Fine-tuning Cross-lingual Language Models for Lexical Semantic Change Detection

Nikolay Arefyev

Plan:

1. RuShiftEval-2021 shared task on Lexical Semantic Change Detection (LSCD) for the Russian language
2. DeepMistake solution ← 2nd best, outperformed the winner in the post-eval
   XLM-R fine-tuned for the WiC task
3. GlossReader solution ← the winning solution
   XLM-R fine-tuned for the gloss-based WSD task
4. (briefly) Results in LSCDiscovery-2022 shared task on Lexical Semantic Change Detection (LSCD) for the Spanish language

Co-authors and publications

LIORI at SemEval-2021 Task 2: Span Prediction and Binary Classification approaches to Word-in-Context Disambiguation

Adis Davletov
RANEPA, Moscow, Russia
Lomonosov Moscow State University, Moscow, Russia
davletov-aa@ranepa.ru
Nikolay Arefyev
Lomonosov Moscow State University, Moscow, Russia
Samsung Research Center Russia, Moscow, Russia
HSE University, Moscow, Russia
nick.arefyev@gmail.com
Denis Gordeev
RANEPA, Moscow, Russia
gordeev-di@ranepa.ru
Alexey Rey
RANEPA, Moscow, Russia
rey-ai@ranepa.ru

SemEval-2021: MCL-WiC (Multilingual and Cross-lingual Word-in-Context)

GlossReader at SemEval-2021 Task 2: Reading Definitions Improves Contextualized Word Embeddings

Maxim Rachinskiy[◇] **and Nikolay Arefyev**[△▽◇]
[◇]HSE University, Moscow, Russia
[△]Samsung Research Center Russia, Moscow, Russia
[▽]Lomonosov Moscow State University, Moscow, Russia
myurachinskiy@edu.hse.ru narefjev@cs.msu.ru

Co-authors and publications

DeepMistake: Which Senses are Hard to Distinguish for a Word-in-Context Model

Nikolay Arefyev[◇△▽] Moscow, Russia nick.arefyev@gmail.com
Maksim Fedoseev[◇] Moscow, Russia maxim.fedoseev13@gmail.com
Vitaly Protasov[○] Moscow, Russia vitaly.protasov@skoltech.ru
Daniil Homskiy[◇] Moscow, Russia homdanil123@gmail.com
Adis Davletov[◇▷] Moscow, Russia dev.davletov@gmail.com
Alexander Panchenko[○] Moscow, Russia A.Panchenko@skoltech.ru

[◇]Lomonosov Moscow State University
[△]Samsung Research Center Russia
[▽]HSE University
[○]Skolkovo Institute of Science and Technology
[▷]RANEPA

RuShiftEval-2021:
LSCD for the Russian language

Zero-shot Cross-lingual Transfer of a Gloss Language Model for Semantic Change Detection

Maxim Rachinskiy
HSE University
Moscow, Russia
myurachinskiy@edu.hse.ru

Nikolay Arefyev
Samsung Research Center Russia
Lomonosov Moscow State University
HSE University
Moscow, Russia
narefjev@cs.msu.ru

DeepMistake at LSCDiscovery: Can a Multilingual Word-in-Context Model Replace Human Annotators?

Daniil Homskiy[▽] **Nikolay Arefyev**[◇,▽,△]

[▽]Lomonosov Moscow State University / Moscow, Russia
[△]National Research University Higher School of Economics / Moscow, Russia
[◇]Samsung Research Center Russia / Moscow, Russia
homdanil123@gmail.com, nick.arefyev@gmail.com

LSCDiscovery-2022:
LSCD for the Spanish language

GlossReader at LSCDiscovery: Train to Select a Proper Gloss in English – Discover Lexical Semantic Change in Spanish

Maxim Rachinskiy[△] **Nikolay Arefyev**[◇,▽,△]

[△]National Research University Higher School of Economics / Moscow, Russia
[◇]Samsung Research Center Russia / Moscow, Russia
[▽]Lomonosov Moscow State University / Moscow, Russia
myurachinskiy@edu.hse.ru, nick.arefyev@gmail.com

Plan:

1. **RuShiftEval-2021 shared task on Lexical Semantic Change Detection (LSCD) for the Russian language**
2. DeepMistake solution ← 2nd best, outperformed the winner in the post-eval
   XLM-R fine-tuned for the WiC task
3. GlossReader solution ← the winning solution
   XLM-R fine-tuned for the gloss-based WSD task

4. (briefly) Results in LSCDiscovery-2022 shared task on Lexical Semantic Change Detection (LSCD) for the Spanish language

RuShiftEval-2021 shared task: LCSD for the Russian language

Andrey Kutuzov & Lidia Pivovarova. RuShiftEval: a shared task on semantic shift detection for Russian. 2021
dataset collection and annotation, and evaluation follows DURel:
Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. Diachronic usage relatedness (DURel): A framework for the annotation of lexical semantic change. 2018

Input: 99 test words; 3 time periods.
Task: for each pair of periods, order test words according to their gold COMPARE scores.
Evaluation metric: Spearman's rank correlation between predicted and gold word scores (*word Spearman*)

Corpora: *Russian National Corpus (RNC):*

#	Period	Dates	Volume
p1	pre-Soviet	1700-1916	0.9 GB
p2	Soviet	1918-1990	1.1 GB
p3	post-Soviet	1991-2016	1.1 GB

Target word	p12	p23	p13
	gold COMPARE score		
тачка (*car* / *cart*)	*3.4*	*1.9*	*1.9*
сверстник (*agemate*)	*3.9*	*3.9*	*3.8*

1.0 — strong semantic shift
4.0 — no semantic shift

RuShiftEval-2021: dataset annotation

Pre-Soviet corpus

Тачка *(car/cart)*

Post-Soviet corpus

Sampling sentences

Sampling sentences

sentence 1	sentence 2
Тачку они взяли у лавочника, рядом. (They took cart from the shopkeeper, nearby)	Ты на машине следуй за моей **тачкой**. (You in the car follow my car)
...	...

30 pairs

Human annotation of **COMPARE pairs**

WiC Model annotation

cart

тачка

car

sentence 1	sentence 2	annot.
Тачку они взяли у лавочника, рядом.	Ты на машине следуй за моей **тачкой**.	1
...	...	...

the COMPARE score calculation (averaging over sentence pairs)

1.899

RuShiftEval-2021: RuSemShift as a training/development set

Rodina Julia, Kutuzov Andrey. RuSemShift: a dataset of historical lexical semantic change in Russian. 2020
dataset collection and annotation, and evaluation follows DURel:
Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. Diachronic usage relatedness (DURel): A framework for the annotation of lexical semantic change. 2018

Data: 2 pairs of time periods (p12, p23), 70 words for each (changed + random); max 60 pairs of sentences sampled per word (min 9).
Annotations: 5 annotators per pair, 1-4 scale + "cannot decide", crowdsourcing (different annotators for diff. pairs)

We made 50% / 50% lexical split: dev1 - 35 words from p12, dev2 - 35 words form p23, train - 70 words mixed

sent 1	sent 2	group	anns	Mean
Она остановила **машину** не совсем ладно, мотор заглох.	Эмгебистская легковая **машина** исчезает за углом, а я остаюсь на мостовой с большим плохо связанным узлом в руках.	LATER	4,4,4,4,4	4.0
Богадельня, или исправительный дом, ночлежка.	Настроит **богаделен**, ещё церковь выстроит, замостит и выровняет улицы.	COMPARE	3,4,3,4,3	3.43
Сафонов подошёл к Максимовичу и протянул ему руку: -- Вы -- **молодец!**	Михаил Юрьевич не безродный, не нищий, он в славе и собой **молодец.**	LATER	2,1,1,4,4	2.4

RuShiftEval-2021: official results

	Team	RuSemShift1	RuSemShift2	RuSemShift3	Mean	Type
1	**GlossReader**	0.781	**0.803**	**0.822**	**0.802**	token
2	**DeepMistake**	**0.798**	0.773	0.803	0.791	token
3	vanyatko	0.678	0.746	0.737	0.720	token
4	**aryzhova**	0.469	0.450	0.453	0.457	token
5	Discovery	0.455	0.410	0.494	0.453	token
6	**UWB**	0.362	0.354	0.533	0.417	type
7	dschlechtweg	0.419	0.373	0.383	0.392	type
8	jenskaiser	0.430	0.310	0.406	0.382	token
9	**SBX-HY**	0.388	0.281	0.439	0.369	type
	Baseline	0.314	0.302	0.381	0.332	type
10	svart	0.163	0.223	0.401	0.262	type
11	**BykovDmitrii**	0.274	0.202	0.307	0.261	token
12	fdzr	0.217	0.251	0.065	0.178	type

automate pairwise ann., *ft. XLM-R / ruBERT*

automate pairwise ann., *orig. ruBERT / ELMo*

fastText + CCA + cos

SGNS + regression

SGNS, Temporal ref. / WI

Temporal referencing

Table 1: Evaluation phase leaderboard (Spearman rank correlations). The Type column shows the type of the used distributional embeddings.

Table from **Andrey Kutuzov & Lidia Pivovarova. RuShiftEval: a shared task on semantic shift detection for Russian, 2021**

Plan:

1. RuShiftEval-2021 shared task on Lexical Semantic Change Detection (LSCD) for the Russian language
2. **DeepMistake solution** ← 2nd best, outperformed the winner in the post-eval
   XLM-R fine-tuned for the WiC task
3. GlossReader solution ← the winning solution
   XLM-R fine-tuned for the gloss-based WSD task
4. (briefly) Results in LSCDiscovery-2022 shared task on Lexical Semantic Change Detection (LSCD) for the Spanish language

DeepMistake: eval. and post-eval. results

Link to code: https://github.com/Daniil153/RuShiftEval ← after polishing the code, the results improved a little bit further

Our post-evaluation improvements

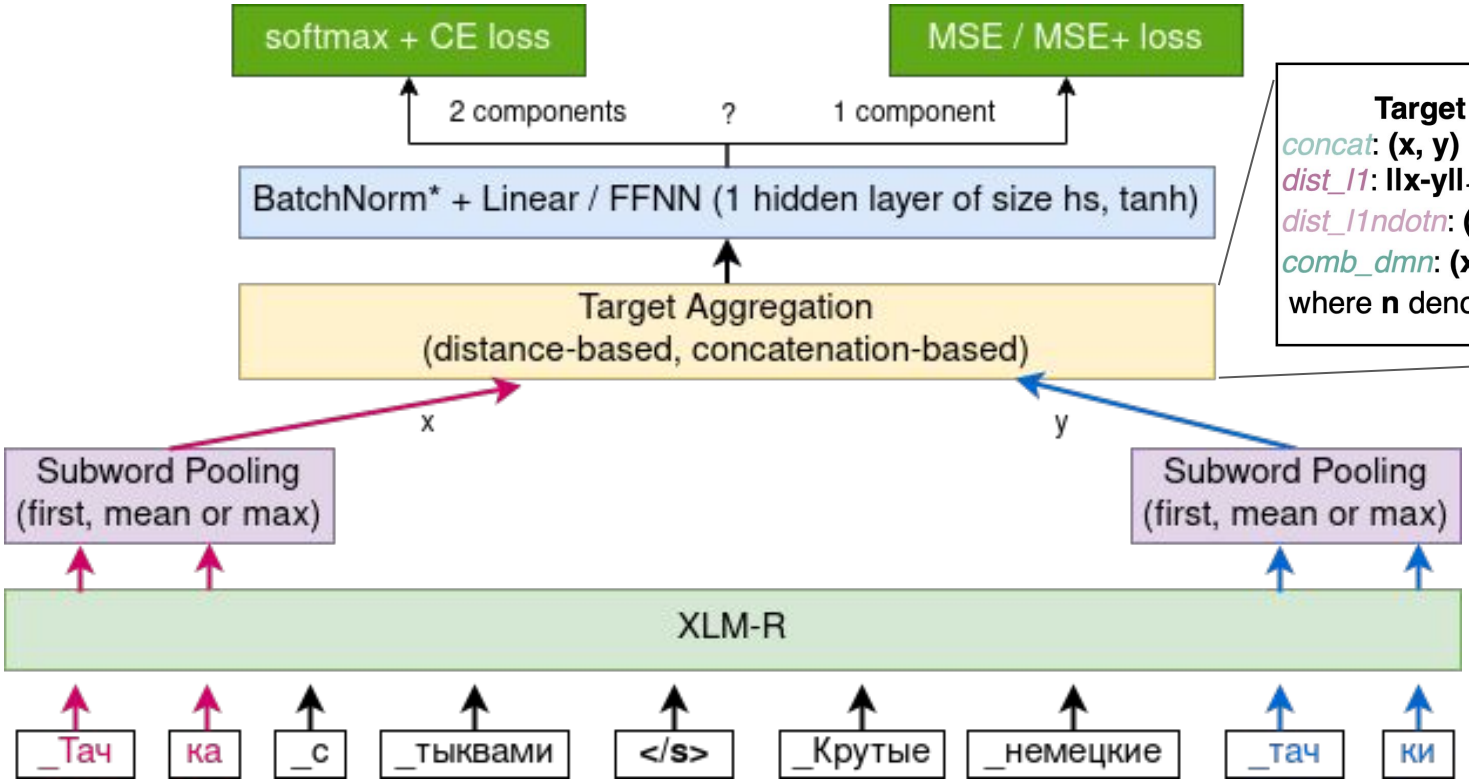
*DeepMistake	**0.863**	**0.854**	**0.834**	**0.85**
DeepMistake	**0.825**	**0.821**	**0.823**	**0.823**

* This result is absent in the paper, it was achieved after the paper was published with mean+dist_l1ndotn-hs0 on $MCL^{nen\text{-}acc}{}_{CE} \rightarrow RSS^{dev2\text{-}sentSpear}{}_{CE}$, Mean

	Team	RuSemShift1	RuSemShift2	RuSemShift3	Mean
1	**GlossReader**	0.781	**0.803**	**0.822**	**0.802**
2	**DeepMistake**	**0.798**	0.773	0.803	0.791
3	vanyatko	0.678	0.746	0.737	0.720
4	**aryzhova**	0.469	0.450	0.453	0.457
5	Discovery	0.455	0.410	0.494	0.453
6	**UWB**	0.362	0.354	0.533	0.417
7	dschlechtweg	0.419	0.373	0.383	0.392
8	jenskaiser	0.430	0.310	0.406	0.382
9	**SBX-HY**	0.388	0.281	0.439	0.369

Table from **Andrey Kutuzov & Lidia Pivovarova. RuShiftEval: a shared task on semantic shift detection for Russian, 2021**

WiC system: architecture



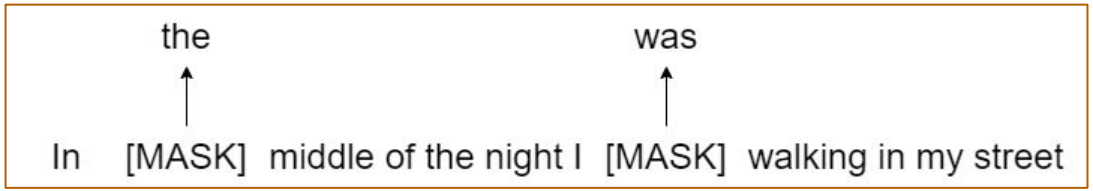
Target aggregation options

concat: **(x, y)** [size: 2 * hs],
dist_l1: **$\|x-y\|_1$** [size: 1],
dist_l1ndotn: **$(\|\bar{x}_n-\bar{y}_n\|_1, \langle \bar{x}_n, \bar{y}_n \rangle)$** [size: 2],
comb_dmn: **$(x - y, \bar{x}_n * \bar{y}_n)$** [size: 2 * hs],
where **n** denotes normalized vectors

* Batch Normalization was used only in the post-competition experiments.

