Finding next of kin

Studies on related languages and dialects

Serge Sharoff

Centre for Translation Studies
University of Leeds

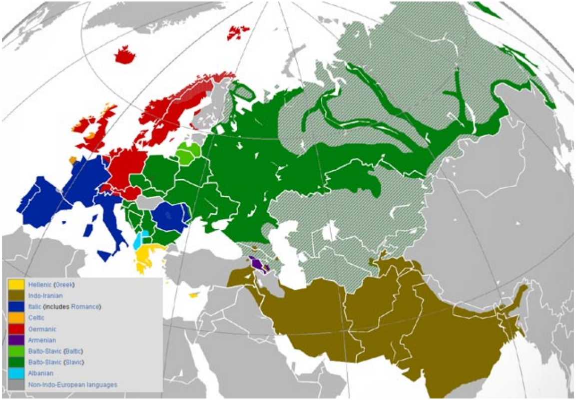
11 December 2022

Outline

1. Language adaptation
   - Rationale for language adaptation
   - My story about related languages
   - Learning shared representations
2. Detection of cognates
   - Weighted Levenshtein Distance
   - False friends vs cognates
3. WLD for contextual embeddings
   - Low data challenge
   - Improving contextual embeddings

Rationale: lack of resources

- In Ethnologue: 5,625 languages with > 1000 speakers
- 100 languages needed to cover 85% world's population
- 98-100. Balochi, Belarusian and Konkani, ≈ 7M speakers
- 40. Ukrainian, 30M native speakers (8. in Europe)



Relations between languages

UD: Roger Bacon (c1250) vs Joakim Nivre (c2015)

- 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
- BUT UD sets: 13K for Belarusian, 1.2M for Russian

My story on representations for related languages

Используйте команду Multiline, чтобы соединить двадцать два отрезка.
'Use$_{imp,pl}$ the Multiline command to connect twenty two lines$_{gen,sg}$'

- Multilingual Slavonic grammars (Bateman and Sharoff, 1998)
- Resources for reading Romanian via French (Ciobanu et al., 2006)
- POS taggers for Kannada via Telugu (Reddy and Sharoff, 2011); for Ukrainian via Russian (Babych and Sharoff, 2016)
- Language adaptation for MT Quality Estimation (Rios and Sharoff, 2016)

Learning representations for MTQE

- en: *A banner notification at the top of the screen indicates an issue.*
  ru: *Баннерное уведомление в верхней части экрана указывает на проблему.*
  pl: *Powiadomienie na pasku u góry ekranu wskazuje na problem.*
- We know which Russian MT output is good
  Polish MT output with similar features is likely to be good
- BUT we have different feature spaces between Polish and Russian

Self-Taught Learning (STL) for adapting feature spaces

1. Build a function for transforming data using *unlabelled* Russian and Polish data (MT without PE)
2. Learning a shared space using variational autoencoders
3. Train a prediction model on transformed Russian data
4. Apply the model to transformed Polish data

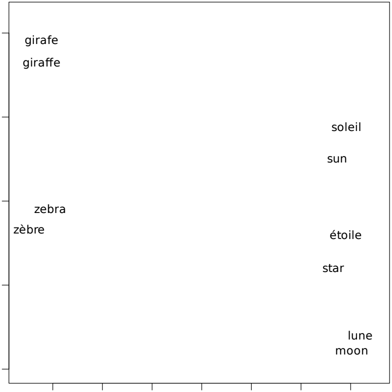
Experimental results (Rios and Sharoff, 2016)

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 Zero-shot: ru → 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 Zero-shot: es → cs/pl	MAE	0.23	0.22
	RSME	0.31	0.29
	Correlation	*0.11*	*0.09*

Cross-lingual word embeddings (word2vec)



- Vector space models (Rapp, 1995; Sharoff et al., 2006)
- SGD (Mikolov et al., 2013), CCA (Faruqui and Dyer, 2014), multivariate regression (Dinu et al., 2014), regression with orthogonalisation constraints (Artetxe et al., 2016)

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

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

- Also WLD works across charsets:

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

Integration of WLD into embeddings

- Cognates can be produced by:
  $score(s_e, s_f) = \alpha \cos(v_e, v_f) + (1 - \alpha) WLD(s_e, s_f)$

Integration of WLD into embeddings

- Cognates can be produced by:
  $score(s_e, s_f) = \alpha \cos(v_e, v_f) + (1 - \alpha) WLD(s_e, s_f)$
- Iterations for refining cross-lingual spaces:

