budha/B-per** je učil"

NER shared task

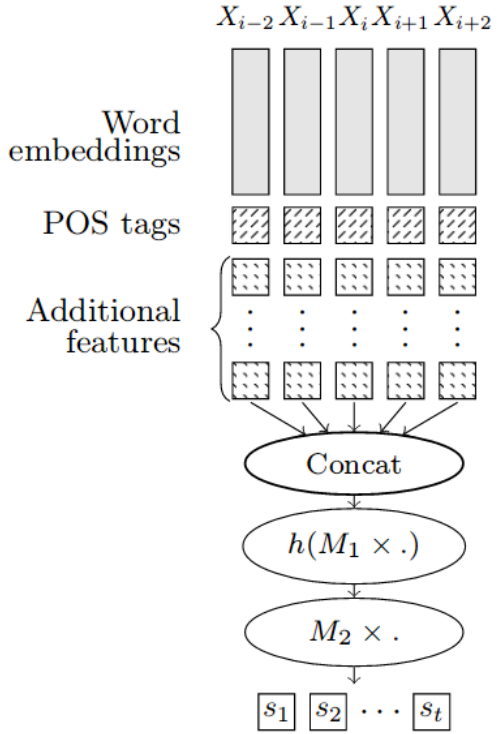
- "A glej , v pradavnem budističnem tekstu **Madhyamika/B-misc Karika/I-misc Vrrti/I-misc** piše : » **Budha/B-per** je učil"



NN model (Lample et al, 2016)
LSTM: 89%; LSTM+CRF: 91%

NER shared task

- "A glej , v pradavnem budističnem tekstu **Madhyamika/B-misc Karika/I-misc Vrrti/I-misc** piše : » **Budha/B-per** je učil"



NN model (Lample et al, 2016)
LSTM: 89%; LSTM+CRF: 91%

- Embeddings from recipient Wikipedias (2015), training on Slovenian SSJ500 (200kW)

Shared task results: F1

- JHU: SVM tagger with projections using on Europarl/UN

Shared task results: F1

- JHU: SVM tagger with projections using on Europarl/UN
- JRC: large gazetteers

Shared task results: F1

- JHU: SVM tagger with projections using on Europarl/UN
- JRC: large gazetteers
- EC:

cs	hr	pl	ru	sl	ua	
47.2	**46.2**	**44.8**	46.5	47.8	10.8	JHU
41.2	30.0	34.6	**53.7**	37.5	**20.8**	JRC
47.7	44.3	44.2	33.6	**59.5**	13.7	Sharoff

Shared task results: F1

- JHU: SVM tagger with projections using on Europarl/UN
- JRC: large gazetteers
- EC:

cs	hr	pl	ru	sl	ua	
47.2	**46.2**	**44.8**	46.5	47.8	10.8	JHU
41.2	30.0	34.6	**53.7**	37.5	**20.8**	JRC
47.7	44.3	44.2	33.6	**59.5**	13.7	Sharoff

- Trump:

cs	hr	pl	ru	sl	ua	
46.1	50.4	41.0	41.8	46.2	33.2	JHU
42.2	37.4	48.0	**55.6**	44.2	**50.8**	JRC
52.6	**52.4**	**55.2**	21.0	**62.6**	20.7	Sharoff

Closer look at Sl-Hr transfer

- Slovenian: Upper baseline

sl→sl	P	R	F1	N
Accuracy: 97.47	74.08	73.87	73.98	
loc	81.92	76.72	79.23	354
misc	30.77	17.52	22.33	78
org	64.06	57.68	60.70	217
per	77.49	88.73	82.73	813

Closer look at Sl-Hr transfer

- Slovenian: Upper baseline

sl→sl	P	R	F1	N
Accuracy: 97.47	74.08	73.87	73.98	
loc	81.92	76.72	79.23	354
misc	30.77	17.52	22.33	78
org	64.06	57.68	60.70	217
per	77.49	88.73	82.73	813

- Croatian SE Times test corpus (180kW)

sl→hr	P	R	F1	N
Accuracy: 93.98	67.40	63.64	65.47	
loc	81.66	61.27	70.01	709
misc	0.00	0.00	0.00	105
org	56.40	59.11	57.73	851
per	77.96	72.85	75.32	599

Gold vs test results

en	hr	Gold	Output
Bosnian	Bosanska	O	O
model	manekenka	O	O
in	u	O	O
dress	modelu	O	O
from	iz	O	O
collection	kolekcije	O	O
Omnia	Omnia	O	B-misc
Mei	Mei	O	I-misc
Serbian	srbijanske	O	I-misc
fashion	modne	O	I-misc
designer	kreatorice	O	I-misc
Slavica	Slavice	B-per	B-per
Aleksijev	Aleksijev	I-per	I-per

en	hr	Gold	Output
at	na	O	O
autumn	jesenjem	O	O
Sarajevo	Sarajevskom	O	B-org
week	tjednu	O	I-org
fashion	mode	O	I-org
2006	2006	O	O
	.	O	O
	(	O	O
Getty	Getty	O	B-misc
Images	Images	O	I-misc
	)	O	O

