A WordNet View on Crosslingual Contextualized Language Models

Wondimagegnhue Tufa
CLTL Lab, VU Amsterdam
w.t.tufa@vu.nl

Lisa Beinborn
CLTL Lab, VU Amsterdam
l.beinborn@vu.nl

Piek Vossen
CLTL Lab, VU Amsterdam
p.t.j.m.vossen@vu.nl

Agenda

- Word Embeddings as WordNet
- Closer to Wordnet via Context
- The Problem with Crosslingual Models
- Crosslingual models approximate meaning
- Language Interaction in Crosslingual Models
- Conclusion

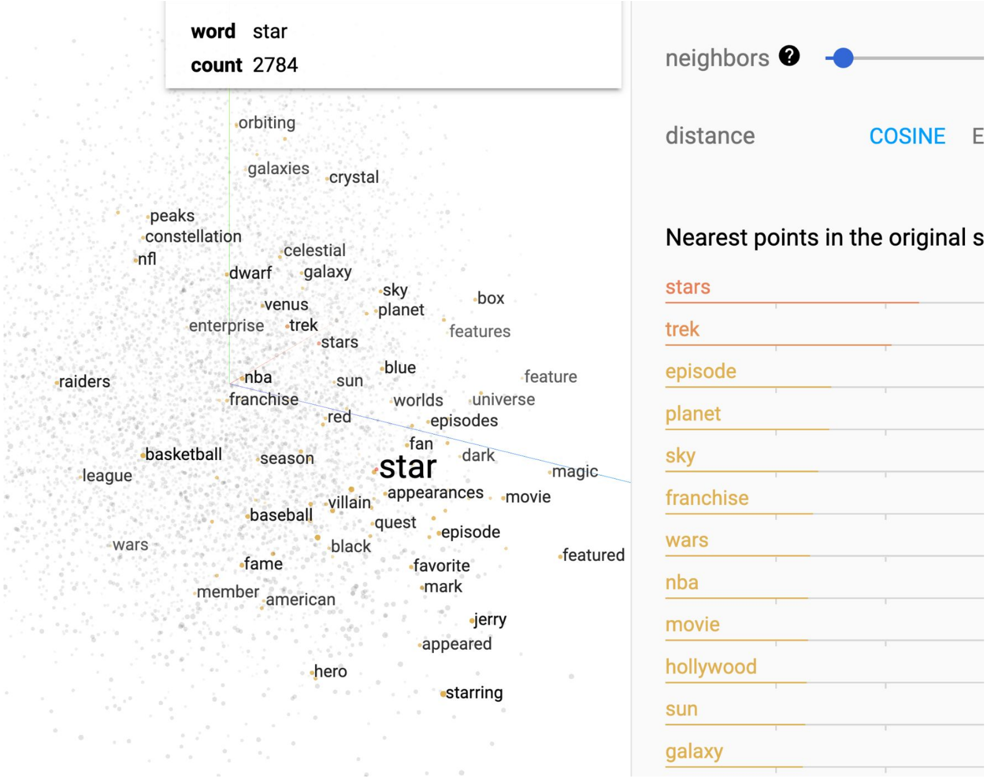
WordNet vs Distributional Representations

