Language Adaptation experiments

Cross-lingual embeddings for related languages

Serge Sharoff

Centre for Translation Studies
University of Leeds

14 June 2018

Outline

1. **Rationale for Language Adaptation**
   - Universal Dependencies
   - Multilingual terminology
   - Limitations of resources
2. **Detection of cognates**
   - Cross-lingual word embeddings
   - Weigted Levenshtein Distance
3. **Predicting morphology**
   - Syncretism across related languages
   - Impact of prediction
4. **Terminology augmentation**
   - Similarity across the forms
   - Cross-lingual prediction methods

Need for language adaptation

- 100 languages needed to cover 85% world's population

Need for language adaptation

- 100 languages needed to cover 85% world's population

  98-100. Balochi, Belarusian and Konkani, ≈8M speakers

Need for language adaptation

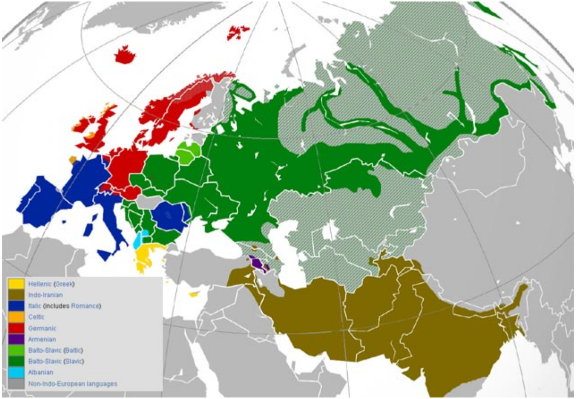
- 100 languages needed to cover 85% world's population
  98-100. Balochi, Belarusian and Konkani, ≈8M speakers
  40. Ukrainian, 30M native speakers (8. in Europe)

Need for language adaptation

- 100 languages needed to cover 85% world's population

  98-100. Balochi, Belarusian and Konkani, ≈8M speakers

  40. Ukrainian, 30M native speakers (8. in Europe)



Universal Dependencies

Roger Bacon, (c1250)

Universal Dependencies

Roger Bacon, (c1250)

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

Universal Dependencies

Roger 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 Dependencies

Roger 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 Dependencies

Roger 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 Dependencies

Roger 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 annotated corpora for 47 languages (Version 2.0)

Universal Dependencies

Roger 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 annotated corpora for 47 languages (Version 2.0)

- Balochi and Konkani not covered yet

Universal Dependencies

Roger 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 annotated corpora for 47 languages (Version 2.0)

- Balochi and Konkani not covered yet

BUT Farsi and Hindi are

UD examples: German and English

Stuttgart tagset (German) vs Penn tagset (English)

1	Ich	ich	PPER	PRON	Case=Nom|Num=Sing|Person=1|Type=Pers
2	kann	können	VMFIN	AUX	Num=Sing|Person=1|Tense=Pres|VerbForm=Fi
3	es	es	PPER	PRON	Case=Acc|Gender=Neut|Num=Sing|Person=3|T
4	nur	nur	ADV	ADV	
5	empfehlen	empfehlen	VVINF	VERB	VerbForm=Inf
6	.	.	.	PUNCT	

1	I	I	PRP	PRON	Case=Nom|Num=Sing|Person=1|Type=Pers
2	ca	can	MD	AUX	Tense=Pres|VerbForm=Fin
3	n't	not	RB	PART	
4	thank	thank	VB	VERB	VerbForm=Inf
5	you	you	PRP	PRON	Case=Acc|Person=2|Type=Pers
6	enough	enough	RB	ADV	
7	.	.	.	PUNCT	

IATE (InterActive Terminology for Europe)

- 1,293,271 term entries over 164 subject domains

IATE (InterActive Terminology for Europe)

- 1,293,271 term entries over 164 subject domains
- Top 10 domains cover 859,181 terms

Min	1stQ	Median	3rdQ	Max
1	435	1676	7457	150900

IATE (InterActive Terminology for Europe)

- 1,293,271 term entries over 164 subject domains
- Top 10 domains cover 859,181 terms

Min	1stQ	Median	3rdQ	Max
1	435	1676	7457	150900

Largest subject domain: id2841

acceptable risk	risque acceptable
acebrochol	acébrochol
acute	aigu
Bayes' theorem	théorème de Bayes

IATE (InterActive Terminology for Europe)

- 1,293,271 term entries over 164 subject domains
- Top 10 domains cover 859,181 terms

Min	1stQ	Median	3rdQ	Max
1	435	1676	7457	150900

Largest subject domain: id2841

acceptable risk	risque acceptable
acebrochol	acébrochol
acute	aigu
Bayes' theorem	théorème de Bayes

Two smallest subject domains: id4206, id360

acquisition cost	coût d'achat
reverse osmosis	osmose inverse

Statistics across languages

- 5,551 term entries of 1,293,271 available in 24 languages

Statistics across languages

- 5,551 term entries of 1,293,271 available in 24 languages

Language count:	Min	1stQ	Median	3rdQ	Max
	0	2	3	8	25

Statistics across languages

- 5,551 term entries of 1,293,271 available in 24 languages

Language count:

Min	1stQ	Median	3rdQ	Max
0	2	3	8	25

Pre-1990 languages

da	de	el	en	es	fr	it	nl	pt
461,133	692,844	401,754	965,785	471,205	957,818	520,751	512,050	401,708

Statistics across languages

- 5,551 term entries of 1,293,271 available in 24 languages

Language count:	Min	1stQ	Median	3rdQ	Max
	0	2	3	8	25

Pre-1990 languages

da	de	el	en	es	fr	it	nl	pt
461,133	692,844	401,754	965,785	471,205	957,818	520,751	512,050	401,708

2004/07 languages

bg	cs	et	hu	pl	ro	sk	sl	**Slav**
33,311	29,382	36,165	33,899	57,725	39,613	35,930	43,706	**16,728**

Statistics across languages

- 5,551 term entries of 1,293,271 available in 24 languages

Language count:

Min	1stQ	Median	3rdQ	Max
0	2	3	8	25

Pre-1990 languages

da	de	el	en	es	fr	it	nl	pt
461,133	692,844	401,754	965,785	471,205	957,818	520,751	512,050	401,708

2004/07 languages

bg	cs	et	hu	pl	ro	sk	sl	**Slav**
33,311	29,382	36,165	33,899	57,725	39,613	35,930	43,706	**16,728**

- *second reading*→*drugie czytanie* (Polish term hunting),

Statistics across languages

- 5,551 term entries of 1,293,271 available in 24 languages

Language count:	Min	1stQ	Median	3rdQ	Max
	0	2	3	8	25

Pre-1990 languages

da	de	el	en	es	fr	it	nl	pt
461,133	692,844	401,754	965,785	471,205	957,818	520,751	512,050	401,708

2004/07 languages

bg	cs	et	hu	pl	ro	sk	sl	**Slav**
33,311	29,382	36,165	33,899	57,725	39,613	35,930	43,706	**16,728**

- *second reading*→*drugie czytanie* (Polish term hunting),

→ Similar terms: *druhé čtení* (cs) or *второ четене* (bg)

Limitations of resources

Limitations of resources

Languages	UD	Wiki	PEMT
Romance			
Catalan	442K	181M	
French	367K	667M	432K
Italian	266K	433M	329K
Portuguese	454K	222M	321K
Romanian	109K	63M	
Spanish	853K	530M	265K
Slavonic			
Belarusian	2K	20M	
Bulgarian	124K	55M	
Czech	1671K	110M	183K
Polish	70K	227M	213K
Russian	928K	420M	266K
Slovenian	136K	321M	
Ukrainian	10K	161M	

Limitations of resources

Languages	UD	Wiki	PEMT
Romance			
Catalan	442K	181M	
French	367K	667M	432K
Italian	266K	433M	329K
Portuguese	454K	222M	321K
Romanian	109K	63M	
Spanish	853K	530M	265K
Slavonic			
Belarusian	2K	20M	
Bulgarian	124K	55M	
Czech	1671K	110M	183K
Polish	70K	227M	213K
Russian	928K	420M	266K
Slovenian	136K	321M	
Ukrainian	10K	161M	

Limitations of resources

Languages	UD	Wiki	PEMT
Romance			
Catalan	442K	181M	
French	367K	667M	432K
Italian	266K	433M	329K
Portuguese	454K	222M	321K
Romanian	109K	63M	
Spanish	853K	530M	265K
Slavonic			
Belarusian	2K	20M	
Bulgarian	124K	55M	
Czech	1671K	110M	183K
Polish	70K	227M	213K
Russian	928K	420M	266K
Slovenian	136K	321M	
Ukrainian	10K	161M	

- Variation in annotation
  Only in Czech:
  Style=Arch|Coll|Slng|Vrnc|Vulg...
  NameType=Geo|Giv|Sur|Nat...
  Inconsistencies:
  Polarity=Pos and Hyph=Yes

Limitations of resources

Languages	UD	Wiki	PEMT
Romance			
Catalan	442K	181M	
French	367K	667M	432K
Italian	266K	433M	329K
Portuguese	454K	222M	321K
Romanian	109K	63M	
Spanish	853K	530M	265K
Slavonic			
Belarusian	2K	20M	
Bulgarian	124K	55M	
Czech	1671K	110M	183K
Polish	70K	227M	213K
Russian	928K	420M	266K
Slovenian	136K	321M	
Ukrainian	10K	161M	

- Variation in annotation
  Only in Czech:
  Style=Arch|Coll|Slng|Vrnc|Vulg...
  NameType=Geo|Giv|Sur|Nat...
  Inconsistencies:
  Polarity=Pos and Hyph=Yes
- Tagsets are sparse:
  685 uk vs 710 ru
  440 ro vs 221 fr

Limitations of resources

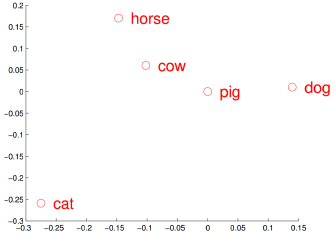
Languages	UD	Wiki	PEMT
Romance			
Catalan	442K	181M	
French	367K	667M	432K
Italian	266K	433M	329K
Portuguese	454K	222M	321K
Romanian	109K	63M	
Spanish	853K	530M	265K
Slavonic			
Belarusian	2K	20M	
Bulgarian	124K	55M	
Czech	1671K	110M	183K
Polish	70K	227M	213K
Russian	928K	420M	266K
Slovenian	136K	321M	
Ukrainian	10K	161M	