Integration of WLD into embeddings

- Cognates can be produced by:
  $score(s_e, s_f) = \alpha \cos(v_e, v_f) + (1 - \alpha) WLD(s_e, s_f)$
- Iterations for refining cross-lingual spaces:
  1. Large dictionary of reliable cognates

Integration of WLD into embeddings

- Cognates can be produced by:
  $score(s_e, s_f) = \alpha \cos(v_e, v_f) + (1-\alpha) WLD(s_e, s_f)$
- Iterations for refining cross-lingual spaces:
  1. Large dictionary of reliable cognates
  2. Re-alignment of spaces using this dictionary

Integration of WLD into embeddings

- Cognates can be produced by:
  $score(s_e, s_f) = \alpha \cos(v_e, v_f) + (1 - \alpha) WLD(s_e, s_f)$
- Iterations for refining cross-lingual spaces:
  1. Large dictionary of reliable cognates
  2. Re-alignment of spaces using this dictionary
- Improving SOTA

State-of-the-art for *en-it* (Artetxe et al., 2016)	0601
Weighted Levenshtein Distance (Sharoff, 2020)	0.692

Integration of WLD into embeddings

- Cognates can be produced by:
  $score(s_e, s_f) = \alpha \cos(v_e, v_f) + (1 - \alpha) WLD(s_e, s_f)$
- Iterations for refining cross-lingual spaces:
  1. Large dictionary of reliable cognates
  2. Re-alignment of spaces using this dictionary
- Improving SOTA

State-of-the-art for *en-it* (Artetxe et al., 2016)	0601
Weighted Levenshtein Distance (Sharoff, 2020)	0.692

- Cross-lingual Panslavonic embeddings for BLI

	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

Integration of WLD into embeddings

- Cognates can be produced by:
  $score(s_e, s_f) = \alpha \cos(v_e, v_f) + (1 - \alpha) WLD(s_e, s_f)$
- Iterations for refining cross-lingual spaces:
  1. Large dictionary of reliable cognates
  2. Re-alignment of spaces using this dictionary
- Improving SOTA

State-of-the-art for *en-it* (Artetxe et al., 2016)	0601
Weighted Levenshtein Distance (Sharoff, 2020)	0.692

- Cross-lingual Panslavonic embeddings for BLI

	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

- Success in zero-shot downstream tasks: NER and POS tagging

False friends vs cognates

Cases of false friends

- *consistently* false friends:
  Mist in German='manure'
  bezcenny Polish='worthless' vs Czech='priceless'

False friends vs cognates

Cases of false friends

- *consistently* false friends:
  Mist in German='manure'
  bezcenny Polish='worthless' vs Czech='priceless'
- *partial* false friends:
  žena = 'wife' or 'woman' in a number of Slavonic languages
  Russian always = 'wife';

False friends vs cognates

Cases of false friends

- *consistently* false friends:
  Mist in German='manure'
  bezcenny Polish='worthless' vs Czech='priceless'
- *partial* false friends:
  žena = 'wife' or 'woman' in a number of Slavonic languages
  Russian always = 'wife';
- *actual* cognates with uncommon divergent senses
  żona Polish='wife', rarely ='woman'

False friends vs cognates

Cases of false friends

- *consistently* false friends:
  Mist in German='manure'
  bezcenny Polish='worthless' vs Czech='priceless'
- *partial* false friends:
  žena = 'wife' or 'woman' in a number of Slavonic languages
  Russian always = 'wife';
- *actual* cognates with uncommon divergent senses
  żona Polish='wife', rarely ='woman'

→ Disagreement between annotators about which friend is false (Fišer and Ljubešić, 2013)

Empirical investigation of false friends

- Monolingual embeddings reflect meaning
  $\rightarrow$ Similar embeddings for words with similar meanings
- WLD scores reflect word forms
  $\rightarrow$ Higher orthographic similarity for false friends
- Starting from "The False Friends of the Slavist"
  `https://en.wikibooks.org/w/index.php?oldid=3417664`
- Overall WLD helps . . .

RQ Does WLD hurt translation predictions for false friends?