Think Word Embeddings as loose form of Wordnet

Star[1], Asteroids, Planets, Satelites

Star[2], Superstar, Whiz, Wizard

Star[3], Principal, Lead, Chief

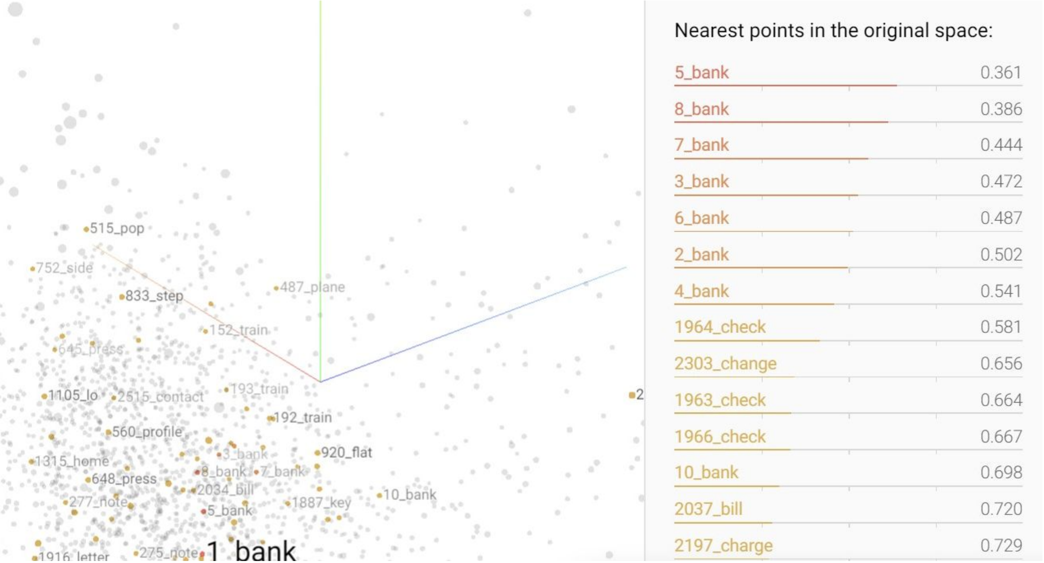


Source https://projector.tensorflow.org/

Word Embeddings to Contextual Models

Contextual Models are more close to Wordnet

sn	sentence/context
1	I have bank account.
2	Loan amount is approved by the bank.
3	He returned to office after he deposited cash in the bank.
4	They started using new software in their bank.
5	he went to bank balance inquiry.
6	I wonder why some bank have more interest rate than others.
7	You have to deposit certain percentage of your salary in the bank.
8	He took loan from a Bank.
9	he is waking along the river bank.
10	The red boat in the bank is already sold.
11	Spending time on the bank of Kaligandaki river was his way of enjoying in his childhood.
12	He was sitting on sea bank with his friend
13	She has always dreamed of spending a vacation on a bank of Caribbean sea.
14	Bank of a river is very pleasant place to enjoy.