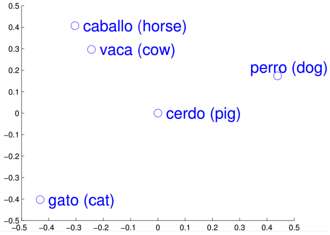
- Variation in annotation
  Only in Czech:
  Style=Arch|Coll|Slng|Vrnc|Vulg...
  NameType=Geo|Giv|Sur|Nat...
  Inconsistencies:
  Polarity=Pos and Hyph=Yes
- Tagsets are sparse:
  685 uk vs 710 ru
  440 ro vs 221 fr
- 45 single examples in ru vs 237 in uk:
  колотыми V,Aspect=Imperf,Case=Inst,
  Num=Plur,Tense=Past,Voice=Passive
  найпотужнішої
  ADJ,Case=Gen,Degree=Sup,Gender=Fem

Outline

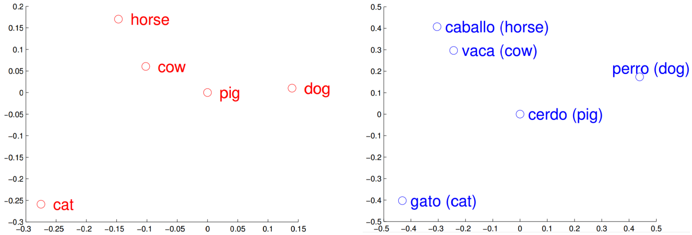
1. Rationale for Language Adaptation
   - Universal Dependencies
   - Multilingual terminology
   - Limitations of resources
2. Detection of cognates
   - Cross-lingual word embeddings
   - Weigted Levenshtein Distance
3. Predicting morphology
   - Syncretism across related languages
   - Impact of prediction
4. Terminology augmentation
   - Similarity across the forms
   - Cross-lingual prediction methods

Cross-lingual word embeddings (Mikolov, 2013)



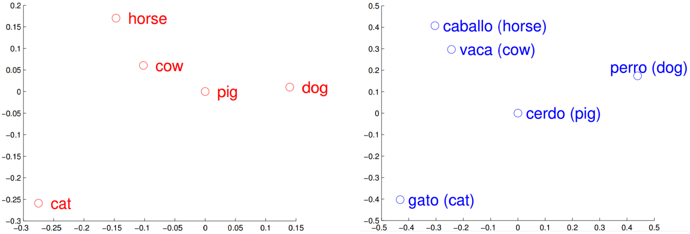


Cross-lingual word embeddings (Mikolov, 2013)



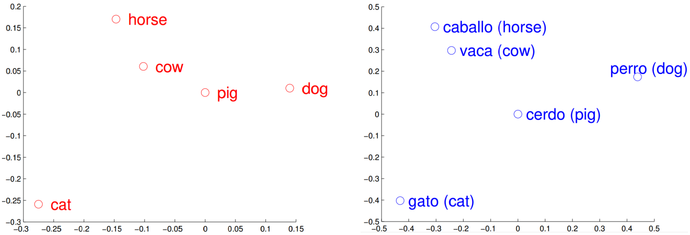
- Earlier vector models (Rapp, 1995; Fung, McKeown, 1997)

Cross-lingual word embeddings (Mikolov, 2013)



- Earlier vector models (Rapp, 1995; Fung, McKeown, 1997)
- Predicting multi-word expressions (Sharoff, et al, 2006)

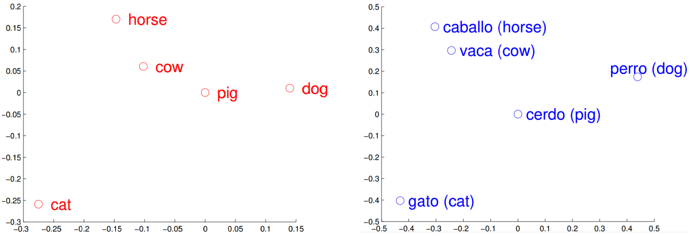
Cross-lingual word embeddings (Mikolov, 2013)



- Earlier vector models (Rapp, 1995; Fung, McKeown, 1997)
- Predicting multi-word expressions (Sharoff, et al, 2006)
- Linear transform or MLP for monolingual embeddings

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

Cross-lingual word embeddings (Mikolov, 2013)



- Earlier vector models (Rapp, 1995; Fung, McKeown, 1997)
- Predicting multi-word expressions (Sharoff, et al, 2006)
- Linear transform or MLP for monolingual embeddings

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

- SGD (Mikolov, et al 2013), CCA (Faruqui, et al 2014), multivariate regression (Dinu, et al 2014), regression with orthogonalisation constraints (Artetxe, et al 2016)

Dictionaries for LA setting

- Dictionary of word forms (if no lemmatisation)