Masked Language Models

MLM objective:



MLM is proposed in:
Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL 2019

Similar ideas of pre-training a neural network to guess a word by its left & right context:

- word2vec CBOW: sum embeddings of k previous and k next words, predict the center word with a linear classifier
  Mikolov et al. Efficient Estimation of Word Representations in Vector Space, 2013

- context2vec: encode left & right context separately with LSTMs, combine with a MLP, predict the center word with a linear classifier
  Melamud et al. context2vec: Learning Generic Context Embedding with Bidirectional LSTM, 2016

- pre-training for WSI: retrieve examples of an ambiguous word, replace it with CENTERWORD, use Transformer-based encoder-decoder (MT) architecture to restore it:

  They were standing on the CENTERWORD of the Volga river → bank

  Struyanskiy & Arefyev. Neural Networks with Attention for Word Sense Induction // AIST'2018 (July 2018, the first draft of the BERT paper appeared in October 2018)

A real example from the IMDB Movie Reviews dataset

This is a really well made movie . Sum it ra B have has always made sensible cinema and this is my favourite film by her . This movie should have won the National Award and would have been my pick to represent India at the Oscars . It is at least a thousand times better than ' Sh a aws ', which is going to the Oscars , from India , this year .< br / >< br /> It is such a pity that the information about this (and all other Indian movies) on IM Db is lacking and sometimes even wrong . Sad ash iv Am rap ur kar played a very important character in this movie and he is not even credited on these pages .

XLM-R MLM pre-training:

1) text is split into tokens (subwords);
2) 15% of positions are sampled **from the Uniform distribution** as target positions, model is asked to guess tokens on these positions (cross-entropy loss is calculated on these positions only);
3) 80% of tokens on target positions are replaced with [MASK], 10% are replaced with random tokens, 10% are left untouched.

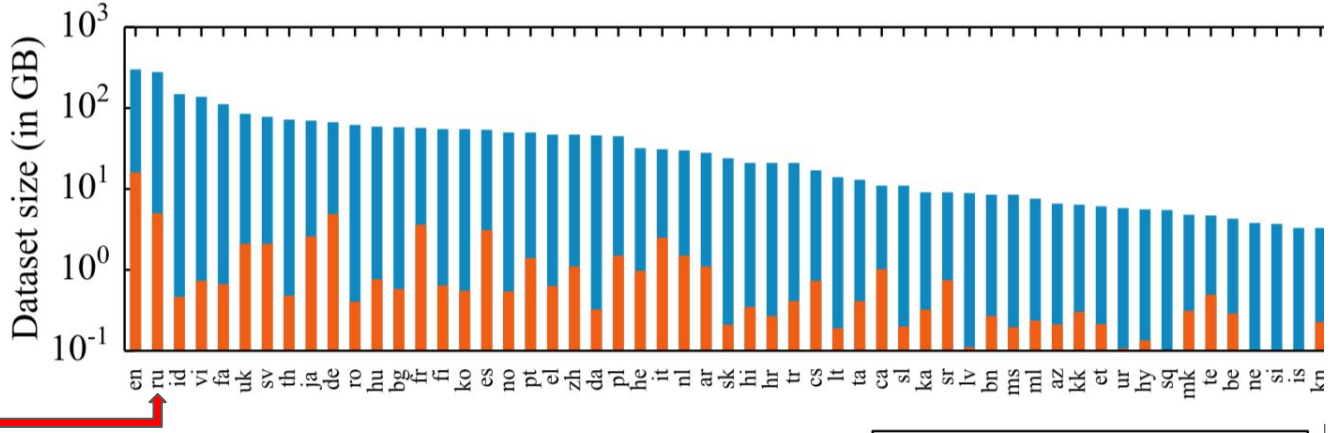
XLM-R Masked Language Model

From BERT to XLM-R:

- BERT (16 GB, en, Wikipedia + BookCorpus)
- RoBERTa (160 GB, en, Wikipedia + BookCorpus + others)
  - more data, longer training;
  - small technical tricks (dynamic masking, no NSP pretrain, byte-level BPE)
- XLM-R[oBERTa] (2.5 TB, 100 languages, Wikipedia + CommonCrawl)
  - pre-trained on 500 V100 GPUs for ~1 week

XLM-R: 1) good initial weights for text encoders in any of 100 languages for almost any NLP task
2) zero-shot cross-lingual transfer ability

mBERT is trained only on Wikipedias



Russian is 2nd largest!

Figure from Alexis Conneau et al. Unsupervised Cross-lingual Representation Learning at Scale, 2020.

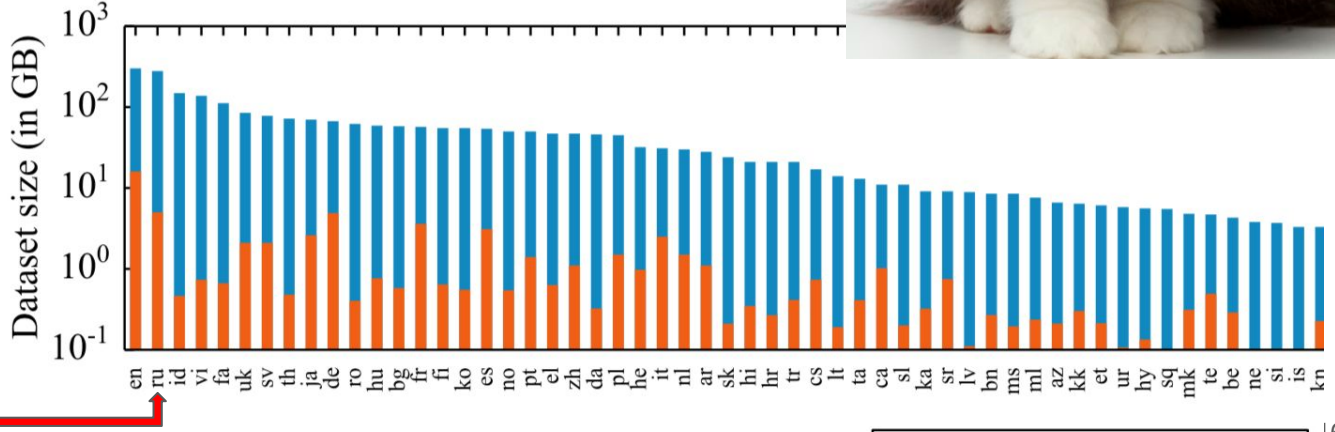
XLM-R Masked Language Model

From BERT to XLM-R:

- BERT (16 GB, en, Wikipedia + BookCorpus)
- RoBERTa (160 GB, en, Wikipedia + BookCorpus + others)
  - more data, longer training;
  - small technical tricks (dynamic masking, no NSP pretrain, byte-level BPE)
- XLM-R[oBERTa] (2.5 TB, 100 languages, Wikipedia + CommonCra
  - pre-trained on 500 V100 GPUs for ~1 week

XLM-R: 1) good initial weights for text encoders in any of 100 languages
2) zero-shot cross-lingual transfer ability

mBERT is trained only on Wikipedias



Russian is 2nd largest!

Figure from Alexis Conneau et al. Unsupervised Cross-lingual Representation Learning at Scale, 2020.

WiC system: training / development data

Dataset	*Training ex.*	*dev1 (p12) / dev2 (p23)*
MCL-WiC (binary)	10.9k (incl. 8k/718/718/718/718 en/ru/zh/ar/fr)	
MCL-WiC en-en	8k	
MCL-WiC ru-ru	1.7k (incl. test set)	
RuSemShift (1-4 scores)	3.9k COMPARE+EARLIER+LATER pairs for 70 words, p12+p23	~600*2 COMPARE pairs for ~35*2 words

Dataset	*Sentence #1*	*Sentence #2*	*Target*
MCL-WiC	On ne peut donc examiner les **questions** de développement... (*So we can't look at development **issues**…*)	En réponse aux **questions** posées sur les visites… (*In response to **questions** asked about the visits ...*)	F
MCL-WiC en-en	He **declared** that the deaths ... would not be in vain...	There she **declared** that she considered the new proposal...	T
MCL-WiC ru-ru	Над Корсаком **нависает** угроза ареста… (*The threat of arrest **hangs** over Korsak...*)	Над залом **нависает** гигантский купол… (*A giant dome **hangs** over the hall...*)	F
RuSemShift	...не на ту **орбиту** забрался. (...climbed on the wrong **orbit**)	Глаза выкатывались из **орбит**. (Eyes rolled out of their **orbits**.)	1.0

WiC system: fine-tuning schemes

Train#1	Loss#1	Train#2	Loss#2
MCL-WiC	CE	-	
MCL-WiC en-en	CE	-	
MCL-WiC ru-ru	CE	-	
RuSemShift	MSE	-	
MCL-WiC	CE	MCL-WiC ru-ru	CE
MCL-WiC	CE	RuSemShift	MSE or CE
MCL-WiC	CE	RuSemShift + MCL-WiC ru-ru	MSE+ or CE

Table: ft. schemes

MSE+:

- for real-value — MSE
- for binary targets:

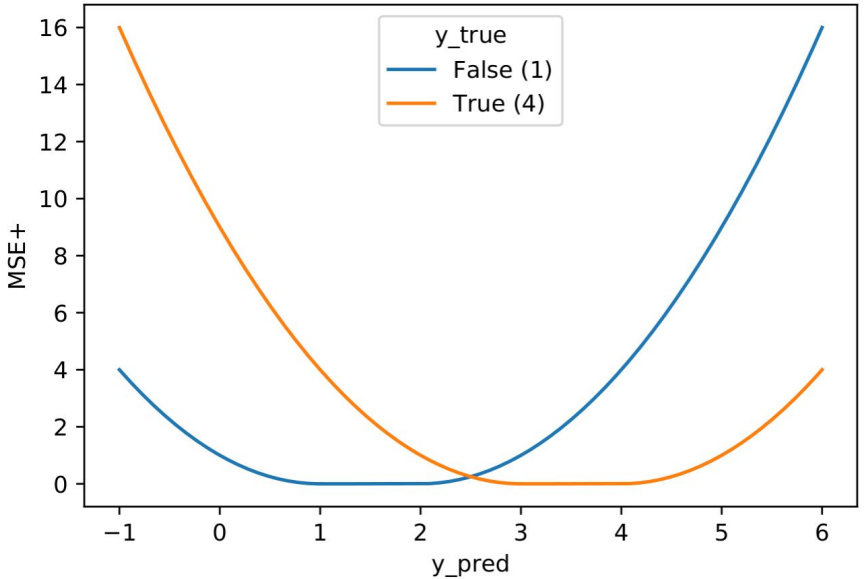


Figure: MSE+ for binary targets

Optimizer: *AdamW*
Learning rate: *1e-05*
Linear warmup: *5%*
Weight decay: *0.1*
Early stopping:

- MCL-WiC — *en-acc (English dev set)* or *nen-acc* (other languages in dev set)
- RuSemShift — *wordSpear* or *sentSpear (Spearman correlation between gold and predicted scores for all instances)*

LSCD solution based on WiC

Pre-Soviet corpus

Тачка *(car/cart)*

Post-Soviet corpus

Retrieve examples from the diachronic corpus. Lemmatization(https://github.com/Koziev/rulemma)

Filter examples: remove k% of the shortest and the longest examples, remove examples where the target word is the first or the last one

Sample pairs of examples of type COMPARE

Scores for sentence pairs: the WiC model

Scores for words: *Mean, IsoReg, LinReg*

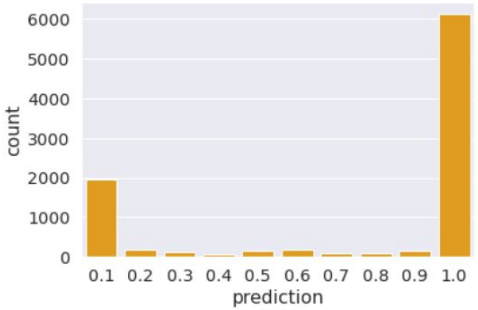
Predicted score: **3.388**

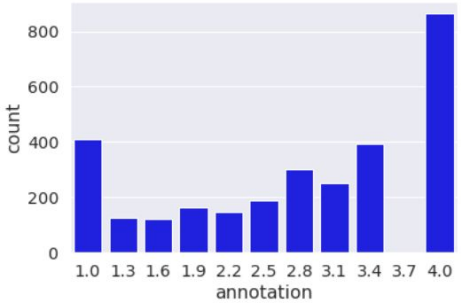
Mean. The mean of the scores for pairs.

IsoReg

Mean

LinReg. Linear Regression on the mean and quartiles of the scores for pairs.