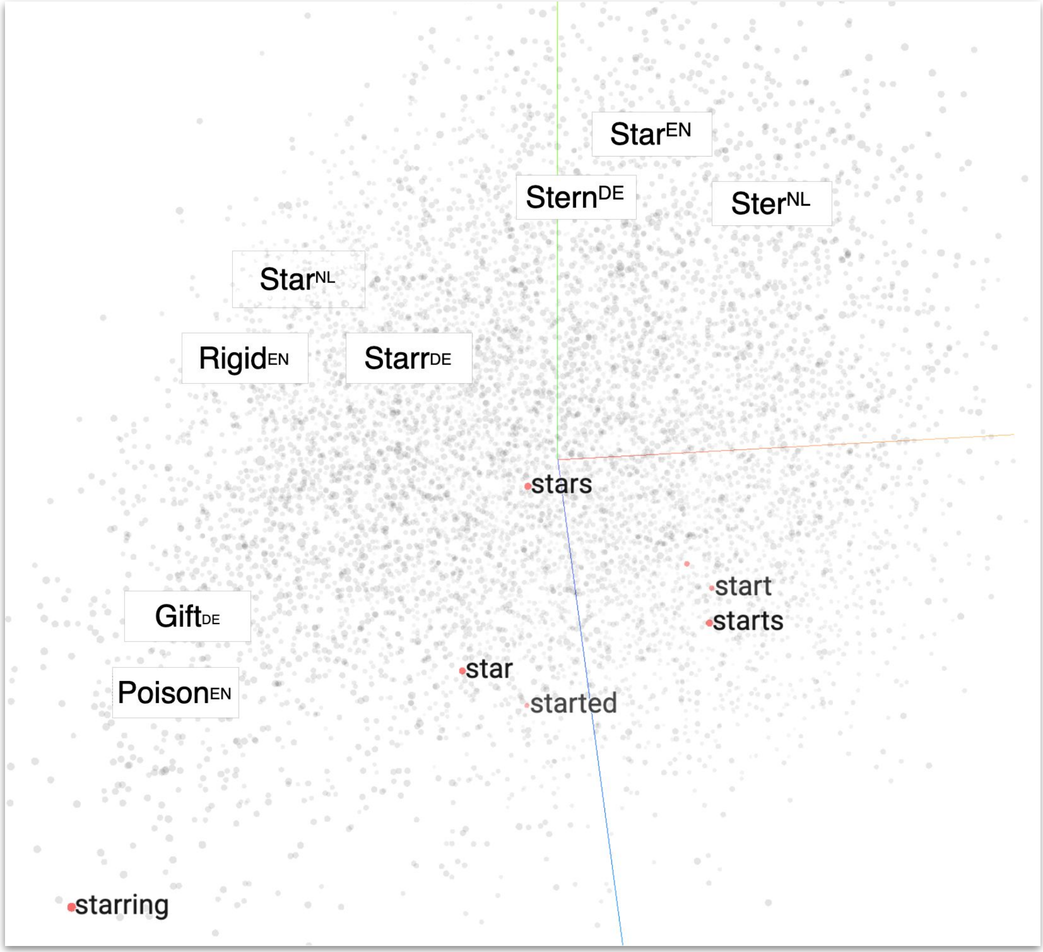


Source - https://projector.tensorflow.org/

Monolingual to Crosslingual

Properties of cross-lingual space:

- Interference (Wang et al., 2020)
- Transfer (Wu & Dredze, 2019)
- Curse of Multilinguality (Conneau et al. ,2020)



Research Question: measuring the degree of “Interference” to be corrected in context

- How well do cross-lingual models approximate meaning?
- How consistent is the relationship between words and concepts with and without the influence of context?
- What are the effects of sharing vocabulary and contexts across languages on the relationship between word and concept ?

Measuring Relationship

Let’s focus on words that illustrate distributional edge cases for the relation between concepts and context.

- Monosemous relation - one to one relation.
- Polysemous relations
  - Balanced - ambiguous words.
  - Skewed - one concept is dominant in language use.

Dataset

For our experiments, we use entities and their respective entity types as a proxy for a more general notion of words and concepts.

	EN	NL	DE
Sentences	17,942,551	12,429,622	5,512,929
Entities	4,219,046	6,737,100	2,917,688
Unique Entities	59,054	60,777	38,930
LOC	512,219	744,024	329,030
ORG	1,690,244	3,282,967	1,580,477
PER	2,016,583	2,710,109	1,008,181

Table 1 Statistics of entities distribution in XLEnt for English, Dutch and German