Dictionaries for LA setting

- Dictionary of word forms (if no lemmatisation)
- Alignment of parallel corpora:
  Europarl, UN, OPUS (subtitles): little for pl-ru

Dictionaries for LA setting

- Dictionary of word forms (if no lemmatisation)
- Alignment of parallel corpora:
  Europarl, UN, OPUS (subtitles): little for pl-ru
- Small (2-5kW) bilingual dictionaries from iWiki links:

(sv) Slaget om Filippinen	(de) Schlacht um die Philippinen
(pl) Z życia marionetek	(ru) Из жизни марионеток
(pl) Wskaźnik jakości życia	(ru) Индекс качества жизни
(sk) Karneval zvierat	(ru) Карнавал животных
(sk) Práva zvierat	(ru) Права животных

Dictionaries for LA setting

- Dictionary of word forms (if no lemmatisation)
- Alignment of parallel corpora:
  Europarl, UN, OPUS (subtitles): little for pl-ru
- Small (2-5kW) bilingual dictionaries from iWiki links:

(sv) Slaget om Filippinen	(de) Schlacht um die Philippinen
(pl) Z życia marionetek	(ru) Из жизни марионеток
(pl) Wskaźnik jakości życia	(ru) Индекс качества жизни
(sk) Karneval zvierat	(ru) Карнавал животных
(sk) Práva zvierat	(ru) Права животных

- Many non-cognates: Sk: *zviera* (animal);
  Ru: *зверь* (wild animal) vs *животное* (animal)

Dictionaries for LA setting

- Dictionary of word forms (if no lemmatisation)
- Alignment of parallel corpora:
  Europarl, UN, OPUS (subtitles): little for pl-ru
- Small (2-5kW) bilingual dictionaries from iWiki links:

(sv) Slaget om Filippinen	(de) Schlacht um die Philippinen
(pl) Z życia marionetek	(ru) Из жизни марионеток
(pl) Wskaźnik jakości życia	(ru) Индекс качества жизни
(sk) Karneval zvierat	(ru) Карнавал животных
(sk) Práva zvierat	(ru) Права животных

- Many non-cognates: Sk: *zviera* (animal);
  Ru: *зверь* (wild animal) vs *животное* (animal)
- Lists of geonames and persons: filtering by frequency
  Alapajevsk, Alarich, Alasdair MacIntyre, Alaska, Alassio, Alastair G.W. Cameron, Alata, Alathfar, Alatri, Alatyr

Levenshtein distances

- Baseline Levenshtein distance (LD):
  Philippinen $\rightarrow$ *Filippinen* : 1 del, 1 sub ($\frac{2}{11}$)
  Schlacht $\rightarrow$ *Slaget* : 2 del, 2 sub ($\frac{4}{8}$)

Levenshtein distances

- Baseline Levenshtein distance (LD):
  Philippinen $\rightarrow$ *Filippinen* : 1 del, 1 sub ($\frac{2}{11}$)
  Schlacht $\rightarrow$ *Slaget* : 2 del, 2 sub ($\frac{4}{8}$)
- 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 del, 1 sub ($\frac{2}{11}$)
  Schlacht $\rightarrow$ *Slaget* : 2 del, 2 sub ($\frac{4}{8}$)
- Weighted Levenshtein Distance (WLD) for cognates

Sch	l	a	ch	t
S	l	a	ge	t

- Alignment probabilities: $p(sch \rightarrow s) = 0.7; p(l \rightarrow s) = 0$

Levenshtein distances

- Baseline Levenshtein distance (LD):
  Philippinen → *Filippinen* : 1 del, 1 sub ($\frac{2}{11}$)
  Schlacht → *Slaget* : 2 del, 2 sub ($\frac{4}{8}$)
- Weighted Levenshtein Distance (WLD) for cognates

  Sch l a ch t
  S l a ge t
- Alignment probabilities: $p(sch \rightarrow s) = 0.7; p(l \rightarrow s) = 0$

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

Levenshtein distances

- Baseline Levenshtein distance (LD):
  Philippinen $\rightarrow$ *Filippinen* : 1 del, 1 sub ($\frac{2}{11}$)
  Schlacht $\rightarrow$ *Slaget* : 2 del, 2 sub ($\frac{4}{8}$)
- Weighted Levenshtein Distance (WLD) for cognates

  Sch l a ch t
  S l a ge t

- Alignment probabilities: $p(sch \rightarrow s) = 0.7; p(l \rightarrow s) = 0$

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

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

Levenshtein distances

- Baseline Levenshtein distance (LD):
  Philippinen → *Filippinen* : 1 del, 1 sub ($\frac{2}{11}$)
  Schlacht → *Slaget* : 2 del, 2 sub ($\frac{4}{8}$)
- Weighted Levenshtein Distance (WLD) for cognates