Results for sampling and sentence filtering

The best result then we deleted 40% of the shortest and 40% of the longest sentences

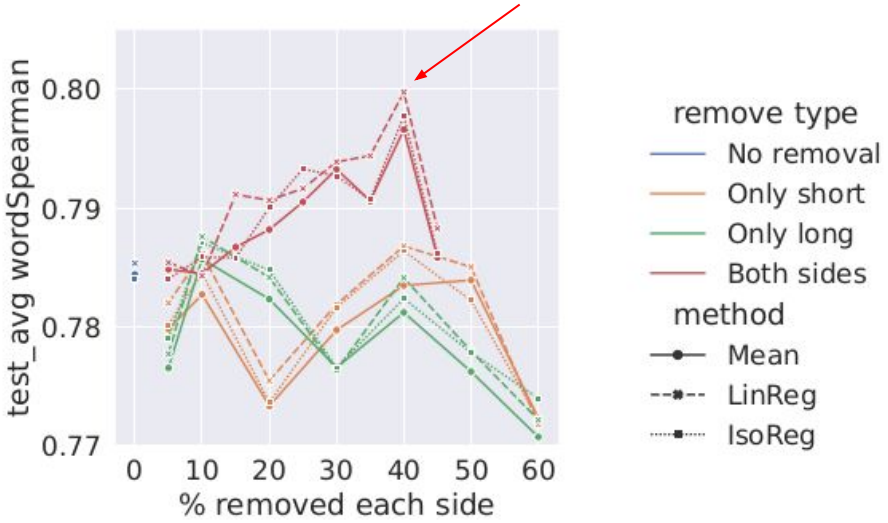


Figure 1

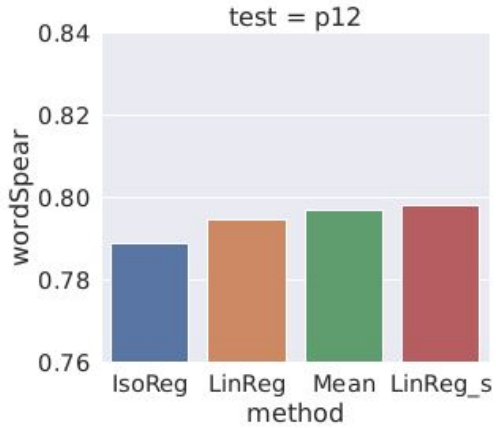


Figure 2

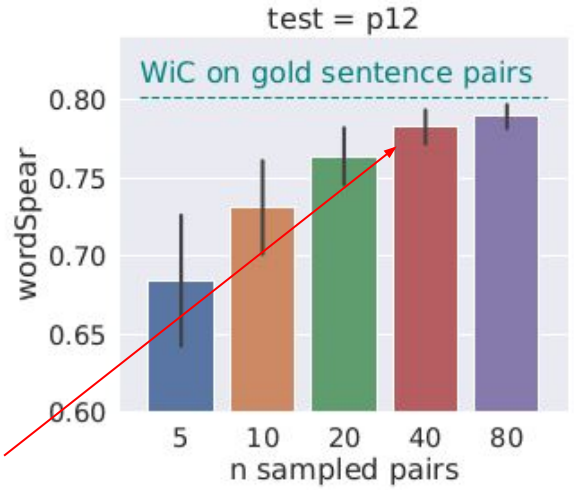


Figure 3

If we sample less than 40 pairs, then the quality of the model is poor and standard deviation(error bars) is high

Evaluation and post-evaluation results (100 sampled sent. pairs per word & periods)

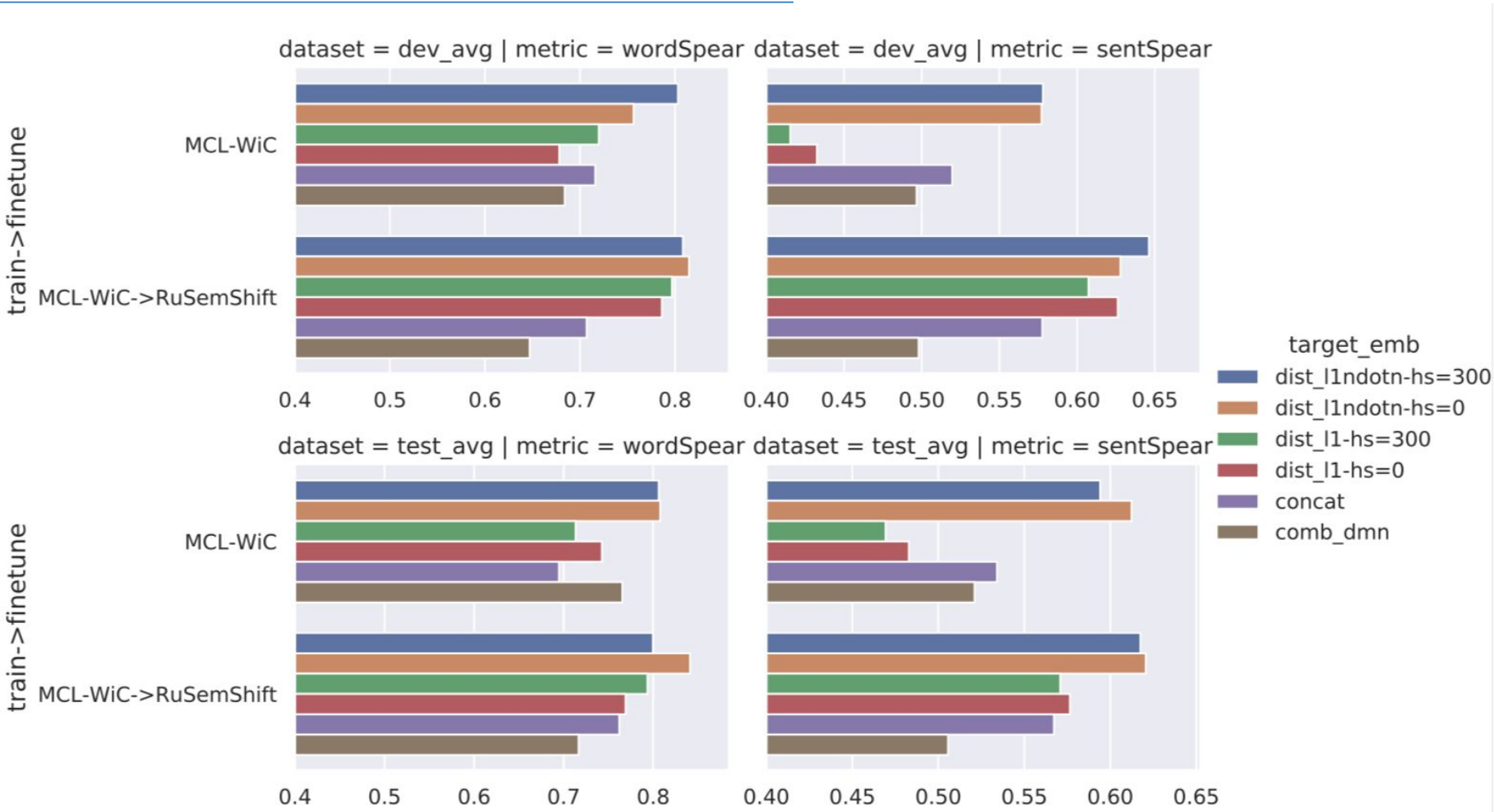
Method/Team	Avg	p12	p23	p13
Best results of other teams				
GlossReader (1st best result)	**0.802**	0.781	**0.803**	**0.822**
vanyatko (3rd best result)	0.720	0.678	0.746	0.737
Our submissions: team *DeepMistake* (2nd best result)				
first+concat on MCL_{CE}^{en-acc} →$\text{RSS}_{MSE}^{dev2-sentSpear}$ (***M1***), LinReg	**0.791**	**0.798**	0.773	0.803
M1, Mean	0.789	0.794	0.773	0.799
M1, IsoReg	0.789	0.793	0.775	0.798
p12, p13: *M2*; *p23*: first+concat on $\text{MCL}_{CE}^{nen-acc}$ →$\text{RSS+ruMCL}_{CE}^{dev1-sentSpear}$, IsoReg	0.785	0.773	**0.802**	0.780
mean+dist_l1ndotn-hs300 on $\text{MCL}_{CE}^{nen-acc}$ → $\text{RSS+ruMCL}_{MSE+}^{dev1-sentSpear}$ (***M2***), Mean	0.780	0.773	0.786	0.780
LinReg on *M1* + *M2* + *M3*	0.780	0.756	0.772	**0.811**
p12, p13: mean+dist_l1 on $\text{MCL}_{CE}^{nen-acc}$ →$\text{RSS}_{MSE}^{dev2-sentSpear}$ (***M3***), Mean *p23*: first+concat on $\text{MCL}_{CE}^{nen-acc}$ →$\text{RSS+ruMCL}_{CE}^{dev2-sentSpear}$, Mean	0.779	0.749	0.801	0.788
p12, p13: *M2*; *p23*: max+concat on $\text{MCL}_{CE}^{nen-acc}$ →$\text{RSS}_{MSE}^{dev2-sentSpear}$, Mean	0.778	0.779	0.775	0.779
p12, p13: *M3*, LinReg* *p23*: mean+comb_dmn on MCL_{CE}^{en-acc} →$\text{RSS}_{MSE}^{dev2-sentSpear}$, LinReg	0.757	0.750	0.732	0.788
Our best models with ablation analysis (* models not from paper)				
*mean+dist_l1ndotn-hs0 on $\text{MCL}_{CE}^{nen-acc}$ →$\text{RSS}_{CE}^{dev2-sentSpear}$, Mean	**0.843**	**0.846**	**0.848**	**0.836**
mean+dist_l1ndotn-hs0 on $\text{MCL}_{CE}^{nen-acc}$ →$\text{RSS}_{MSE}^{dev2-sentSpear}$, Mean	0.823	0.825	0.821	0.823
*mean+dist_l1ndotn-hs0 on $\text{MCL}_{CE}^{nen-acc}$ →$\text{RSS+ruMCL}_{CE}^{dev2-sentSpear}$, Mean	0.822	0.809	0.833	0.825
mean+dist_l1ndotn-hs0 on $\text{MCL}_{CE}^{nen-acc}$ →$\text{RSS+ruMCL}_{MSE+}^{dev2-sentSpear}$, Mean	0.803	0.800	0.798	0.811
mean+dist_l1ndotn-hs0 on $\text{MCL}_{CE}^{nen-acc}$, Mean	0.776	0.777	0.778	0.772
mean+concat on $\text{MCL}_{CE}^{nen-acc}$ →$\text{RSS}_{MSE}^{dev2-sentSpear}$, Mean	0.768	0.760	0.759	0.784
mean+concat on $\text{MCL}_{CE}^{nen-acc}$ →$\text{RSS+ruMCL}_{MSE+}^{dev2-sentSpear}$, Mean	0.791	0.790	0.786	0.797

-6.7pt. if 2nd ft. step removed

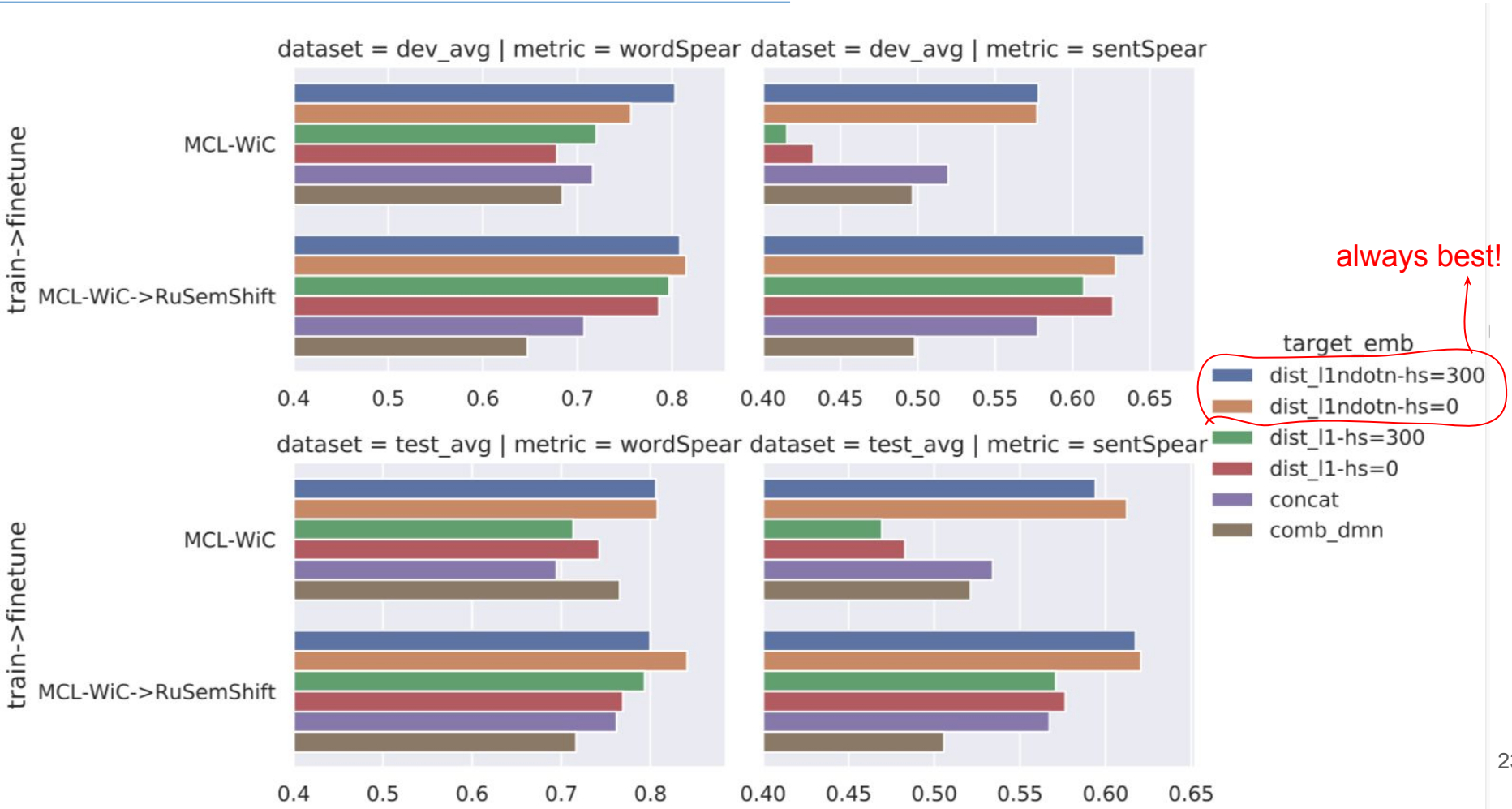
+2pt. from MSE->CE loss at the 2nd step

+5.5pt. from a change in WiC arch. (target aggregation)

Experiments: Target aggregation (gold sent. pairs)