Measuring Relationship through Probing

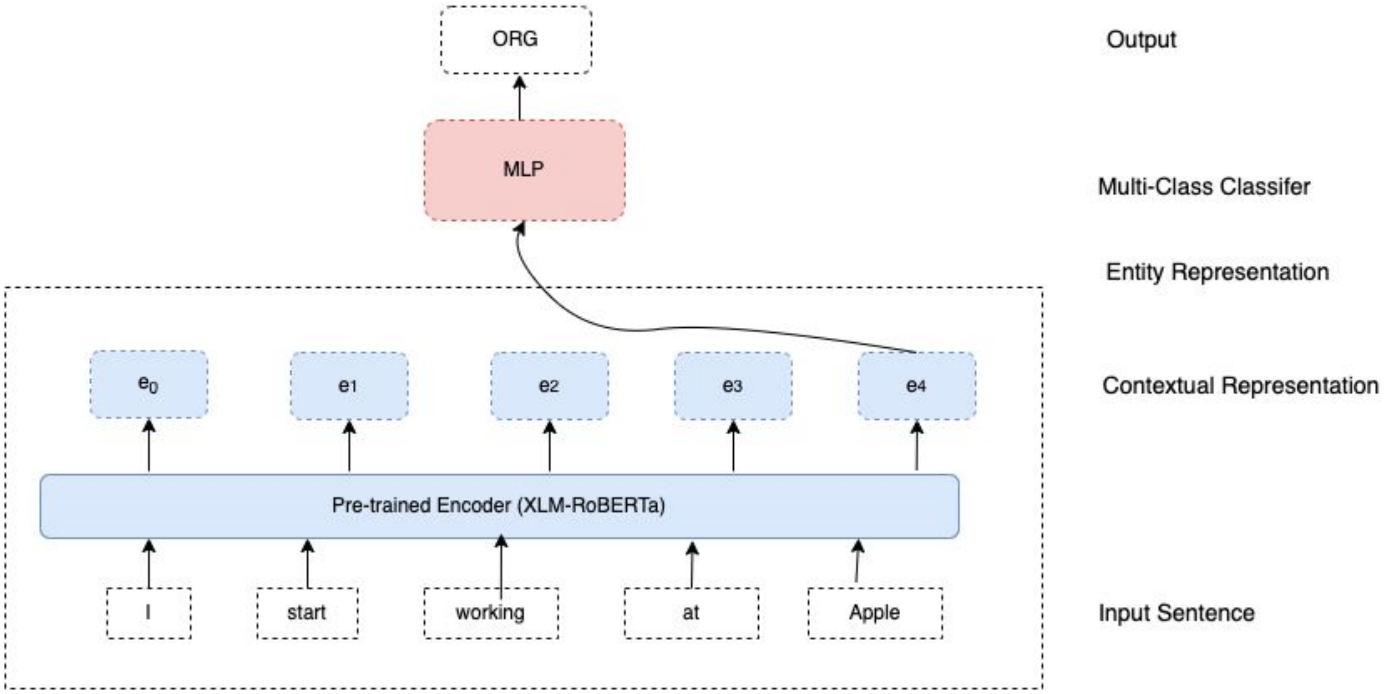


Figure 1: Architecture of our probing classifier

Hypothesis

Hypothesis I : Mono relationship

- minor differences between the lexical initialization level and higher contextual levels.

Hypothesis II : Skewed relationship

- Matching distribution for test cases: same as mono
- Diverging distribution for test cases: low probing accuracy on the lexical level, strong indications of concept sensitivity in higher levels

Hypothesis III : Balanced relationship

- Low probing accuracy at the lexical level, improved concept knowledge in higher levels in all cases but not as strong as for diverging

Measuring Ambiguity

- Effect of ambiguity is reflected in lower layers.
- Context is utilized in correcting concept ambiguity.

	LOC	ORG	PER
Balanced			
Layer-0	0.65	0.58	0.52
Layer-3	0.81	0.78	0.79
Skewed			
Layer-0	0.61	0.75	0.76
Layer-3	0.86	0.87	0.9

Table 1: F1 scores for probing the different layers of XLM-RoBERTa on Polysemy words

Measuring Bias

- Effect of bias is reflected in lower layers
- Context is utilized in correcting concept bias.

	LOC	ORG	PER
Skewed to LOC			
Layer-0	0.82	**0.38**	**0.25**
Layer-3	0.9	**0.63**	**0.73**
Skewed to ORG			
Layer-0	**0.24**	0.81	**0.34**
Layer-3	**0.85**	0.93	**0.75**
Skewed to PER			
Layer-0	**0**	**0**	0.97
Layer-3	**0.67**	**0.29**	0.97

Table 2: F1 scores for probing the different layers of XLM-RoBERTa on Polysemy skewed words

Measuring Interference - Directly

Interference is detected when there is diverging relation between words and concept.