Consistently false friends

Russian	Czech False	WLD	Cos	$\alpha W + C$
заход 'visit'	záchod 'toilet'	0.473	0.009	0.149
рок 'destiny'	rok 'year'	0.112	0.037	0.267
обход 'diversion'	obchod 'shop'	0.287	0.084	0.254
столица 'capital'	stolice 'chair'	0.248	0.106	0.280
заказ 'order'	zákaz 'prohibition'	0.417	0.131	0.253
урок 'lesson'	úrok 'interest'	0.289	0.131	0.288
дело 'business, case'	dělo 'cannon'	0.272	0.154	0.309
красный 'red'	krásný 'beautiful'	0.443	0.155	0.264
повесть 'novel'	pověst 'legend'	0.345	0.185	0.312
живот 'stomach'	život 'life'	0.219	0.197	0.354
родина 'homeland'	rodina 'family'	0.123	0.199	0.382
ел 'ate'	jel 'went'	0.351	0.235	0.346
век 'century'	věk 'age'	0.394	0.238	0.337
князь 'prince'	kněz 'priest'	0.489	0.261	0.329
враг 'enemy'	vrah 'murderer'	0.304	0.281	0.393

Possible cognates with divergencies

Russian	Czech False	WLD	Cos	$\alpha W + C$
скоро 'soon'	skoro 'almost'	0.132	0.245	0.413
злодей 'villain'	zloděj 'thief'	0.380	0.314	0.396
склеп 'crypt'	sklep 'cellar'	0.157	0.323	0.463
петроград 'Petrograd'	petrohrad	0.201	0.330	0.457
тыква 'pumpkin'	tykev 'melon'	0.531	0.411	0.426
словенский 'Slovenian'	slovenský 'Slovak'	0.321	0.415	0.486
стул 'chair'	stůl 'table'	0.277	0.419	0.501
палец 'finger'	palec 'thumb'	0.135	0.428	0.546
постель 'bed, linen'	postel 'bed'	0.230	0.490	0.566
запах 'smell'	zápach 'foul smell'	0.461	0.509	0.517
овощи 'vegetables'	ovoce 'fruits'	0.417	0.518	0.535
угол 'angle,corner'	úhel 'angle'	0.617	0.611	0.549
слышать 'to hear, to sense'	slyšet 'to hear'	0.468	0.625	0.600

Да и кстати третий день не слышу запахи и вкус. Что кофе пью, что воду один хрен. Даже духи не слышу. (I don't sense 'hear' smell and taste for the third day in a row. I don't even sense 'hear' perfume.)

Best translations: false friends

Russian	Czech	False W+C	Best Cos		Best $\alpha W + C$	
заход	záchod	0.149	mezipřistání	0.411	hod	0.359
рок	rok	0.267	punkrockové	0.658	rock	0.580
обход	obchod	0.254	obcházení	0.467	obcházení	0.429
столица	stolice	0.280	*město*	0.489	*město*	0.423
заказ	zákaz	0.253	zakázka	0.608	zakázka	0.562
урок	úrok	0.288	školník	0.383	školník	0.368
дело	dělo	0.309	obvinění	0.361	delikt	0.361
красный	krásný	0.264	červený	0.599	červený	0.503
выход	východ	0.273	výstup	0.404	přechod	0.384
повесть	pověst	0.312	povídka	0.698	povídka	0.640
живот	život	0.354	nohy	0.542	nohy	0.444
родина	rodina	0.382	domovina	0.447	domovina	0.457
ел	jel	0.346	vypil	0.416	jedl	0.428
век	věk	0.337	stol	0.454	století	0.386
князь	kněz	0.329	kníže	0.703	kníže	0.635
враг	vrah	0.393	nepřítel	0.624	nepřítel	0.486

Best translations: divergent cognates

Russian	Czech False	W+C	Best Cos		Best $\alpha W + C$	
скоро	skoro	0.413	brzy	0.595	brzo	0.508
злодей	zloděj	0.396	padouch	0.513	zloduch	0.474
склеп	sklep	0.463	hrob	0.583	hrob	0.475
петроград	petrohrad	0.457	bolševiků	0.390	petrohrad	0.457
тыква	tykev	0.426	kdoule	0.463	tykve	0.436
словенский	slovenský	0.486	chorvatský	0.703	slovinský	0.635
стул	stůl	0.501	stůl	0.419	stůl	0.501
палец	palec	0.546	prst	0.552	palec	0.546
постель	postel	0.566	postel	0.490	postel	0.566
запах	zápach	0.517	vůně	0.521	*zápach*	0.517
овощи	ovoce	0.535	zeleniny	0.633	ovoce	0.535
угол	úhel	0.549	úhel	0.611	úhel	0.549
слышать	slyšet	0.600	slyšet	0.625	slyšet	0.600