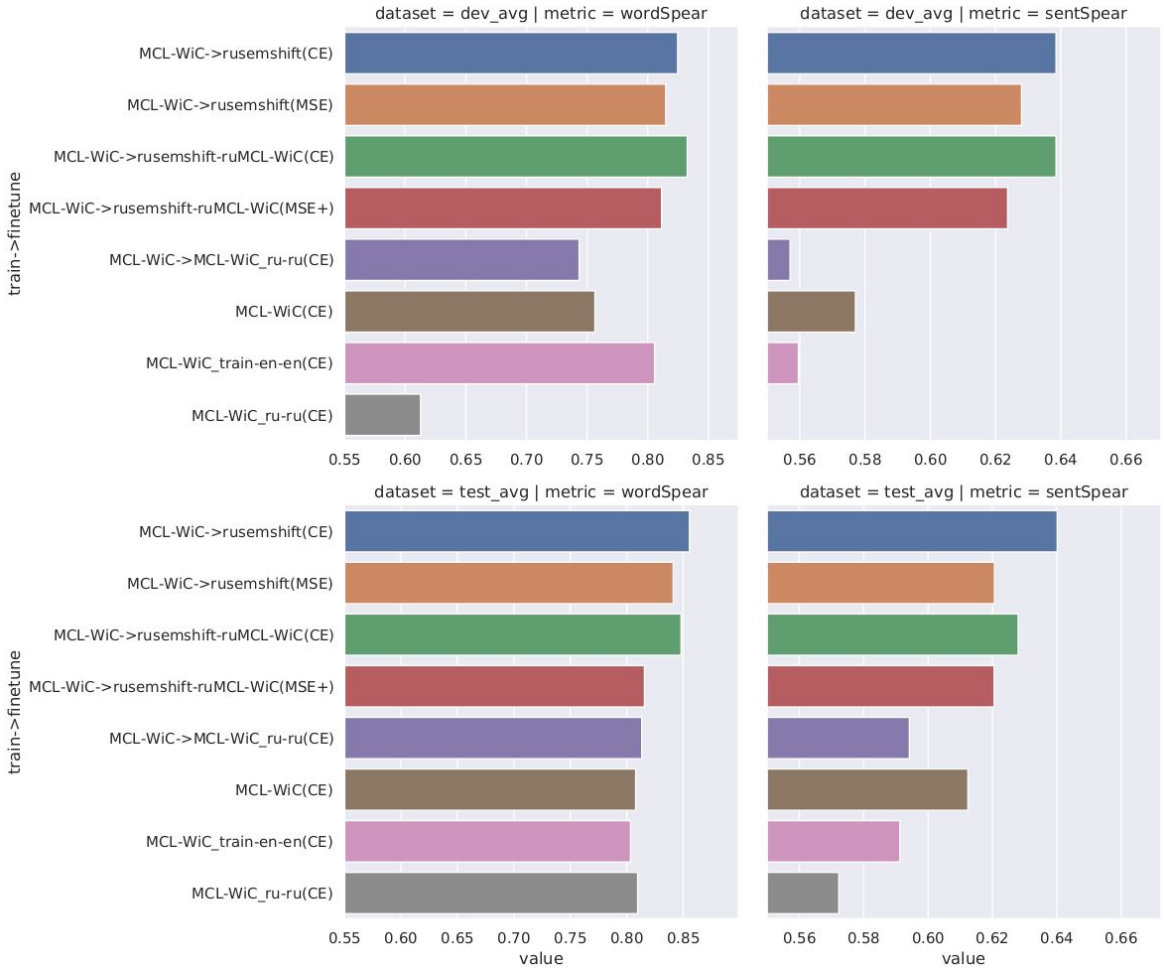


Experiments: Target aggregation (gold sent. pairs)



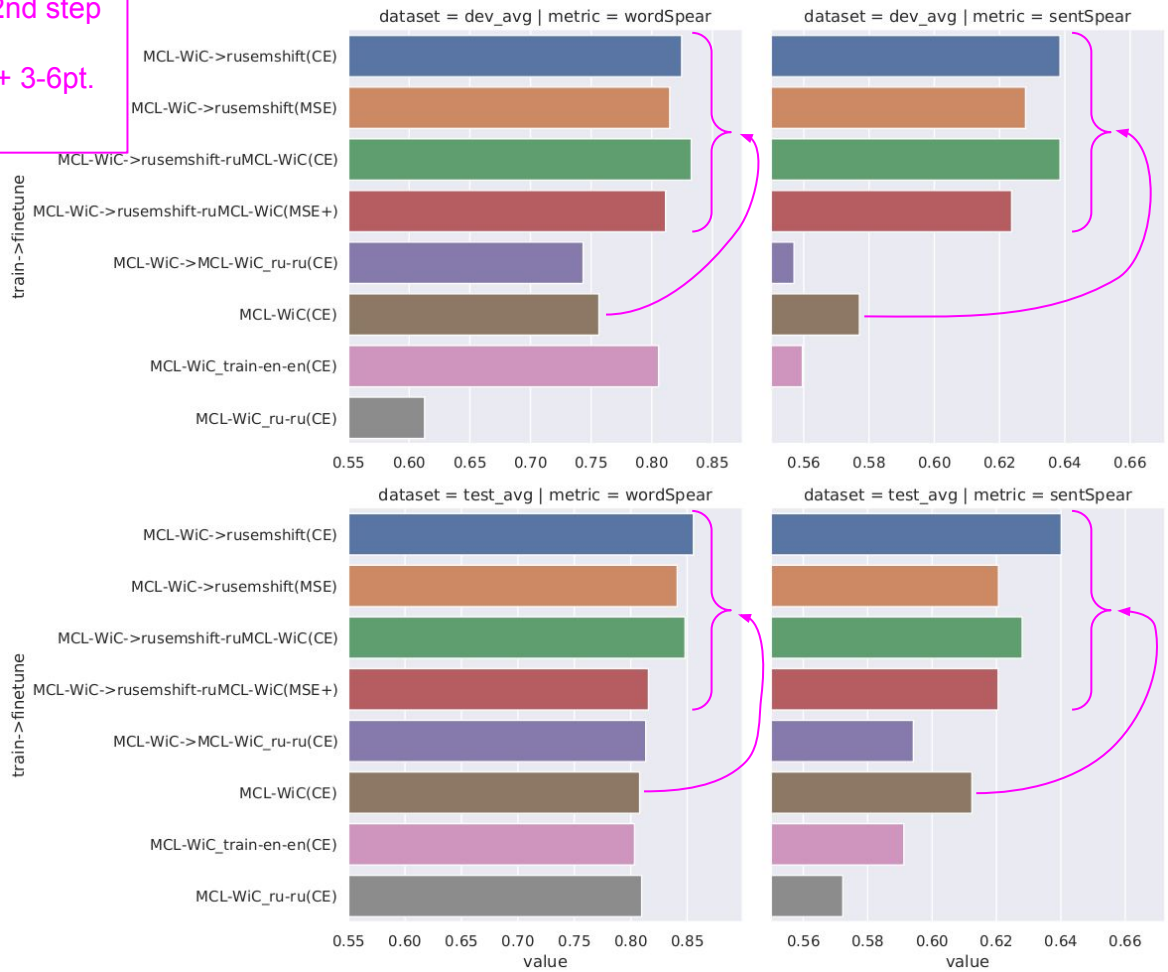
Experiments: Target aggregation (gold sent. pairs)

Experiments: Ft. schemes for *l1ndotn-hs=0* target aggregation (gold sent. pairs)

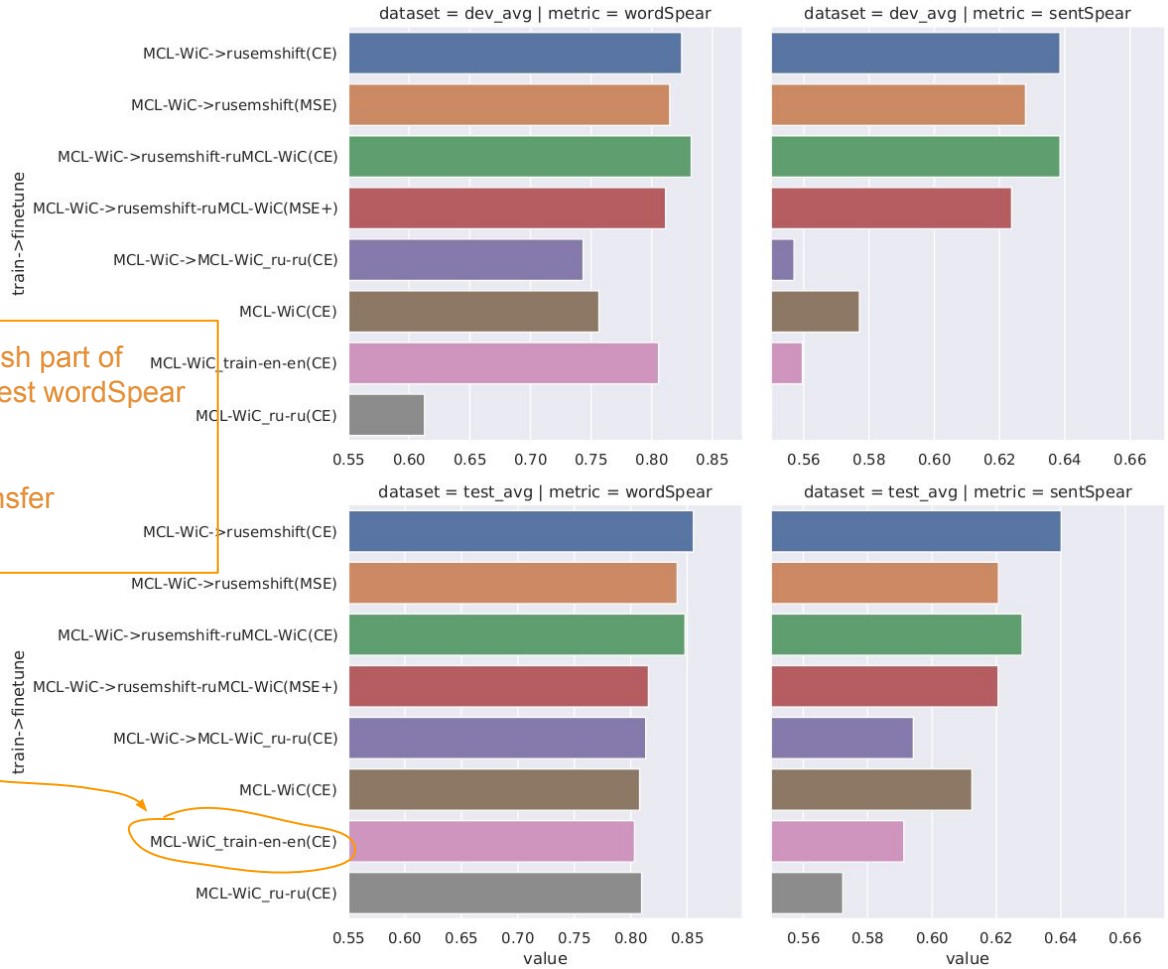


Experiments: Ft. schemes for *l1ndotn-hs=0* target aggregation (gold sent. pairs)

2-step ft. with RSS on the 2nd step consistently helps:
approx. +5pt. wordSpear / + 3-6pt. sentSepar



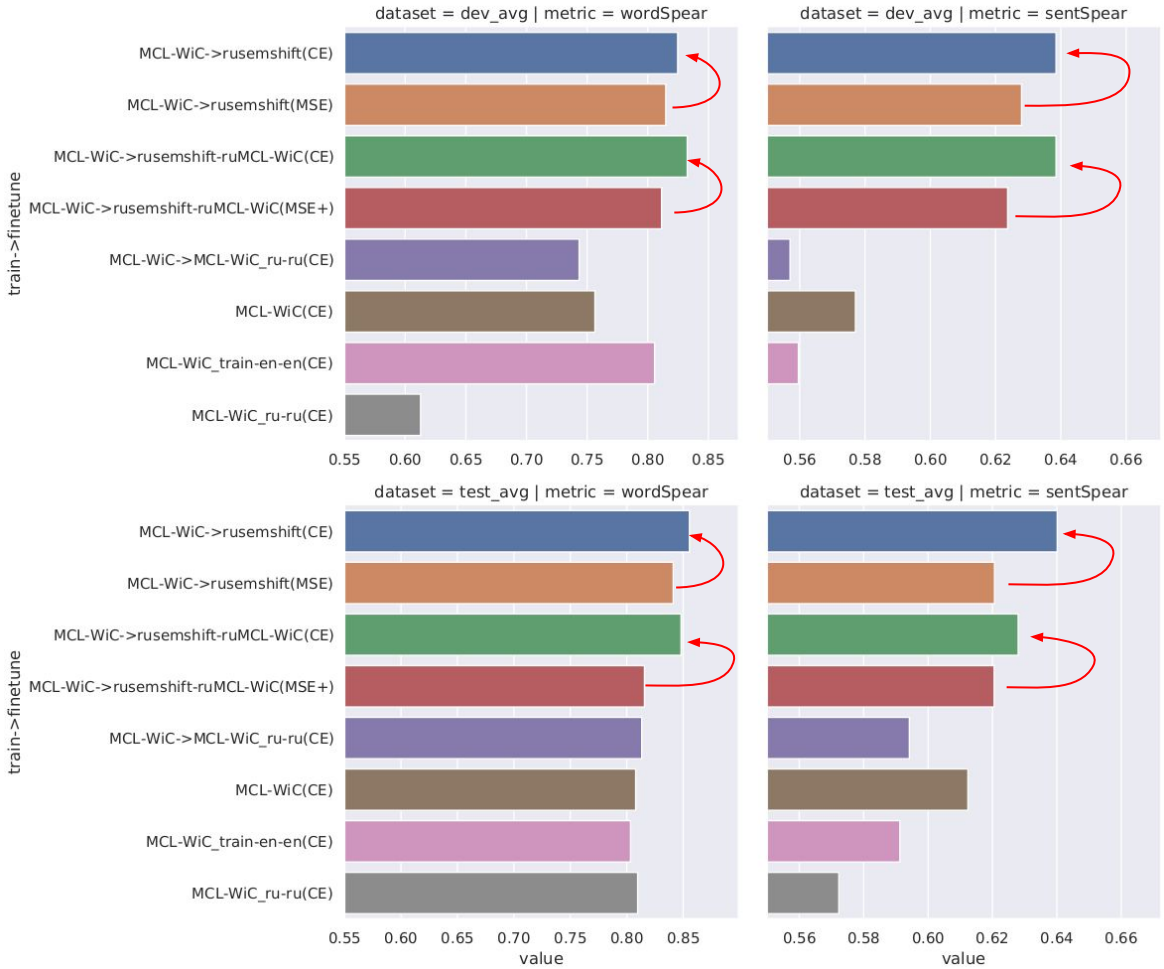
Experiments: Ft. schemes for *l1ndotn-hs=0* target aggregation (gold sent. pairs)



But even 1-step ft. on English part of MCL-WiC only gives ~0.8 test wordSpear (cmp. 0.802 of the winner)!

Zero-shot cross-lingual transfer works for LSCD!