	LOC	ORG	PER
Similar			
Layer-0	**0.76**	**0.67**	**0.78**
Layer-3	0.83	0.83	0.84
Diverging			
Layer-0	**0.57**	**0.53**	**0.42**
Layer-3	0.78	0.82	0.68

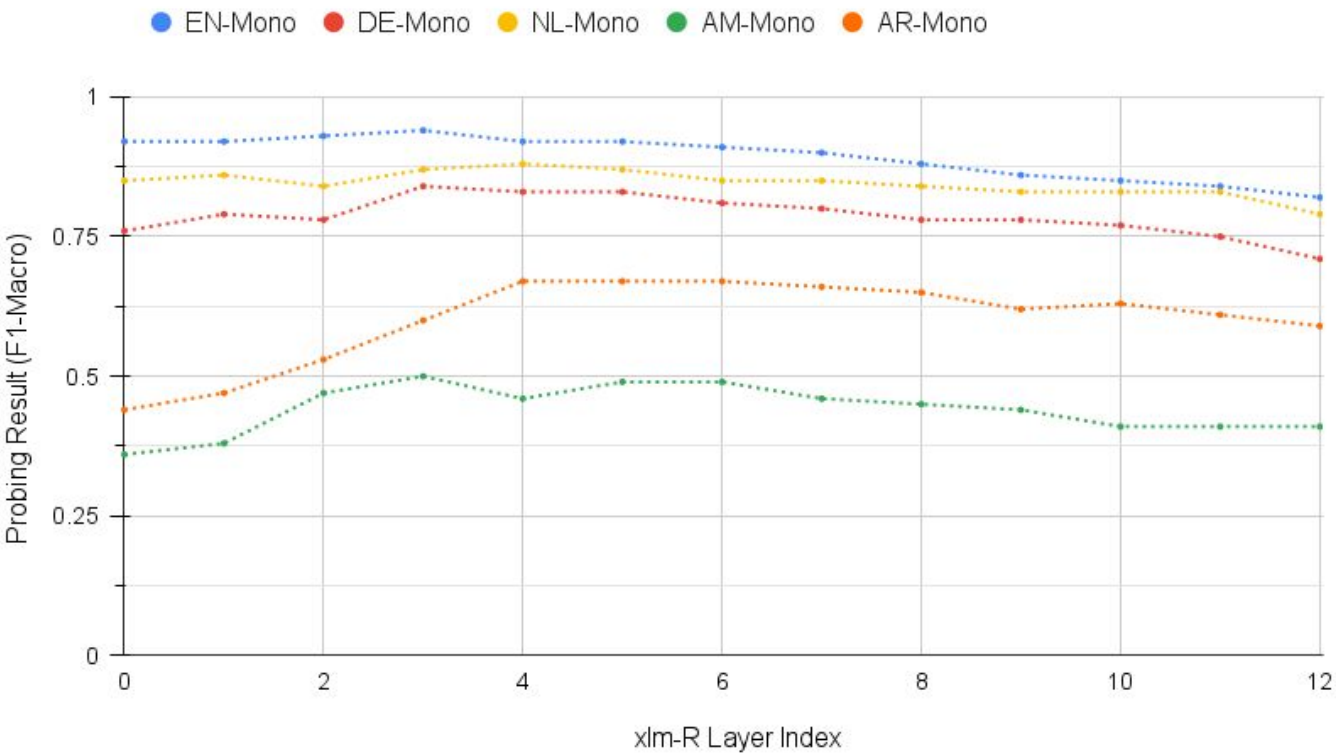
Table 3: (a) F1 scores for probing the different layers of XLM-RoBERTa on Polysemy & Shared words

	LOC	ORG	PER
Similar			
Layer-0	**0.74**	**0.59**	**0.79**
Layer-3	0.82	0.83	0.84
Diverging			
Layer-0	**0.54**	**0.51**	**0.45**
Layer-3	0.76	0.81	0.72

Table 3: (b) F1 scores for probing the different layers of m-BERT on Polysemy & Shared words

Measuring Transfer - Directly

Transfer is more strong between related languages



Zero-shot Transfer Evaluation

Transfer in Zero-shot (Pires et al., 2019;Wu and Dredze, 2019; Conneau et al., 2018b)

Conclusion

- Prior probabilities of polysemy profiles are reflected in the lexical initialization
- Contexts can recover the correct relationship between an ambiguous word and a concept to different degrees.
- Shared polysemy words either help or interfere with the model recovery capacity depending on the similarity of distribution across languages.
- Typological relationships between languages have a measurable impact on transfer.

Thank You