Sch	l	a	ch	t
S	l	a	ge	t

- Alignment probabilities: $p(sch \rightarrow s) = 0.7; p(l \rightarrow s) = 0$

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

- Also WLD works across charsets:

m	a	r	i	o	n	e	t	e	k	ż	y	c	∅	i	a
м	а	р	и	о	н	е	т	о	к	ж	и	з	н		и

- Two alignment cycles: most likely **cognate** pairs

Integration of WLD into embeddings

- Cross-lingual spaces (Mikolov et al.,2013):

$$\min_{W} \sum ||We_i - f_i||^2 \quad (2)$$

Integration of WLD into embeddings

- Cross-lingual spaces (Mikolov et al.,2013):

$$\min_{W} \sum ||We_i - f_i||^2 \quad (2)$$

- Orthogonality constraint (Artetxe et al.,2016):

$$W = V \times U^T \quad (3)$$

Integration of WLD into embeddings

- Cross-lingual spaces (Mikolov et al.,2013):

$$\min_{W} \sum ||We_i - f_i||^2 \quad (2)$$

- Orthogonality constraint (Artetxe et al.,2016):

$$W = V \times U^T \quad (3)$$

when V and U come from SVD factorisation of $F \times E^T$

Integration of WLD into embeddings

- Cross-lingual spaces (Mikolov et al.,2013):

$$\min_{W} \sum ||We_i - f_i||^2 \qquad (2)$$

- Orthogonality constraint (Artetxe et al.,2016):

$$W = V \times U^T \qquad (3)$$

when V and U come from SVD factorisation of $F \times E^T$

- Dictionary can be produced by:

$$score(s_e, s_f) = \alpha \cos(v_e, v_f) + (1 - \alpha) WLD(s_e, s_f) \qquad (4)$$

Integration of WLD into embeddings

- Cross-lingual spaces (Mikolov et al.,2013):

$$\min_{W} \sum ||We_i - f_i||^2 \tag{2}$$

- Orthogonality constraint (Artetxe et al.,2016):

$$W = V \times U^T \tag{3}$$

when V and U come from SVD factorisation of $F \times E^T$

- Dictionary can be produced by:

$$score(s_e, s_f) = \alpha \cos(v_e, v_f) + (1 - \alpha) WLD(s_e, s_f) \tag{4}$$

- Refinement for building cross-lingual spaces:

Integration of WLD into embeddings

- Cross-lingual spaces (Mikolov et al.,2013):

$$\min_{W} \sum ||We_i - f_i||^2 \qquad (2)$$

- Orthogonality constraint (Artetxe et al.,2016):

$$W = V \times U^T \qquad (3)$$

when V and U come from SVD factorisation of $F \times E^T$

- Dictionary can be produced by:

$$score(s_e, s_f) = \alpha \cos(v_e, v_f) + (1 - \alpha) WLD(s_e, s_f) \qquad (4)$$

- Refinement for building cross-lingual spaces:
    1. Large dictionary of reliable cognates via (4)

Integration of WLD into embeddings

- Cross-lingual spaces (Mikolov et al.,2013):

$$\min_{W} \sum ||We_i - f_i||^2 \tag{2}$$

- Orthogonality constraint (Artetxe et al.,2016):

$$W = V \times U^T \tag{3}$$

when V and U come from SVD factorisation of $F \times E^T$

- Dictionary can be produced by:

$$score(s_e, s_f) = \alpha \cos(v_e, v_f) + (1 - \alpha) WLD(s_e, s_f) \tag{4}$$

- Refinement for building cross-lingual spaces:
  1. Large dictionary of reliable cognates via (4)
  2. Re-alignment of spaces using this dictionary

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+Orth+LD	0.501
GC+Orth+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+Orth+LD	0.501
GC+Orth+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+Orth+LD	0.501
GC+Orth+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 (cognates)	0.562
FT+Orth+GC (cognates)	0.601
FT+Orth+GC+WLD (cognates)	**0.681**

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+Orth+LD	0.501
GC+Orth+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 (cognates)	0.562
FT+Orth+GC (cognates)	0.601
FT+Orth+GC+WLD (cognates)	**0.681**
Adversarial NN (Conneau et al, 2017)	0.451
CSLS cost (Joulin et al, 2018)	0.453

Dictionaries for Slavonic languages

en-it	State-of-the-art (Artetxe, et al 2016)	0.393
	Weighted Levenshtein Distance	**0.531**

Dictionaries for Slavonic languages

en-it	State-of-the-art (Artetxe, et al 2016)	0.393
	Weighted Levenshtein Distance	**0.531**

en-it When selecting cognates only (45%)

This removes questionable translation equivalents:
absolve / esimere or *abysmally / malo* ('bad(ly)')

State-of-the-art (Artetxe, et al 2016)	0.601
Weighted Levenshtein Distance	**0.692**

Dictionaries for Slavonic languages

en-it

State-of-the-art (Artetxe, et al 2016)	0.393
Weighted Levenshtein Distance	**0.531**

en-it When selecting cognates only (45%)
This removes questionable translation equivalents:
absolve / esimere or *abysmally / malo* ('bad(ly)')

State-of-the-art (Artetxe, et al 2016)	0.601
Weighted Levenshtein Distance	**0.692**

- Producing cross-lingual Panslavonic embeddings:

	sl-hr	sl-cs	sl-pl	sl-ru	ru-uk	cs-sk
SOTA:	0.429	0.611	0.584	0.566	0.929	0.814
With WLD:	0.840	0.763	0.751	0.662	0.945	0.910