NLP: slipping into the Dark Ages

- Hand-made rules capture meaning
  Human intuition and knowledge

NLP: slipping into the Dark Ages

- Hand-made rules capture meaning
  Human intuition and knowledge
- Machine learning: interpretable feature spaces and black boxes of algorithms (Random Forest or SVM$^{??}$)

NLP: slipping into the Dark Ages

- Hand-made rules capture meaning
  Human intuition and knowledge
- Machine learning: interpretable feature spaces and black boxes of algorithms (Random Forest or SVM$^{??}$)

NLP: slipping into the Dark Ages

- Hand-made rules capture meaning
  Human intuition and knowledge
- Machine learning: interpretable feature spaces and black boxes of algorithms (Random Forest or SVM$^{??}$)

Jelinek vs Church

NLP: slipping into the Dark Ages

- Hand-made rules capture meaning
  Human intuition and knowledge
- Machine learning: interpretable feature spaces and
  black boxes of algorithms (Random Forest or SVM$^{??}$)

Jelinek vs Church

Jelinek: Every time I fire a linguist, the performance goes up

NLP: slipping into the Dark Ages

- Hand-made rules capture meaning
  Human intuition and knowledge
- Machine learning: interpretable feature spaces and
  black boxes of algorithms (Random Forest or SVM$^{??}$)

Jelinek vs Church

Jelinek: Every time I fire a linguist, the performance goes up

Church: Fire everybody and buy more data

NLP: slipping into the Dark Ages

- Hand-made rules capture meaning
  Human intuition and knowledge
- Machine learning: interpretable feature spaces and
  black boxes of algorithms (Random Forest or SVM$^{??}$)

Jelinek vs Church

Jelinek: Every time I fire a linguist, the performance goes up

Church: Fire everybody and buy more data

NLP: slipping into the Dark Ages

- Hand-made rules capture meaning
  Human intuition and knowledge
- Machine learning: interpretable feature spaces and
  black boxes of algorithms (Random Forest or SVM$^{??}$)

Jelinek vs Church

Jelinek: Every time I fire a linguist, the performance goes up

Church: Fire everybody and buy more data

Neil Lawrence quoted by Chris Manning

NLP: slipping into the Dark Ages

- Hand-made rules capture meaning
  Human intuition and knowledge
- Machine learning: interpretable feature spaces and
  black boxes of algorithms (Random Forest or SVM$^{??}$)

Jelinek vs Church

Jelinek: Every time I fire a linguist, the performance goes up

Church: Fire everybody and buy more data

Neil Lawrence quoted by Chris Manning

NLP is kind of like a rabbit in the headlights waiting to be flattened by the Deep Learning steam train.

NLP: slipping into the Dark Ages

- Hand-made rules capture meaning
  Human intuition and knowledge
- Machine learning: interpretable feature spaces and
  black boxes of algorithms (Random Forest or SVM$^{??}$)

Jelinek vs Church

Jelinek: Every time I fire a linguist, the performance goes up

Church: Fire everybody and buy more data

Neil Lawrence quoted by Chris Manning

NLP is kind of like a rabbit in the headlights waiting to be flattened by the Deep Learning steam train.

- In Deep Learning everything is a black box

The Zero Theorem by Terry Gilliam

Counteracting the Dark Ages with linguistics

- "Unreasonable effectiveness of Neural Networks"
  Easy to outperform other approaches

Counteracting the Dark Ages with linguistics

- "Unreasonable effectiveness of Neural Networks"
  Easy to outperform other approaches

?? Lots of hype and long training times
  Often Deep Learning is much more shallow

Counteracting the Dark Ages with linguistics

- "Unreasonable effectiveness of Neural Networks"
  Easy to outperform other approaches
- ?? Lots of hype and long training times
  Often Deep Learning is much more shallow
- ?? Winning by a small margin:
  Design new architecture or Buy more data?
  'Language is a large number of rare events'

Counteracting the Dark Ages with linguistics

- "Unreasonable effectiveness of Neural Networks"
  Easy to outperform other approaches
- ?? Lots of hype and long training times
  Often Deep Learning is much more shallow
- ?? Winning by a small margin:
  Design new architecture or Buy more data?
  'Language is a large number of rare events'
- From lots of data to one-shot learning:
  Marco Baroni's wampimuk

Counteracting the Dark Ages with linguistics

- "Unreasonable effectiveness of Neural Networks"
  Easy to outperform other approaches
- ?? Lots of hype and long training times
  Often Deep Learning is much more shallow
- ?? Winning by a small margin:
  Design new architecture or Buy more data?
  'Language is a large number of rare events'
- From lots of data to one-shot learning:
  Marco Baroni's wampimuk
- Very simple data integration with NNs
  $\rightarrow$ Easy to create shared representations

Counteracting the Dark Ages with linguistics

- "Unreasonable effectiveness of Neural Networks"
  Easy to outperform other approaches
- ?? Lots of hype and long training times
  Often Deep Learning is much more shallow
- ?? Winning by a small margin:
  Design new architecture or Buy more data?
  'Language is a large number of rare events'
- From lots of data to one-shot learning:
  Marco Baroni's wampimuk
- Very simple data integration with NNs
  $\rightarrow$ Easy to create shared representations
- Place for linguistics: what is shared?