Interpretability of Sentence Embeddings in low-resource Languages

Master thesis final presentation

Supervisors: Dr. Steffen Eger, Dr. Johannes Daxenberger

TECHNISCHE UNIVERSITÄT DARMSTADT

Agenda

TECHNISCHE UNIVERSITÄT DARMSTADT

Section:

Introduction

TECHNISCHE
UNIVERSITÄT
DARMSTADT

Introduction

- A plethora of sentence embedding techniques has been developed
- **Problem:**
  The knowledge about what is captured by sentence embeddings is limited!
- **Probing tasks come to the rescue:**
  - *'Classification problem that focuses on simple linguistic properties of sentences'* (Conneau.2018)
  - Conneau.2018 introduced a **set of ten probing tasks**
  - E. g. sentence length, containment of words, subject number, tense, etc.
  - Conneau and colleagues mainly drew inspiration from Ettinger.2016, Shi.2016 and Adi.2017

Scope of this Thesis

- Most research in this domain is done for English/high-resource languages
- **Low-resource languages are mainly neglected**
- Languages considered in this thesis:

English	EN		high-resource
German	DE	Deutsch	high-resource
Russian	RU	русский язык	low-resource
Turkish	TR	Türkçe	low-resource
Georgian	KA	ქართული ენა	low-resource

- **Are patterns for English reproducible in low-resource languages?**

High-Level Process

TECHNISCHE UNIVERSITÄT DARMSTADT

❶ **Embeddings**	Train sentence encoders in multiple languages
❷ **Probing**	Data generation / Evaluation on probing tasks
❸ **Downstream**	Data generation / Evaluation on downstream applications
❹ **Stability**	Discrepancies with literature / different setups in literature: Investigate the rank stability of embeddings in various setups

Section:

Sentence Embeddings

TECHNISCHE
UNIVERSITÄT
DARMSTADT

Sentence Embedding Algorithms

TECHNISCHE UNIVERSITÄT DARMSTADT

- Vanilla average (300 d)
- *p*-Means (1,500 d)
- Geometric embeddings (300 d)
- Smooth inverse frequency (300 d)
- Hierarchical pooling (300 d)

Non-parametric

- InferSent (4,096 d)
- Quick-Thought (2,400 d)
- sent2vec (700 d)
- LASER (1,024 d)
- BERT (768 d)
- Random encoders (4,096/8,192 d)

Parametric

- Non-parametric: Aggregation of word embeddings **without training**
- Parametric models are **trained from scratch** on top of word embeddings

Section:

Probing and Downstream Tasks

TECHNISCHE
UNIVERSITÄT
DARMSTADT

Probing Task Examples

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Sentence Length (SentLen):**

  E. g.: **Label:** *short* **Sentence:** *It felt good to smile .*

  (A binning approach is used for the labels. Think of classes like *'short'*, *'medium'*, *'long'*)