Contextual embeddings for ambiguities

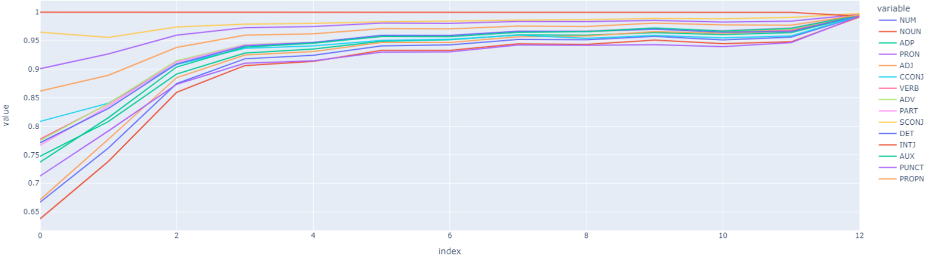
Multilingual models: mBERT, XLM-Roberta

- *I put my glass on the kitchen table.* vs
  The table lists all the products.
- Shared parameters *Consult the .. of beam sizes below* vs
  Vous pouvez consulter le .. des rémunérations des professeurs
- BUT we cannot use WLD to align *word* spaces, we need to fine-tune transformer parameters

Uneven data distribution for training mBERT

	# Texts	# Tokens	#L10
Be	75345	22857203	118639
Bg	160884	60643545	191724
Pl	807576	242688746	524924
Ru	1170755	459637736	988900
Uk	553255	193180812	527722

Representations over layers



Average cosine similarity of ru and uk POS classes for parallel sentences

Matching WECHSEL and WLD

- WECHSEL (Minixhofer et al., 2022): Improving 0-layer embeddings e^r of recipient languages by initializations aligned with embeddings e^d of the donor language:

$$e_x^r = \frac{\sum_{y \in \mathcal{J}_x} \exp\left(s_{x,y}/\tau\right) \cdot e_y^d}{\sum_{y' \in \mathcal{J}_x} \exp\left(s_{x,y'}/\tau\right)}$$

where $\mathcal{J}_x$ is the set of k neighbouring subwords in the donor language; τ is temperature, $s_{x,y}$ is the cosine similarity.

Matching WECHSEL and WLD

- WECHSEL (Minixhofer et al., 2022): Improving 0-layer embeddings e^r of recipient languages by initializations aligned with embeddings e^d of the donor language:

$$e_x^r = \frac{\sum_{y \in \mathcal{J}_x} \exp\left(s_{x,y}/\tau\right) \cdot e_y^d}{\sum_{y' \in \mathcal{J}_x} \exp\left(s_{x,y'}/\tau\right)}$$

  where $\mathcal{J}_x$ is the set of k neighbouring subwords in the donor language; τ is temperature, $s_{x,y}$ is the cosine similarity.
- Our improvements:

Matching WECHSEL and WLD

- WECHSEL (Minixhofer et al., 2022): Improving 0-layer embeddings e^r of recipient languages by initializations aligned with embeddings e^d of the donor language:

$$e^r_x = \frac{\sum_{y \in \mathcal{J}_x} \exp\left(s_{x,y}/\tau\right) \cdot e^d_y}{\sum_{y' \in \mathcal{J}_x} \exp\left(s_{x,y'}/\tau\right)}$$

  where $\mathcal{J}_x$ is the set of k neighbouring subwords in the donor language; τ is temperature, $s_{x,y}$ is the cosine similarity.
- Our improvements:
  1. Better initial word embeddings with WLD;

Matching WECHSEL and WLD

- WECHSEL (Minixhofer et al., 2022): Improving 0-layer embeddings e^r of recipient languages by initializations aligned with embeddings e^d of the donor language:

$$e^r_x = \frac{\sum_{y \in \mathcal{J}_x} \exp\left(s_{x,y}/\tau\right) \cdot e^d_y}{\sum_{y' \in \mathcal{J}_x} \exp\left(s_{x,y'}/\tau\right)}$$

  where $\mathcal{J}_x$ is the set of k neighbouring subwords in the donor language; τ is temperature, $s_{x,y}$ is the cosine similarity.
- Our improvements:
  1. Better initial word embeddings with WLD;
  2. Improving the nearest neigbours by replacing $s_{x,y}$ with:

$$s(x, y) = \alpha \cos(x, y) + (1 - \alpha) WLD(x, y)$$

Applying it to Named Entity Recognition

Few-shot testing with our gold dataset (W) and SlavicNER (S)