Dictionaries for Slavonic languages

en-it

State-of-the-art (Artetxe, et al 2016)	0.393
Weighted Levenshtein Distance	**0.531**

en-it When selecting cognates only (45%)

This removes questionable translation equivalents:
absolve / esimere or *abysmally / malo* ('bad(ly)')

State-of-the-art (Artetxe, et al 2016)	0.601
Weighted Levenshtein Distance	**0.692**

- Producing cross-lingual Panslavonic embeddings:

	sl-hr	sl-cs	sl-pl	sl-ru	ru-uk	cs-sk
SOTA:	0.429	0.611	0.584	0.566	0.929	0.814
With WLD:	0.840	0.763	0.751	0.662	0.945	0.910

- In-family embedding spaces are better than multilingual ones: Success in NER Shared task at BSNLP'17

Outline

1. Rationale for Language Adaptation
   - Universal Dependencies
   - Multilingual terminology
   - Limitations of resources
2. Detection of cognates
   - Cross-lingual word embeddings
   - Weigted Levenshtein Distance
3. Predicting morphology
   - Syncretism across related languages
   - Impact of prediction
4. Terminology augmentation
   - Similarity across the forms
   - Cross-lingual prediction methods

Prediction from cross-lingual embeddings

- Syncretism: one form can serve several syntactic functions
  Fr:*je/il anticip**e*** vs Es:*yo anticipo/el anticipa*

Prediction from cross-lingual embeddings

- Syncretism: one form can serve several syntactic functions
  Fr:*je/il anticip**e*** vs Es:*yo anticipo/el anticipa*

Prediction from cross-lingual embeddings

- Syncretism: one form can serve several syntactic functions
  Fr:*je/il anticipe* vs Es:*yo anticipo/el anticipa*

Forms of *green*	Russian		Ukrainian	
	Masc	Fem	Masc	Fem
Nominative	зелёный	зелёная	зелений	зелена
Genitive	зелёного	зелён**ой**	зеленого	зеленої
Dative	зелёному	зелён**ой**	зелен**ому**	зеленій
Instrumental	зелёным	зелён**ой**	зеленим	зеленою
Locative	зелёном	зелён**ой**	зелен**ому**	зеленій

Prediction from cross-lingual embeddings

- Syncretism: one form can serve several syntactic functions
  Fr:*je/il anticipe* vs Es:*yo anticipo/el anticipa*

Forms of *green*	Russian Masc	Russian Fem	Ukrainian Masc	Ukrainian Fem
Nominative	зелёный	зелёная	зелений	зелена
Genitive	зелёного	зелён**ой**	зеленого	зеленої
Dative	зелёному	зелён**ой**	зелен**ому**	зеленій
Instrumental	зелёным	зелён**ой**	зеленим	зеленою
Locative	зелёном	зелён**ой**	зелен**ому**	зеленій

- **Problem:** Cross-lingual mappings between the forms are not one-to-one even across closely related languages

Prediction of morphology

RQ Do embeddings knows about morphology?
Does this knowledge remain after the linear transform?

Prediction of morphology

RQ Do embeddings knows about morphology?
Does this knowledge remain after the linear transform?

- (Linzen, et al, 2016), (Belinkov, et al, 2017):
  predicting properties from embeddings

Prediction of morphology

RQ Do embeddings knows about morphology?
Does this knowledge remain after the linear transform?

- (Linzen, et al, 2016), (Belinkov, et al, 2017):
  predicting properties from embeddings

ru зелёному=(-0.047 -0.032 -0.101 0.007 0.021 -0.046 0.0066 0.095...)
→Case=Dat|Gender=Masc,Neut|Number=Sing

Prediction of morphology

RQ Do embeddings knows about morphology?
Does this knowledge remain after the linear transform?

- (Linzen, et al, 2016), (Belinkov, et al, 2017):
predicting properties from embeddings

ru зелёному=(-0.047 -0.032 -0.101 0.007 0.021 -0.046 0.0066 0.095...)
→Case=Dat|Gender=Masc,Neut|Number=Sing

uk зеленому=(-0.044 -0.062 -0.137 -0.035 -0.019 0.058 0.106 0.017...)
→Case=Dat,Loc|Gender=Masc,Neut|Number=Sing

Prediction of morphology

RQ Do embeddings knows about morphology?
Does this knowledge remain after the linear transform?

- (Linzen, et al, 2016), (Belinkov, et al, 2017):
  predicting properties from embeddings

ru зелёному=(-0.047 -0.032 -0.101 0.007 0.021 -0.046 0.0066 0.095. . .)
→Case=Dat|Gender=Masc,Neut|Number=Sing

uk зеленому=(-0.044 -0.062 -0.137 -0.035 -0.019 0.058 0.106 0.017. . .)
→Case=Dat,Loc|Gender=Masc,Neut|Number=Sing

- Direct prediction and by using cross-lingual embedding for training: Cs→Sk, Ru→Uk

Prediction results: Language adaptation

- Prediction is by Multi-layer Perceptron (300, 75, tanh)