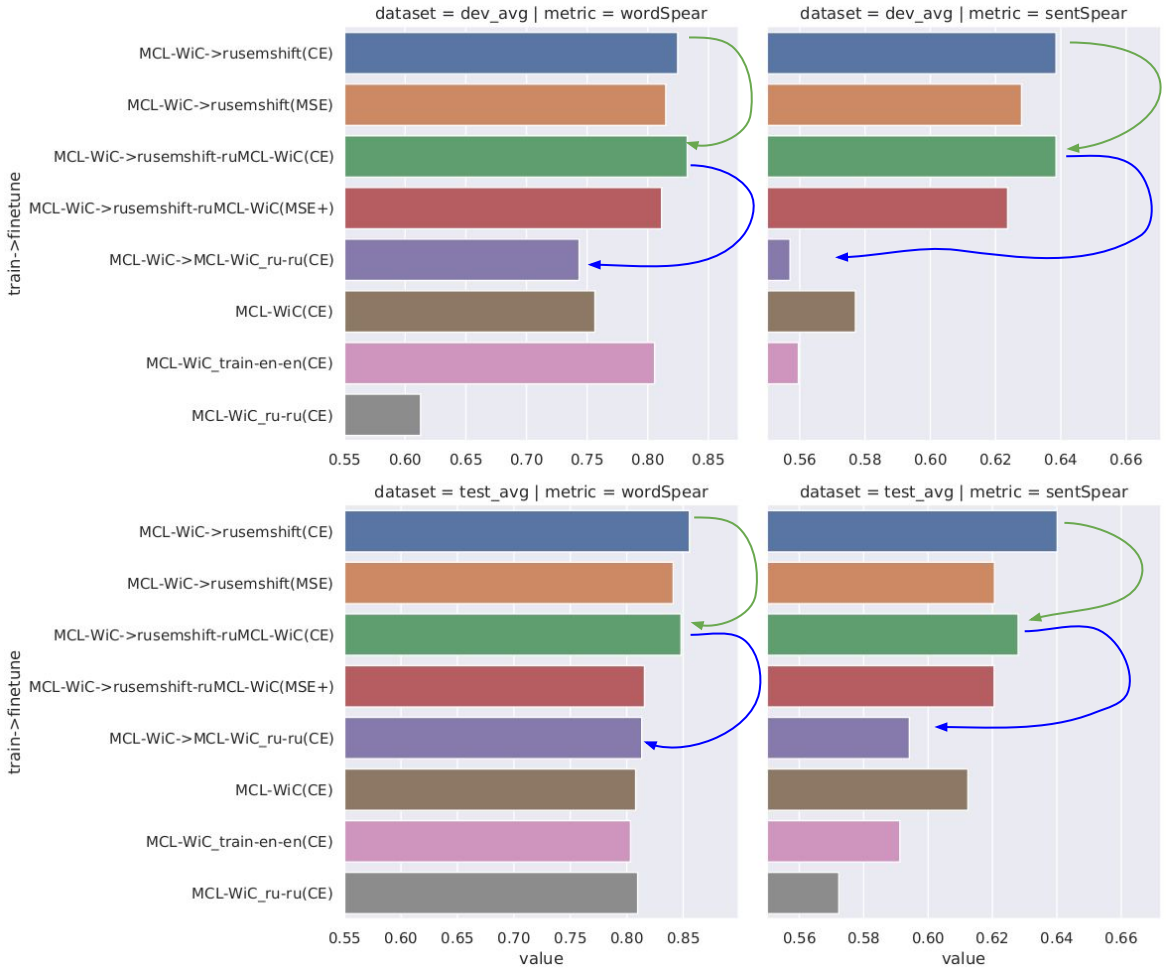
Experiments: Ft. schemes for *l1ndotn-hs=0* target aggregation (gold sent. pairs)



CE loss is the best loss on the 2nd step (despite binarization),

no need for MSE+

Experiments: Ft. schemes for *l1ndotn-hs=0* target aggregation (gold sent. pairs)



adding Russian part of MCL-WiC to RSS on the 2nd step is useless and replacing RSS with it hurts

Plan:

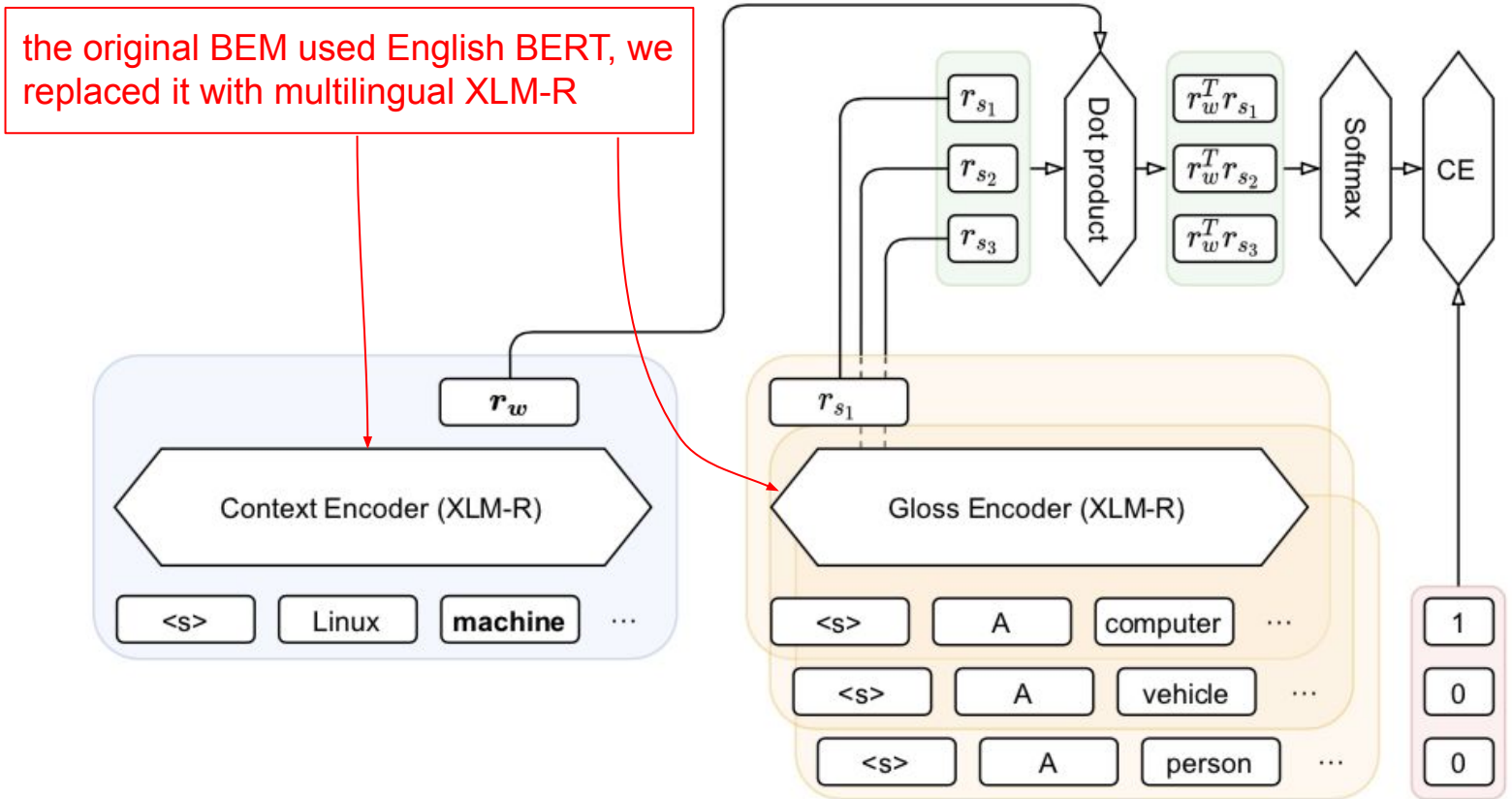
1. RuShiftEval-2021 shared task on Lexical Semantic Change Detection (LSCD) for the Russian language
2. DeepMistake solution ← 2nd best, outperformed the winner in the post-eval
   XLM-R fine-tuned for the WiC task
3. **GlossReader solution** ← the winning solution
   XLM-R fine-tuned for the gloss-based WSD task
4. (briefly) Results in LSCDiscovery-2022 shared task on Lexical Semantic Change Detection (LSCD) for the Spanish language

GlossReader: gloss-based WSD fine-tuning

BEM (bi-encoder model): a simple and SOTA (for 2020) gloss-based WSD system:

Blevins and Zettlemoyer. Moving Down the Long Tail of Word Sense Disambiguation with Gloss Informed Bi-encoders, 2020

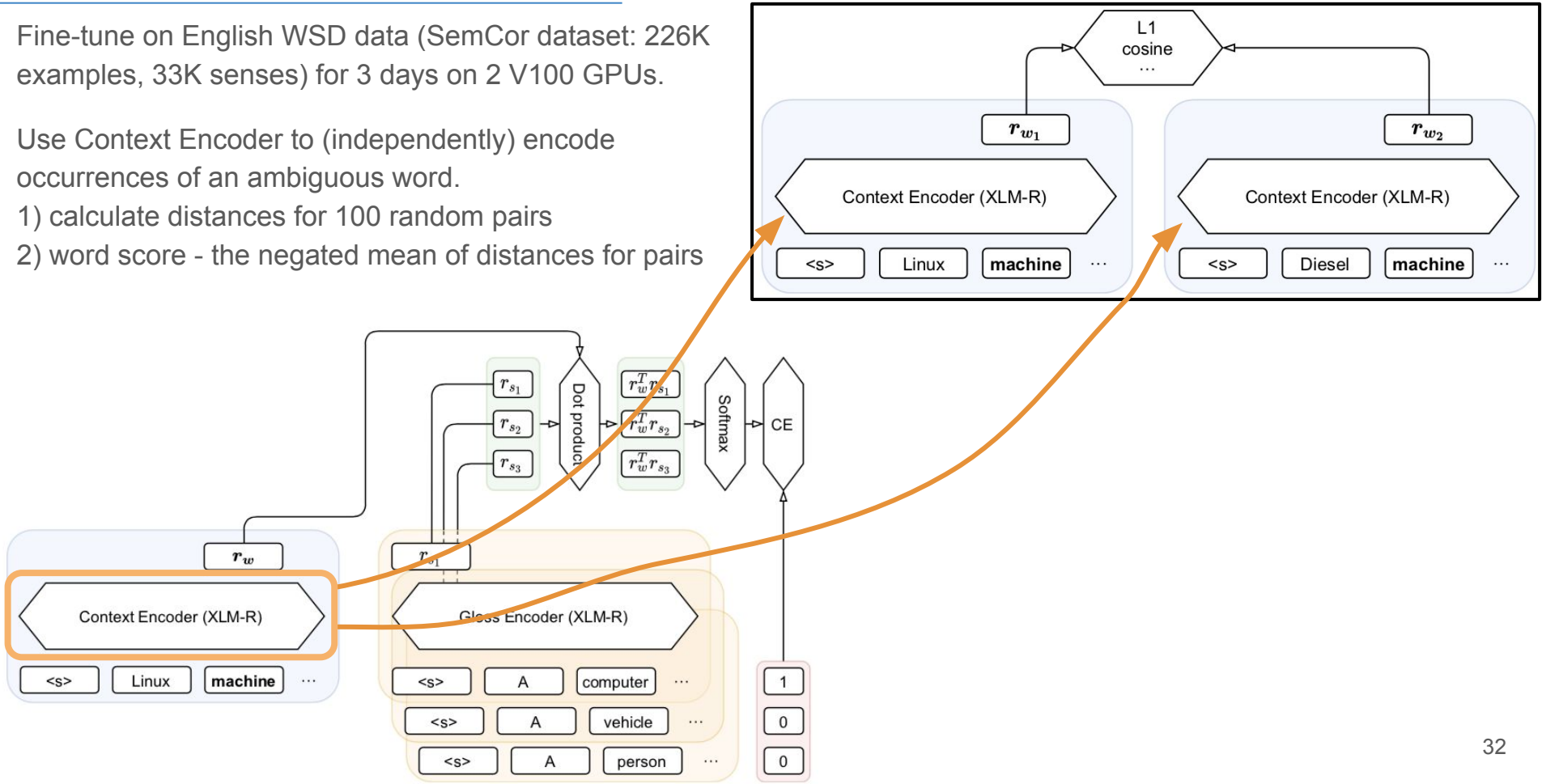
the original BEM used English BERT, we replaced it with multilingual XLM-R



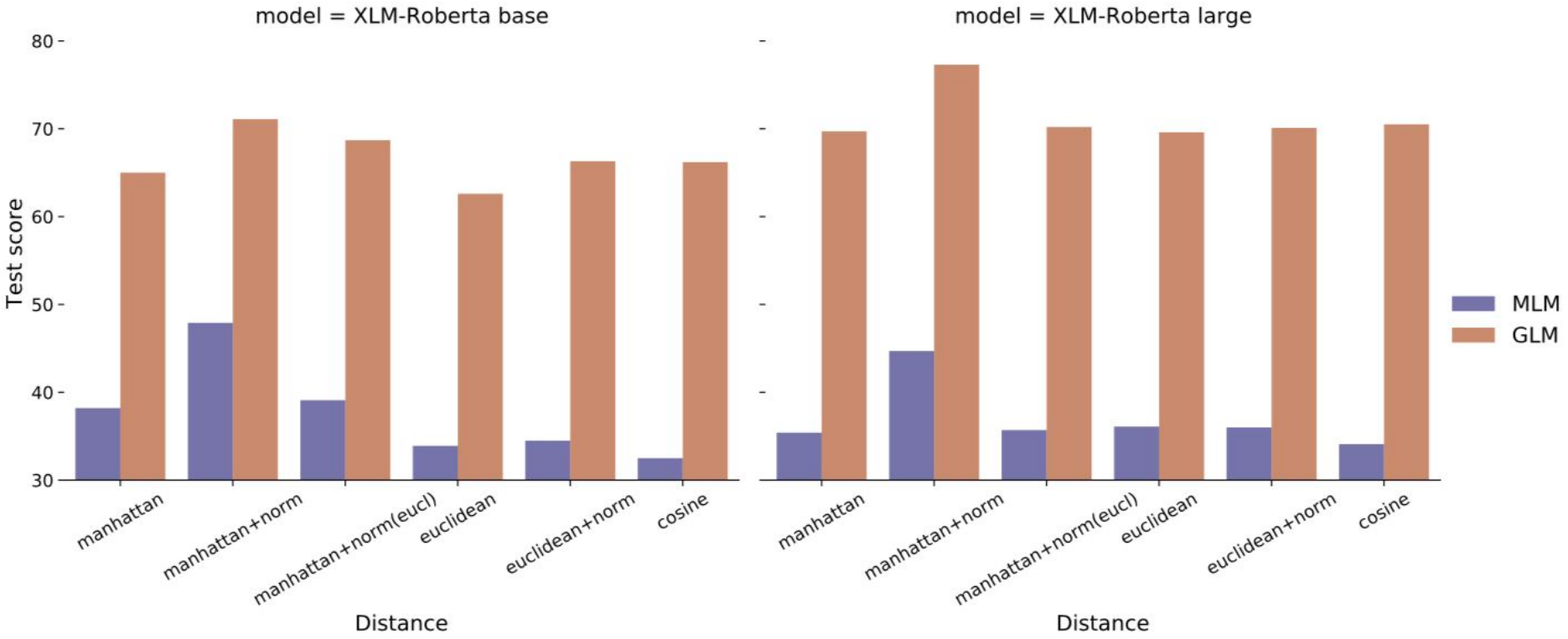
GlossReader: gloss-based WSD fine-tuning

Fine-tune on English WSD data (SemCor dataset: 226K examples, 33K senses) for 3 days on 2 V100 GPUs.