Polish as L_{Donor}	bg		cs		ru		sl		uk		be
	W	S	W	S	W	S	W	S	W	S	W
Baseline	0.77	0.82	0.84	0.87	0.76	0.82	0.80	0.83	0.78	0.81	0.77
Wechsel	0.83	0.83	0.89	0.88	0.81	0.83	**0.85**	0.84	0.84	0.84	0.82
Wechsel+WLD	**0.85**	**0.84**	**0.90**	**0.91**	**0.84**	**0.85**	**0.85**	**0.86**	**0.86**	**0.86**	**0.84**

Russian as L_{Donor}	bg		cs		pl		sl		uk		be
	W	S	W	S	W	S	W	S	W	S	W
Baseline	0.79	0.80	0.78	0.84	0.79	0.85	0.77	0.82	0.83	0.79	0.81
Wechsel	0.81	0.80	0.85	0.86	0.83	0.85	0.82	0.83	0.84	0.84	0.84
Wechsel+WLD	**0.82**	**0.81**	**0.86**	**0.88**	**0.84**	**0.86**	**0.83**	**0.84**	**0.86**	**0.87**	**0.85**

Take-home message

- Improved embeddings via Weighted Levenshtein Distance
- They can be used in downstream tasks:
  POS tagging, NER or terminology extraction
- False friends do not get into the way (mostly)
- Morphology is preserved in transformation
- Challenges in building source embeddings with very little data
- Alignment of cross-lingual contextual embeddings for related languages can also be improved

References I

Artetxe, M., Labaka, G., and Agirre, E. (2016).
Learning principled bilingual mappings of word embeddings while preserving monolingual invariance.
In *Proc EMNLP*, Austin, Texas.

Babych, B. and Sharoff, S. (2016).
Ukrainian part-of-speech tagger for hybrid MT: Rapid induction of morphological disambiguation resources from a closely related language.
In *Proc Fifth Workshop on Hybrid Approaches to Translation (HyTra)*.

Bateman, J. A. and Sharoff, S. (1998).
Multilingual grammars and multilingual lexicons for multilingual text generation.
In *Multilinguality in the lexicon II*, ECAI'98 Workshop, pages 1–8, Brighton, UK. European Conference on Artificial Intelligence.

References II

Ciobanu, D., Hartley, A., and Sharoff, S. (2006).
Using richly annotated trilingual language resources for acquiring reading skills in a foreign language.
In *Proc Fifth Language Resources and Evaluation Conference, LREC*, Genoa.

Dinu, G., Lazaridou, A., and Baroni, M. (2014).
Improving zero-shot learning by mitigating the hubness problem.
arXiv preprint arXiv:1412.6568.

Faruqui, M. and Dyer, C. (2014).
Improving vector space word representations using multilingual correlation.
In *Proc EACL*, pages 462–471, Gothenburg, Sweden.

References III

Fišer, D. and Ljubešić, N. (2013).
Best friends or just faking it? Corpus-based extraction of Slovene-Croatian translation equivalents and false friends.
Slovenščina 2.0, 1.

Mikolov, T., Le, Q. V., and Sutskever, I. (2013).
Exploiting similarities among languages for machine translation.
arXiv preprint arXiv:1309.4168.

Minixhofer, B., Paischer, F., and Rekabsaz, N. (2022).
WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models.
In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3992–4006, Seattle, United States. Association for Computational Linguistics.

References IV

Rapp, R. (1995).
Identifying word translations in non-parallel texts.
In *Proc. of the 33rd ACL*, pages 320–322, Cambridge, MA.

Reddy, S. and Sharoff, S. (2011).
Cross language POS taggers (and other tools) for Indian languages: An experiment with Kannada using Telugu resources.
In *Proc IJCNLP'11*, Chiang Mai, Thailand.

Rios, M. and Sharoff, S. (2016).
Language adaptation for extending post-editing estimates for closely related languages.
The Prague Bulletin of Mathematical Linguistics, 106:181–192.

Sharoff, S. (2020).
Finding next of kin: Cross-lingual embedding spaces for related languages.
Journal of Natural Language Engineering, 26:163–182.

References V

Sharoff, S., Babych, B., and Hartley, A. (2006).
Using comparable corpora to solve problems difficult for human translators.
In *Proc International Confenrence on Computational Linguistics and Association of Computational Linguistics, COLING-ACL 2006*, pages 739–746, Sydney.