O training on the original UD lexicon

T using cross-lingual embedding by transfer from related languages: Cs→Sk, Ru→Uk

	POS	$Tags_O$	$Tags_T$	$Train_O$	$Train_T$	MLP_O	MLP_T
Slovak	adj	23	202	1061	10778	45%	52%
	nouns	45	78	3537	8919	31%	43%
	verbs	30	61	1333	4695	49%	54%
Ukrainian	adj	45	54	1394	6235	40%	47%
	nouns	47	58	4187	14054	50%	58%
	verbs	32	54	2123	5765	55%	59%

Impact of prediction

Proportion of OOV words in the lexicons

	Cs	Ru	Pl	Sk	Be	Uk
Train	108257	97749	19344	19100	1628	5080
Test	32461	26567	4778	5425	662	271
OOV #	7891	8034	2327	3385	436	192
OOV %	24.31%	30.24%	48.70%	62.40%	65.86%	70.85%

Impact of prediction

Proportion of OOV words in the lexicons

	Cs	Ru	Pl	Sk	Be	Uk
Train	108257	97749	19344	19100	1628	5080
Test	32461	26567	4778	5425	662	271
OOV #	7891	8034	2327	3385	436	192
OOV %	24.31%	30.24%	48.70%	62.40%	65.86%	70.85%

- Predicting OOV as open-class words (Noun, Verb, Adj, Adv, X)

Impact of prediction

Proportion of OOV words in the lexicons

	Cs	Ru	Pl	Sk	Be	Uk
Train	108257	97749	19344	19100	1628	5080
Test	32461	26567	4778	5425	662	271
OOV #	7891	8034	2327	3385	436	192
OOV %	24.31%	30.24%	48.70%	62.40%	65.86%	70.85%

- Predicting OOV as open-class words (Noun, Verb, Adj, Adv, X)

Precision of UDPipe POS taggers

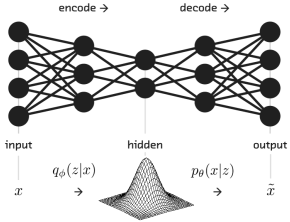
	Pl (Cs)	Sk (Cs)	Sk (Ru)	Be (Ru)	Uk (Ru)
Baseline (train only)	70.33	79.82	79.82	58.79	70.01
With added lexicon	**82.34**	**83.03**	**81.42**	**71.20**	**82.79**

Future cross lingual morphology prediction

- Signals beyond embeddings: endings, morphology clusters
  ой is a strong signal in Russian (60% adjectives)
  Signals differ: **ій,ої,ою** in Ukrainian

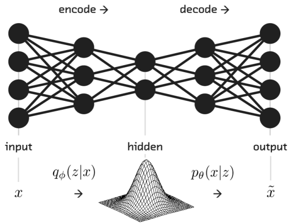
Future cross lingual morphology prediction

- Signals beyond embeddings: endings, morphology clusters
  ой is a strong signal in Russian (60% adjectives)
  Signals differ: **ій,ої,ою** in Ukrainian
- Variational Autoencoder: inference of regularities



Future cross lingual morphology prediction

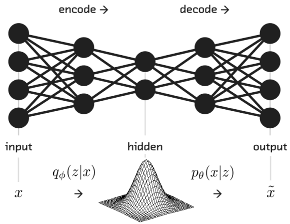
- Signals beyond embeddings: endings, morphology clusters
  ой is a strong signal in Russian (60% adjectives)
  Signals differ: **ій,ої,ою** in Ukrainian
- Variational Autoencoder: inference of regularities



- Adversarial training: faking similarities

Future cross lingual morphology prediction

- Signals beyond embeddings: endings, morphology clusters
  ой is a strong signal in Russian (60% adjectives)
  Signals differ: **ій,ої,ою** in Ukrainian
- Variational Autoencoder: inference of regularities



- Adversarial training: faking similarities
- Proper transfer learning:
  train on related languages with morph prediction

Outline

1. Rationale for Language Adaptation
   - Universal Dependencies
   - Multilingual terminology
   - Limitations of resources
2. Detection of cognates
   - Cross-lingual word embeddings
   - Weigted Levenshtein Distance
3. Predicting morphology
   - Syncretism across related languages
   - Impact of prediction
4. Terminology augmentation
   - Similarity across the forms
   - Cross-lingual prediction methods

Similarity across the forms

Single-word terms

English	Polish	Slovenian
minority	mniejszość	manjšina
homelessness	bezdomność	brezdomstvo
admissibility	dopuszczalność	dopustnost
drug, narcotic	narkotyk	droga, narkotik

Similarity across the forms

Single-word terms

English	Polish	Slovenian
minority	mniejszość	manjšina
homelessness	bezdomność	brezdomstvo
admissibility	dopuszczalność	dopustnost
drug, narcotic	narkotyk	droga, narkotik

Multiword terms

English	Polish	Slovenian
Graham's salt	sól Grahama	grahamova sol
Maddrell's salt	sól Maddrella	maddrellova sol
sodium hexametaphosphate	heksametafosforan sodu	natrijev heksametafosfat
sodium metaphosphate	metafosforan sodu	natrijev metafosfat
glassy sodium polyphosphate	szklisty polifosforan sodu	steklast natrijev polifosfat

Single-word term augmentation

- Test set: single-word terms in the shared set
  Corpus: combined Wikipedias and Europarl