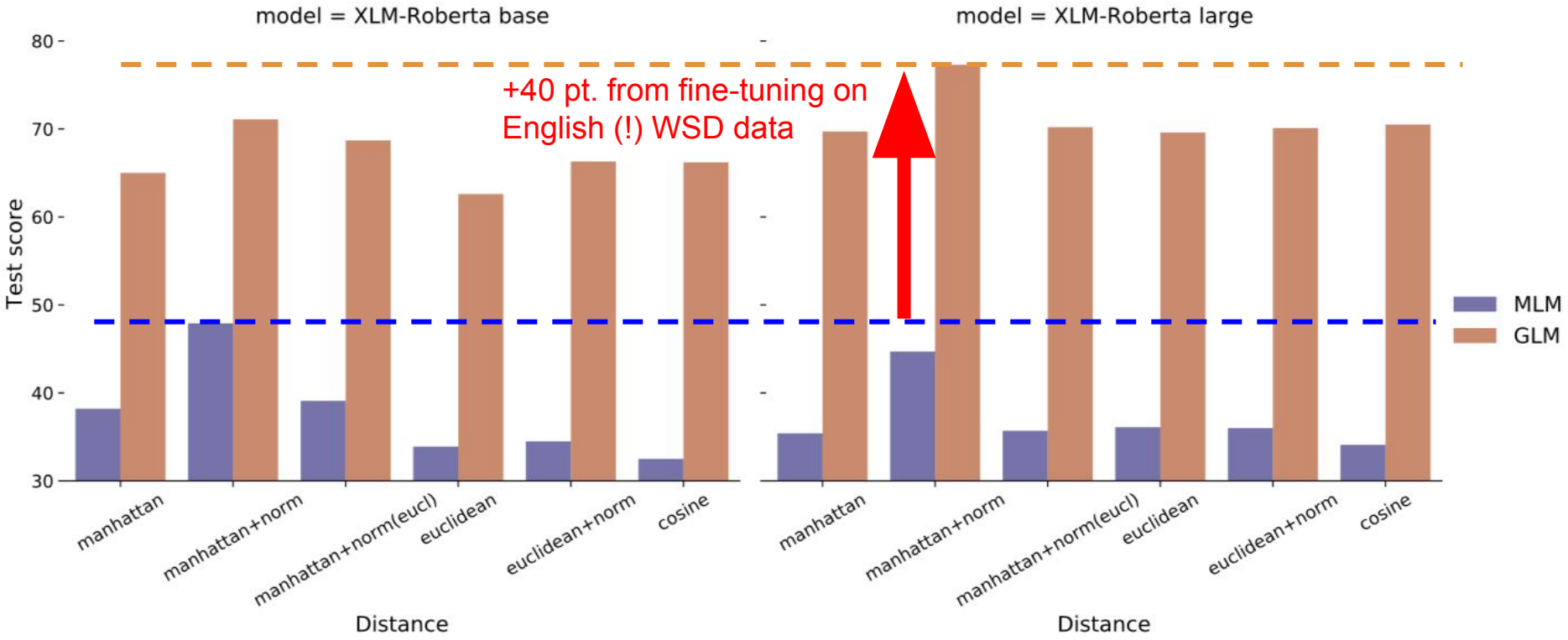
Use Context Encoder to (independently) encode occurrences of an ambiguous word.
1) calculate distances for 100 random pairs
2) word score - the negated mean of distances for pairs



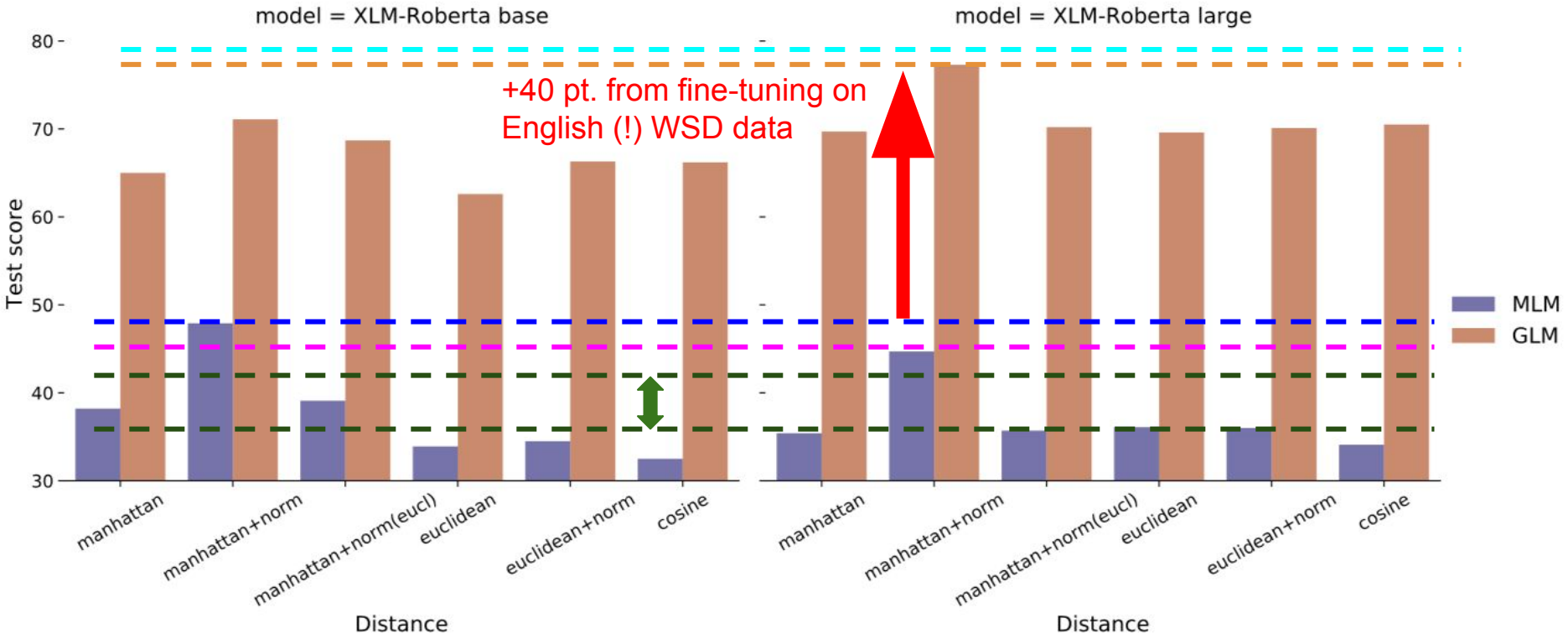
GlossReader: gloss-based WSD fine-tuning



GlossReader: gloss-based WSD fine-tuning



GlossReader: gloss-based WSD fine-tuning

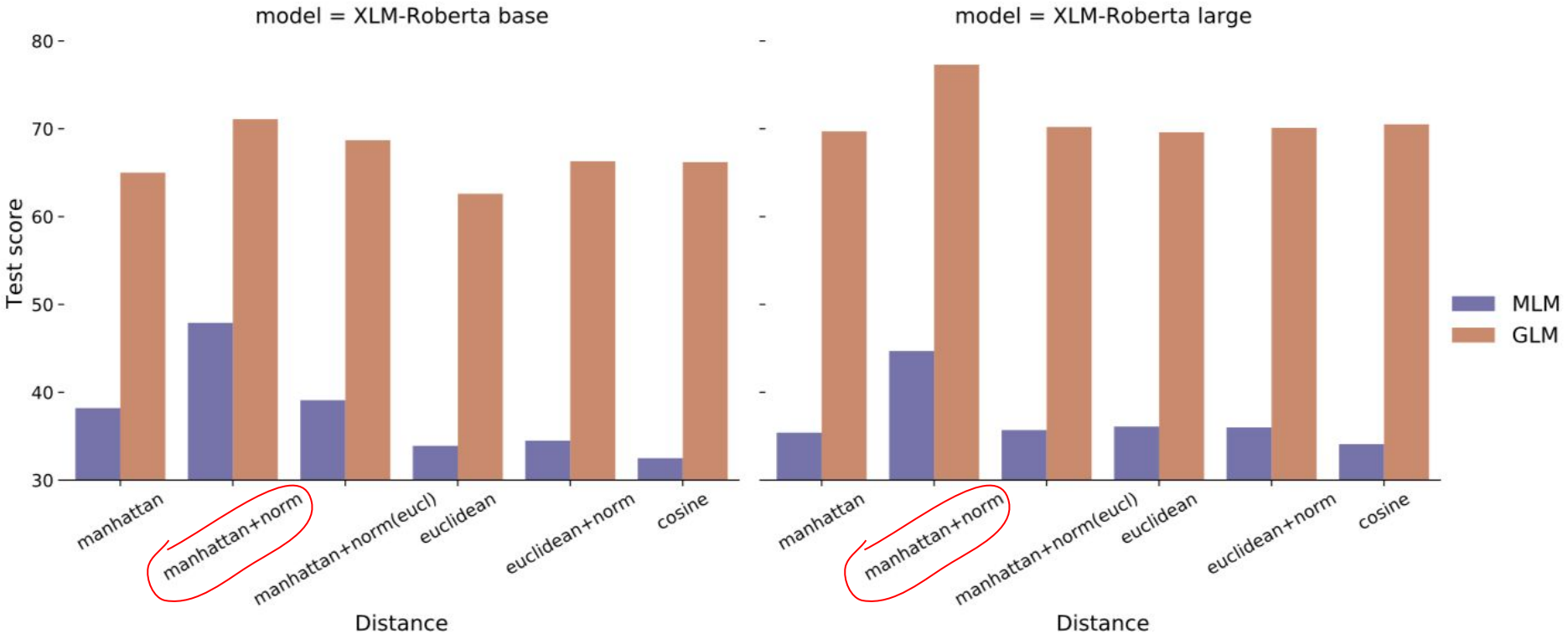


~0.45: non-finetuned ruBERT/ELMo (ranks 4-5)

0.37-0.42: type-based systems (ranks 6-9)

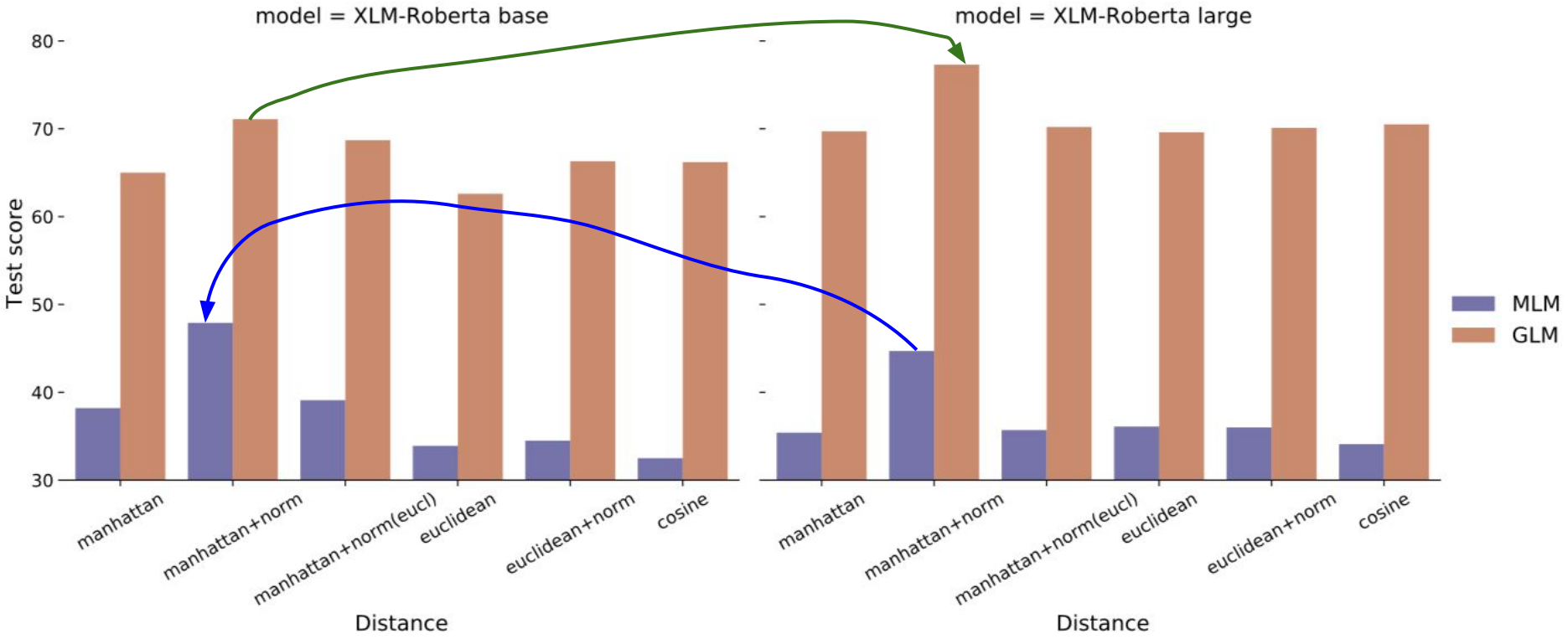
0.79: DeepMistake's best submission

GlossReader: gloss-based WSD fine-tuning



manhattan distance between L1-normalized embeddings consistently outperforms others (for base & large, fine-tuned and non-funetuned)