Single-word term augmentation

- Test set: single-word terms in the shared set
  Corpus: combined Wikipedias and Europarl

	bg		cs		sl	
Test #	2229		2186		2194	
Found #	792		862		766	
	Orth	WLD	Orth	WLD	Orth	WLD
prec@1	0.225	0.480	0.413	0.541	0.251	0.433
prec@5	0.393	0.595	0.580	0.668	0.422	0.555
prec@10	0.458	0.621	0.633	0.701	0.490	0.584
recall@1	0.220	0.467	0.397	0.519	0.234	0.408
recall@5	0.383	0.576	0.557	0.644	0.395	0.527
recall@10	0.447	0.604	0.609	0.678	0.460	0.555

Cross-lingual term prediction: future

- Embedding spaces: good for single-word terms

Cross-lingual term prediction: future

- Embedding spaces: good for single-word terms
- Problems with OOV and morphology in MWEs:
  pl: *metafosforan sodu* (Case=Gen)

Cross-lingual term prediction: future

- Embedding spaces: good for single-word terms
- Problems with OOV and morphology in MWEs:
  pl: *metafosforan sodu* (Case=Gen)
- Terms contain cognates and non-cognates
  sl: *natrijev metafosfat*

Cross-lingual term prediction: future

- Embedding spaces: good for single-word terms
- Problems with OOV and morphology in MWEs:
  pl: *metafosforan sodu* (Case=Gen)
- Terms contain cognates and non-cognates
  sl: *natrijev metafosfat*
- Word-level MT: same OOV problems, small training set

Cross-lingual term prediction: future

- Embedding spaces: good for single-word terms
- Problems with OOV and morphology in MWEs:
  pl: *metafosforan sodu* (Case=Gen)
- Terms contain cognates and non-cognates
  sl: *natrijev metafosfat*
- Word-level MT: same OOV problems, small training set
- Character-level MT: term/morphology hallucination

Cross-lingual term prediction: future

- Embedding spaces: good for single-word terms
- Problems with OOV and morphology in MWEs:
  pl: *metafosforan sodu* (Case=Gen)
- Terms contain cognates and non-cognates
  sl: *natrijev metafosfat*
- Word-level MT: same OOV problems, small training set
- Character-level MT: term/morphology hallucination
- (Iwai, et al, 2017): term inference on a graph
  information processing, information retrieval, data retrieval → *data processing*

Constraints on term structure

- Domain relevance and specialised corpora

Constraints on term structure

- Domain relevance and specialised corpora
- Vast search space in comparison to single words

Constraints on term structure

- Domain relevance and specialised corpora
- Vast search space in comparison to single words

BUT Regular term formation via compounding

Constraints on term structure

- Domain relevance and specialised corpora
- Vast search space in comparison to single words

BUT Regular term formation via compounding

- Prediction with embeddings, morphology and syntax

Constraints on term structure

- Domain relevance and specialised corpora
- Vast search space in comparison to single words

BUT Regular term formation via compounding

- Prediction with embeddings, morphology and syntax

Term variation

brass plate company	compagnie écran
dummy company	entreprise boîte aux lettres
front company	filiale sans support matériel
letterbox company	société boîte aux lettres
money box company	société boîte à lettres
paper company	société coquille
shell company	société de façade
shell corporation	société fantôme
	société fictive

Take-home message

- Cross-lingual embeddings can be improved via cognates

Take-home message

- Cross-lingual embeddings can be improved via cognates
- They can be used in downstream tasks:
  POS tagging, NER or terminology extraction

Take-home message

- Cross-lingual embeddings can be improved via cognates
- They can be used in downstream tasks:
  POS tagging, NER or terminology extraction
- Problems with multi-word expressions and terms

Take-home message

- Cross-lingual embeddings can be improved via cognates
- They can be used in downstream tasks:
  POS tagging, NER or terminology extraction
- Problems with multi-word expressions and terms

Take-home message

- Cross-lingual embeddings can be improved via cognates
- They can be used in downstream tasks:
  POS tagging, NER or terminology extraction
- Problems with multi-word expressions and terms

Jelinek vs Church

Take-home message

- Cross-lingual embeddings can be improved via cognates
- They can be used in downstream tasks:
  POS tagging, NER or terminology extraction
- Problems with multi-word expressions and terms

Jelinek vs Church

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

Take-home message

- Cross-lingual embeddings can be improved via cognates
- They can be used in downstream tasks:
  POS tagging, NER or terminology extraction
- Problems with multi-word expressions and terms

Jelinek vs Church

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

Church: Fire everybody and buy more data

Take-home message

- Cross-lingual embeddings can be improved via cognates
- They can be used in downstream tasks:
  POS tagging, NER or terminology extraction
- Problems with multi-word expressions and terms

Jelinek vs Church

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

Church: Fire everybody and buy more data

- Share information across tasks and languages

Take-home message

- Cross-lingual embeddings can be improved via cognates
- They can be used in downstream tasks:
  POS tagging, NER or terminology extraction
- Problems with multi-word expressions and terms

Jelinek vs Church

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

Church: Fire everybody and buy more data

- Share information across tasks and languages
- Place for linguistics: what is shared?
  UD annotation or Term structure