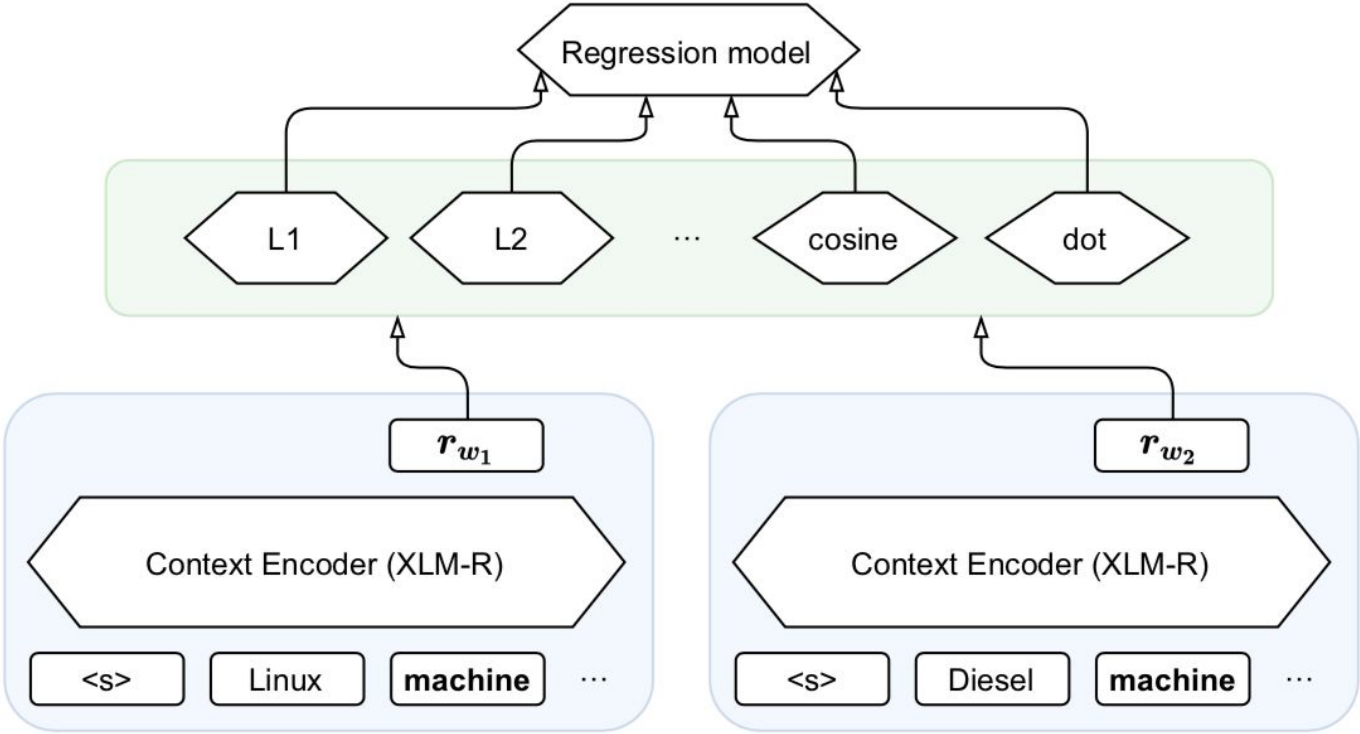
GlossReader: gloss-based WSD fine-tuning



when non-finetuned, XLM-R base outperforms large! (the larger model - the stronger grammatical bias?)
when fine-tuned, large outperforms base, as usual

GlossReader: regression

To benefit from RuSemShift training data: train a regression model on vectors of 14 distances: (L1, L1+norm, L2, L2+norm, Dot product, Dot product+L1-norm, Cosine) X (base, large)



GlossReader: regression

ID	Model	P1	P2	P3	Aver.
GLM, zero-shot cross-lingual transfer to Russian					
1	Manhattan+norm GLM xlmr.large	74.1	77.9	79.8	77.3
GLM + regression to human scores trained on RuSemShift					
2	Linear regression on GLM xlmr.large distances	77.0	80.1	81.8	79.6
3	Linear regression on GLM xlmr.large+base distances	78.1	**80.3**	**82.2**	**80.2**
4	Knn regression on GLM xlmr.large+base distances	71.8	76.2	80.9	76.3
5	Random.forest regr. (1K est.) on GLM xlmr.large+base dist.	75.2	78.7	81.6	78.5
6	Random.forest regr. (2K est.) on GLM xlmr.large+base dist.	75.0	78.7	81.7	78.5
7	Random.forest regr. (5K est.) on GLM xlmr.large+base dist.	75.8	78.4	81.3	78.5
The top3 best results of other teams					
-	DeepMistake	**79.8**	77.3	80.3	79.1
-	vanyatko	67.8	74.6	73.7	72.0
-	aryzhova	46.9	45.0	45.3	45.7

Linear regression outperforms other regressions and gives +3pt. (the winning solution)

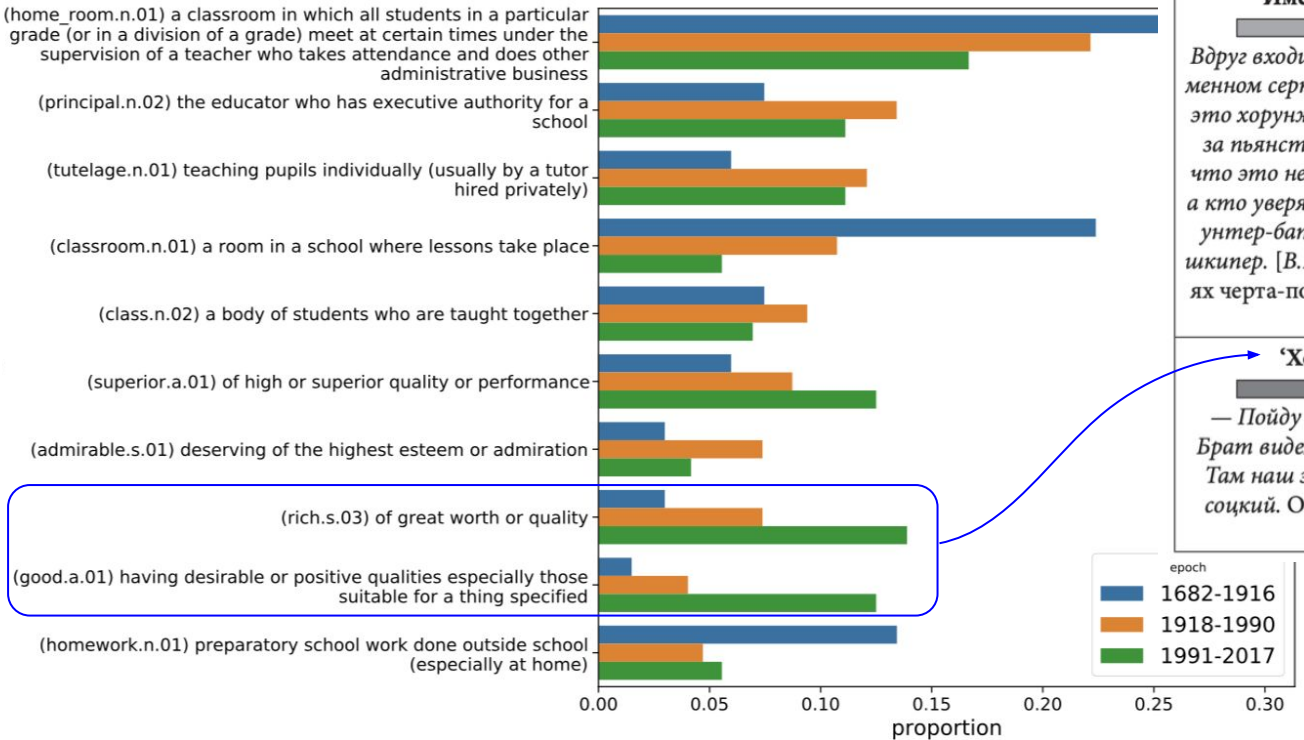
GlossReader: interpretation - preliminary experiment

Maybe use glosses for interpretation of LSCD decisions?

1) for each occurrence of the target word (in **Russian**), among **all ~120K glosses** from WordNet (in **English**) take 3 top ranking glosses ← *P(gloss | w, ctx)* from BEM w. XLM-R
2) for each gloss, plot the proportion of examples obtaining that gloss from each time period

GlossReader: glosses for the word ‘классный’

The chronology of senses for the word ‘классный’ from Daniehl’ M. A., Dobrushina N. R. Dva veka v dvadtsati slovakh [Two centuries in twenty words] (the translation of glosses is ours)



Хронология значений слова *классный*

Значение и пример (*Meaning and example*)	Период использования (*Time period of use*)
‘Имеющий отношение к школьному обучению’ (*Related to school education*) *Алеша возвратился в дом и весь вечер просидел один в* ***классных*** *комнатах...* [*Антоний Погорельский.* Черная курица (1829)]	1820-е гг.
‘Имеющий класс (разряд)’ (*Having a rank*) *Вдруг входит человек в изодранном форменном сертучишке, — кто говорил, что это хорунжий, отставленный три раза за пьянство и буянство; кто говорил, что это небольшой* ***классный*** *чиновник, а кто уверял, что это отставной клерк, унтер-баталер, а может быть, и подшкипер.* [*В.И. Даль.* Сказка о похождениях черта-послушника, Сидора Поликарповича (1832)]	1830-е гг.
‘Хороший, отличный’ (*Good, Excellent*) *— Пойду смотреть «Дело пестрых». Брат видел — говорит,* **классное** *кино. Там наш этому ка-а-ак дал!* [*В.С. Высоцкий.* О любителях «приключений» (1955–1960)]	1950-е гг.

Plan:

1. RuShiftEval-2021 shared task on Lexical Semantic Change Detection (LSCD) for the Russian language
2. DeepMistake solution ← 2nd best, outperformed the winner in the post-eval
   XLM-R fine-tuned for the WiC task
3. GlossReader solution ← the winning solution
   XLM-R fine-tuned for the gloss-based WSD task

4. **(briefly) Results in LSCDiscovery-2022 shared task on Lexical Semantic Change Detection (LSCD) for the Spanish language**

LSCDiscovery-2022: leaderboard for graded subtasks

GlossReader: no adaptation to Spanish, no linear regression, just average distance between the contextualized embeddings after WSD fine-tuning on English WSD dataset!

DeepMistake: 2nd step of fine-tuning on Spanish WiC data from the task

Task		Change graded	COMPARE
#	Team name	SPR	SPR
1	GlossReader	**0.735 (1)**	0.842 (1)
2	DeepMistake	**0.702 (2)**	0.829 (2)
3	HSE	**0.553 (3)**	0.558 (4)
4	baseline1	0.543 (4)	0.561 (3)
5	baseline3	0.508 (5)	0.459 (5)
6	Rombek	0.497 (6)	0.456 (6)
7	CoToHiLi	0.282 (7)	–
8	baseline2	0.092 (8)	0.088 (7)
9	baseline5	0.064 (9)	-0.072 (8)
10	BOS	-0.125 (10)	-0.129 (9)

Tables from **Frank D. Zamora-Reina, Felipe Bravo-Marquez, and Dominik Schlechtweg. 2022. LSCDiscovery: A shared task on semantic change discovery and detection in Spanish**

LSCDiscovery-2022: leaderboard for binary subtasks

Both models: thresholding of graded predictions

Task		Change binary		
#	Team name	F1	P	R
1	GlossReader	**0.716 (1)**	0.615 (3)	0.857 (3)
2	UAlberta	**0.709 (2)**	0.549 (7)	1.000 (1)
3	Rombek	**0.687 (3)**	0.590 (4)	0.821 (4)
4	BOS	0.658 (4)	0.510 (8)	0.929 (2)
5	DeepMistake	0.655 (5)	0.633 (2)	0.679 (6)
6	CoToHiLi	0.636 (6)	0.553 (6)	0.750 (5)
7	baseline4	0.636 (6)	0.467 (11)	1.0 (1)
8	HSE	0.586 (7)	0.567 (5)	0.607 (7)
9	baseline3	0.548 (8)	0.500 (9)	0.607 (7)
10	baseline1	0.537 (9)	0.846 (1)	0.393 (9)
11	baseline5	0.508 (10)	0.484 (10)	0.536 (8)
12	baseline2	0.222 (11)	0.500 (9)	0.143 (10)

Task		Sense gain			Sense loss		
#	Team name	F1	P	R	F1	P	R
1	GlossReader	**0.511 (3)**	0.333 (5)	0.929 (2)	**0.688 (1)**	0.564 (2)	0.880 (2)
2	DeepMistake	**0.591 (1)**	0.433 (1)	0.929 (2)	0.582 (5)	0.533 (3)	0.640 (4)
3	HSE	0.250 (8)	0.192 (9)	0.357 (5)	0.364 (7)	0.421 (5)	0.320 (5)
4	baseline1	–	–	–	–	–	–
5	Rombek	0.50 (4)	0.409 (2)	0.643 (4)	**0.681 (2)**	0.727 (1)	0.640 (4)
6	baseline3	–	–	–	–	–	–
7	BOS	**0.520 (2)**	0.361 (4)	0.929 (2)	**0.610 (3)**	0.529 (4)	0.720 (3)
8	baseline2	0.211 (9)	0.400 (3)	0.143 (6)	0 (8)	0 (8)	0 (7)
9	UAlberta	0 (10)	0 (10)	0 (7)	0 (8)	0 (8)	0 (7)
10	CoToHiLi	0.462 (5)	0.316 (6)	0.857 (3)	0 (8)	0 (8)	0 (7)
11	baseline4	0.378 (6)	0.23 (8)	1.0 (1)	0.588 (4)	0.416 (6)	1.0 (1)
12	baseline5	0.333 (7)	0.313 (7)	0.357 (5)	0.367 (6)	0.375 (7)	0.36 (6)

Tables from **Frank D. Zamora-Reina, Felipe Bravo-Marquez, and Dominik Schlechtweg. 2022. LSCDiscovery: A shared task on semantic change discovery and detection in Spanish**

Conclusions / take-aways

1) XLM-R (and likely other LMs/MLMs also) doesn't give good representations of word meaning in context out-of-the-box (high orthographic / grammatical bias).

   We previously observed this for WSI, WiC, and now for LSCD also.

2) Fine-tuning for the WiC or gloss-based WSD tasks gives a huge boost in performance for the LSCD task.

   Even if the labeled data is in different language!

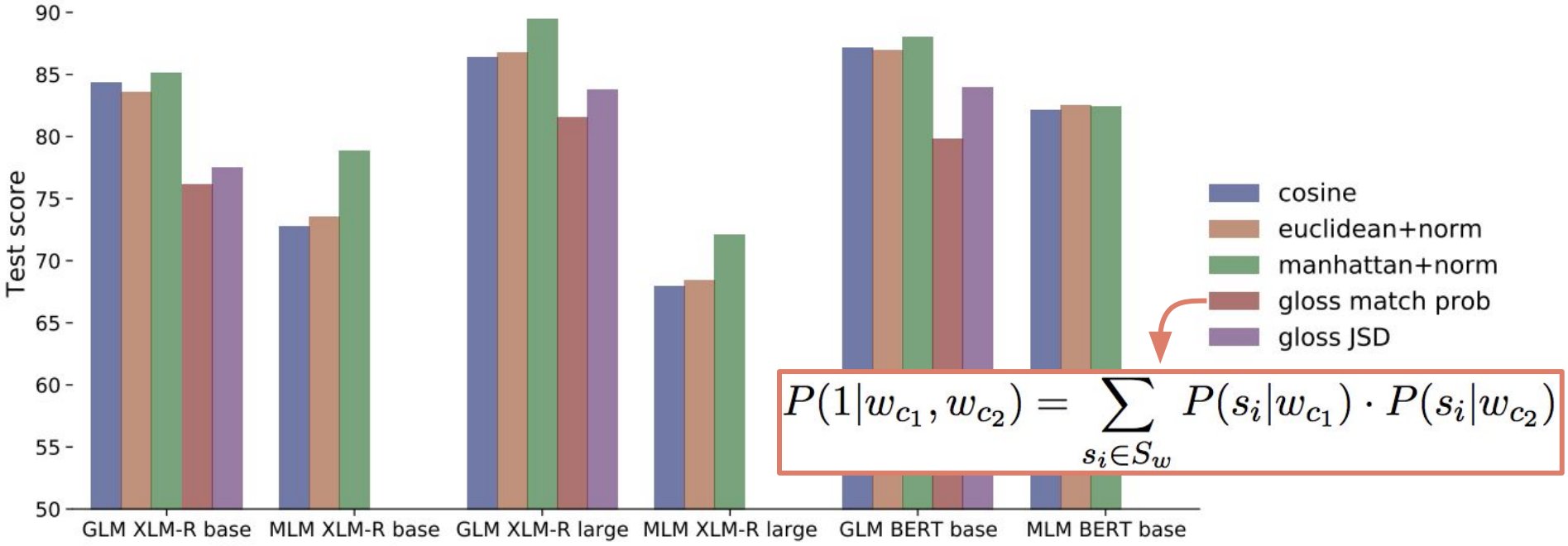
   a) Maybe fine-tuning for some other sense-related tasks is even better?
   b) Can we think about some self-supervision similar to MLM, but resulting in good representations of word meaning in context out-of-the-box?

3) The deepest mistake of the DeepMistake model - it confuses senses that we can clearly distinguish. The most noticeable case are the senses under the 'human'/'person' hypernym.
   a) Maybe we achieved the limits of the distributional hypothesis?
   b) Can we observe similar trends for other languages?

Thank you for your attention!

GlossReader: glosses for WiC

Gloss encoder was jointly trained with Context encoder, but not used then.
For English part of MCL-WiC: run full WSD and compare distributions over senses/glosses of the target word in each context (JSD or gloss matching probability).



accuracy on MCL-WiC, en-en test set

Error analysis: selecting words

- Test set of 99 unique target words with annotated sentence pairs
- Results by the model $\mathrm{MCL}_{CE}^{nen-acc} \rightarrow \mathrm{RSS+ruMCL}_{CE}^{dev1-sentSpear}$
- ΔRank - the difference between the predicted and the gold word ranks

We consider only words with **|ΔRank| ≥ 25** at least for one pair of periods.

1. p12 - 24 words
2. p23 - 18 words
3. p13 - 13 words

=> 27 unique words for analysis

Error analysis: selecting sentence pairs

For these 27 words, we annotated sentence pairs with high disagreement:

$$|\Delta Score| = |S_{model} - S_{annotator}| \geq 1.5$$

- Overall 171 pairs of sentence

Error analysis: types of errors

- **Model can not find the difference. 39%**

The model incorrectly classifies two word occurrences as having the same meaning

Gold scores	Model score	Sentence pairs	Meanings
2 1 2	4	Скоро прибыли к нему **братья** его, Андрей и Борис, с их многочисленною дружиною: не было ни упреков, ни извинений, ни условий; динокровные обнялися с видом искренней любви, чтобы вместе служить отечеству и христианству. Он говорил ей: ”Ты, **брат**”.	***brothers - relatives*** ***brothers - form of address***
1 1 1	4	Злополучный иеромонах был вытащен из **огня** со слабыми признаками жизни. Ночью немцы обрушили на наше расположение массированный артиллерийский **огонь**.	***flame, fire*** ***firing, shooting***
1 4 3	4	Команда, за исключением вахтенных, ушла в **увольнение**, в город. Первым плохим признаком стал запущенный в прессу слух, что сразу же после **увольнения** Примакова Рапота сам подал в отставку, а на его место уже подбирается новая кандидатура.	***vacation*** ***dismissal***

Error analysis: types of errors

- **Model sees wrong difference. 22.8%**

The model incorrectly classifies two word occurrences as having different meanings

Gold scores	Model score	Sentence pairs	Meanings
4 4 4	1	Тут в **кармане** тысяча рублев положена. Вместо птиц он приносил домой целые **карманы** камней и сваливал их под навесом в ящик.	***pocket (clothing)***
4 4 4	1	Когда Их Величества повернули к той стороне **стены**, которая ведет к Спасским воротам, на площади уже стояла тысячная толпа, приветствовавшая Царя и Царицу восторженными кликами ”ура” и бросаньем в воздух шапок. В конце этого двора у **стены** поставлены бочки, на которые наложены доски.	***wall (part of the building)***
4 3 4	1	На руках она с усилием тащила Федьку, прижав его поперек **живота**, чем он нисколько не смущался. Боли в **животе** не то стали слабее, не то он к ним привык	***belly (body part)***

Error analysis: types of errors

- **Model seems to be right. 23.2%**

Pairs of sentences, that in our opinion were correctly classified by the model, but incorrectly annotated by one or more annotators

Gold scores	Model score	Sentence pairs	Meanings
4 4 3	1	… он сильно сочувствует вопросам своего времени, страдает всеми недугами **века**, болезненно мучится несовершенствами общества. Во второй половине прошлого **века** рытье колодцев было заменено бурением скважин.	***epoch*** ***century***
3 3 1	4	… если человек, как бесноватый, одержим одною дикою страстью и ко всему другому восприимчивость в нем потухла, -- он или **маньяк**, которого надо лечить, или воплощенный дьявол, для которого все человеческое кончено. Любое слово, любой жест истолковывается с железной логикой **маньяка** как направленный к моему уничтожению.	***dangerous, ill person***

Error analysis: types of errors

- **Ambiguity. 15%**

From the context we could not understand whether two word occurrences have the same meaning

Gold scores	Model score	Sentence pairs	Meanings
4 1 4	4	Ребята, с **тачками** туда!.. Скотник, насквозь напитанный запахом навозной жижи, тот самый, что не довез своей **тачки**, а пошел смеяться с мужиками, когда мужики еще добродушно покуривали...	***cart/car*** ***cart***
4 2 4	1	Возрас Александр, и по многих победах Аравию завоевав, прислал **дядьке** своему несколько сот возов ладону, написав ему: не будь скуп, где дело идет до Богопочтения.", Это был **дядька** средних лет в пятнистой робе.	***uncle (relative)/teacher*** ***stranger***

Error analysis: results

1. Model can not find the difference. **39%**
2. Model sees wrong difference. **22.8%**
3. Model seems to be right. **23.2%**
4. Ambiguity. **15%**


амбиция
брат
век
дядька
дядя
живот
карман
крыша
лечение
p12
p23
p13
ΔRank
brother: relative / form of address
uncle / teacher / stranger
uncle / “sir” / stranger
линейка
маньяк
монстр
наволочка
наложение
огонь
риторика
совершенство
советчик
maniac: direct / indirect
стена
тачка
увольнение
углеводород
уклад
фаворит
хозяйка
цензура
цифра
cart / car
female owner / housekeeper / form of address

Error analysis

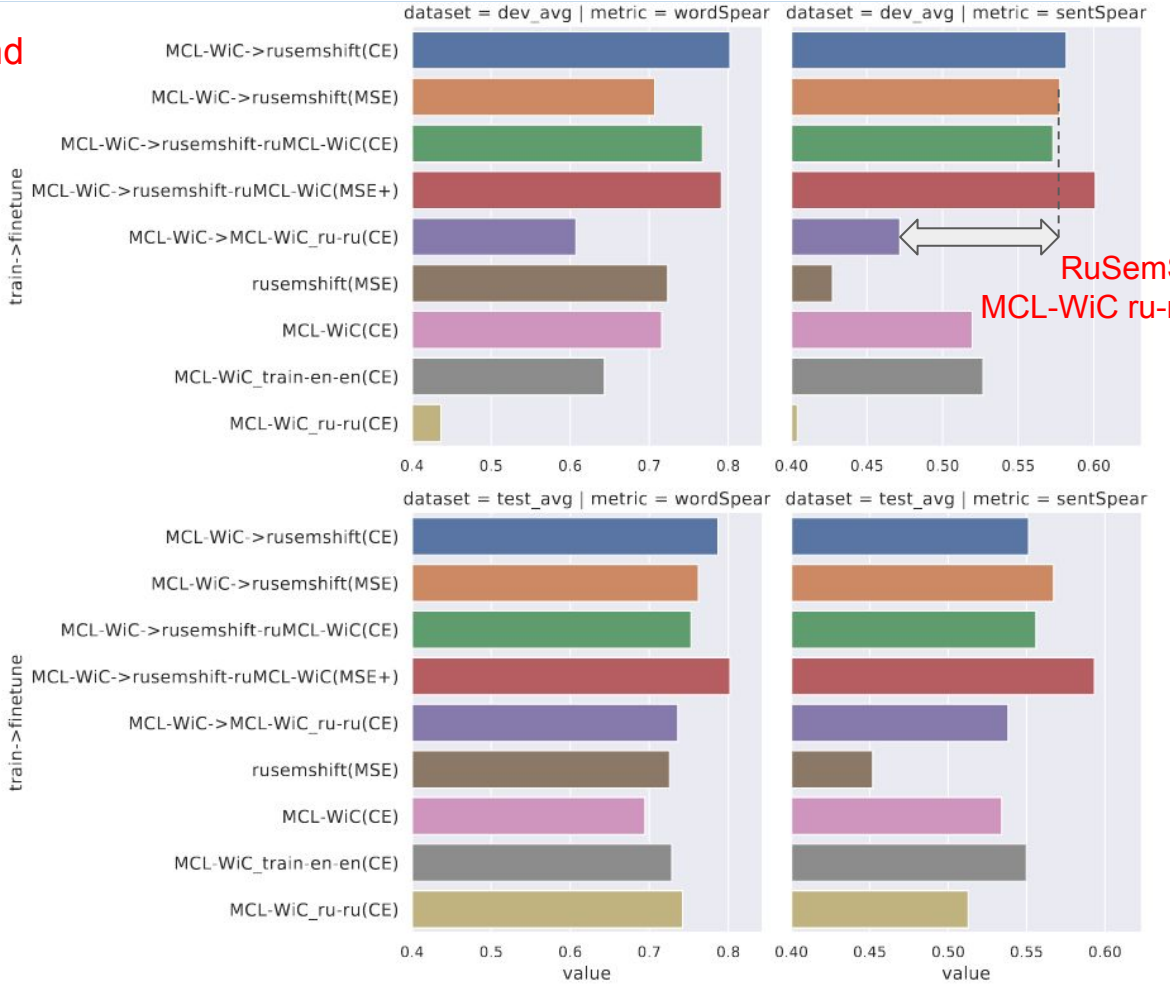
Quantity of human related words:

- Among all: 8/99 ~ 0.08
- Among errors: 6/27 ~ 0.22

Sentence pair	Error type
1) "Я, разумеется, не про ремесла говорю, не про технику, не про математические знания, -- этому и немцы заезжие по найму научат, если мы не научим, нет, а мы-то чему Мы ведь русские, **братья** этому народу, а стало быть, обязаны просветить его.", 2) Их четыре **брата**.	*Ambiguity*
1) "Взошел, шаркая костылем, дряхлый **дядька** Василия Иваныча, отставной камердинер Тихон." 2) Давай тогда искать твоего **дядьку**.	*Model can't find the difference*
1) Мы поднялись с **дядей** Мишей поздно. 2) Сколько их было -Две тети и один **дядя**.	*Model can't find the difference*
1) "Сухонькая голубоглазая **хозяйка** в темном повойнике, с кротким выражением на безответном лице, хлопочет у печки: ворочает уголья, гремит хватом, двигает чугуники." 2) "Иногда мне казалось, что эти заколдованные предметы несли в себе не только характер своей **хозяйки**, но и становились ее глазами и ушами."	*Model can't find the difference*
1) "Унять было невозможно, по крайней мере в ту минуту, и -- вдруг окончательная катастрофа как бомба разразилась над собранием и треснула среди его: третий чтец, тот **маньяк**, который все махал кулаком за кулисами, вдруг выбежал на сцену." 2) "Не надо было быть балетным **маньяком**, чтобы понять, что балериной эта особа никогда не будет."	*Model seems to be right*

Experiments: Ft. schemes for *concat* target aggregation (gold sent. pairs)

2-step ft. with MCL-WiC and RuSemShift is better than 1-step training



RuShiftEval-2021: Examples of annotated sentences

sent 1	sent 2	a1	a2	a3	Mean
Но зато он и выпускает это мясо на шестнадцатичасовой работе с **тачкой** в тридцать пудов. (But on the other hand, he produces this meat at sixteen-hour work with a **cart** of thirty poods.)	В Думе теперь шерстят за лишние **тачки**. (The members of the Lower House are now carefully inspected for too many cars.)	1	1	1	1
В Воеводской тюрьме содержатся прикованные к **тачкам**. (In the Voivod Prison, people are kept chained to **carts**.)	Это те, кто мотается на своих разнокалиберных **тачках** по городу в поисках приключений, а когда оно им подворачивается, то с удовольствием в него бросаются. (These are those who dangle in their different-sized **cars** around the city in search of adventure, and when it turns up for them, they rush into it with pleasure.)	2	1	1	1.33
Перевернутый стул вполне заменил **тачку**. (The overturned chair has completely replaced the **cart**.)	Моей бригаде поручено было следить за тем, чтобы паводковые и сточные воды, стекавшие в забой, не мешали работе забойщиков, особенно тех, кто гонял **тачки** на транспортер. (My team was instructed to ensure that the flood and waste water flowing into the face did not interfere with the work of the miners, especially those who drove the **carts** to the conveyor.)	3	1	4	2.66

The mean of all scores for sentence pairs for the word *тачка* in Pre-Soviet — Post-Soviet is 1.899