- **Word Content (WC):**

  E. g.: **Label:** *everybody* **Sentence:** *Everybody should step back .*

- **Subject-Verb Agreement (SVAgree):**

  E. g.: **Label:** *disagree* **Sentence:** *They works together .*

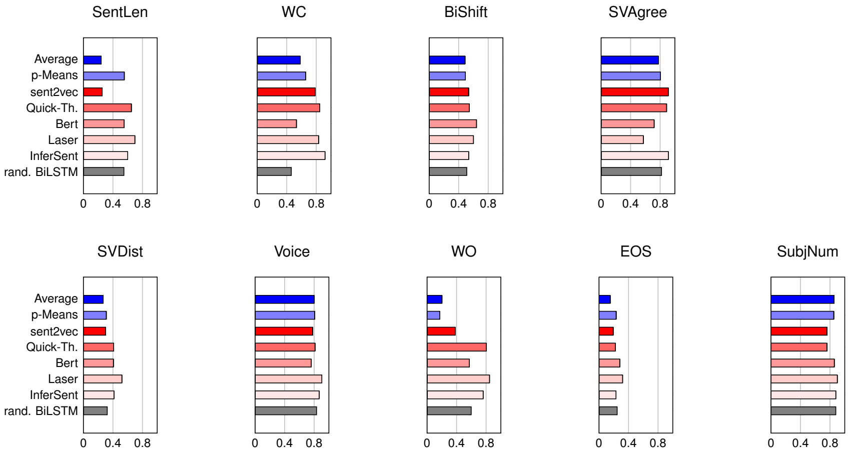
Probing Task Setup

TECHNISCHE UNIVERSITÄT DARMSTADT

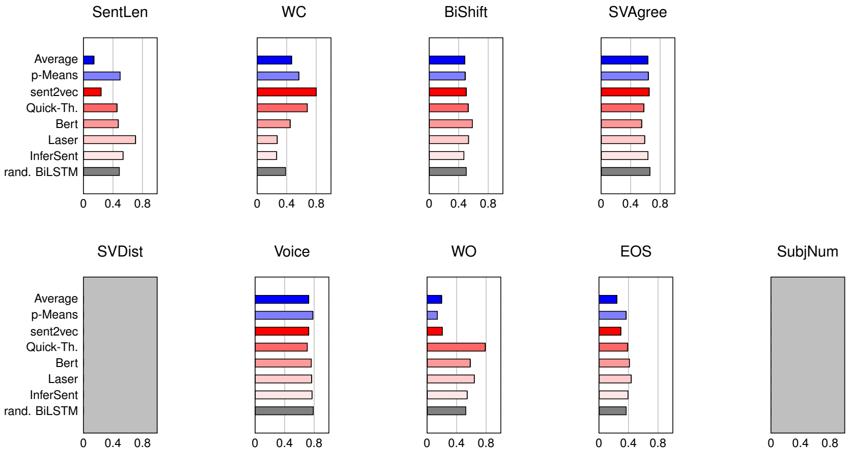
- Implementation of 9 probing tasks for EN, DE, RU, TR as well as 7 for KA (SENTLEN, WC, BISHIFT, SVAGREE, SVDIST*, VOICE, WO, EOS and SUBJNUM*)
- **New:** SVAGREE and SVDIST *(inspired by Linzen.2016)*
- Many tasks require corpora with **morpho-syntactic** annotations
- Universal Dependencies offers tree banks for many languages / GNC
- **Evaluation: MLP with 5-fold x-val**
  - One hidden layer with 50 hidden units
  - Dropout: 0.00
  - Activation: Sigmoid
  - Optimizer: Adam

* not implemented for KA

Probing Task Results for English



Probing Task Results for Georgian

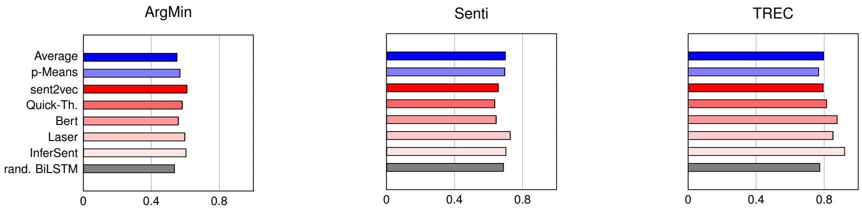


Downstream Tasks

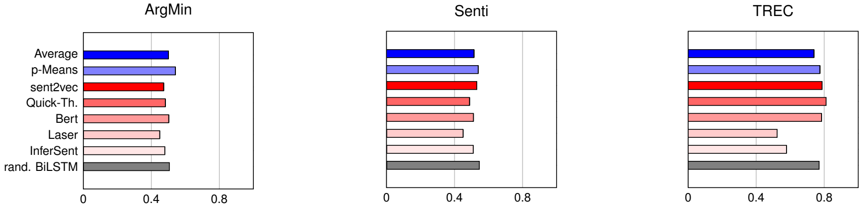
TECHNISCHE UNIVERSITÄT DARMSTADT

- **Downstream tasks:**
  1. Sentential argumentation mining (ARGMIN) $\rightarrow$ *translation necessary*
  2. Sentiment analysis (SENTI)
     - **EN:** US Airline Twitter data
     - **KA:** Own Twitter data set using Emojis as label indication *(Choudhary.2018)*
  3. Question type detection (TREC) $\rightarrow$ *translation necessary*
- **Evaluation:**
  Analogously to probing tasks, except for TREC
  ($\rightarrow$ pre-defined splits from *SentEval*)

Downstream Task Results (EN)



Downstream Task Results (KA)

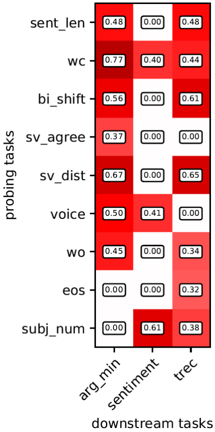


Summary Observations

- **More volatility** in probing tasks *(e. g.* SENTLEN*,* WO*)*
- **No** universal embedding *(Perone.2018)*
- Trained encoders tend to **work best** for English
- Averaging methods often provide **strong baseline** *(e. g.* SUBJNUM*)*
- Random encoders work surprisingly well *(Wieting.2019)*
- **Worse performance of trained models in low-resource languages** *(lack of training data)*

Correlations of Probing and Downstream Tasks

TECHNISCHE UNIVERSITÄT DARMSTADT



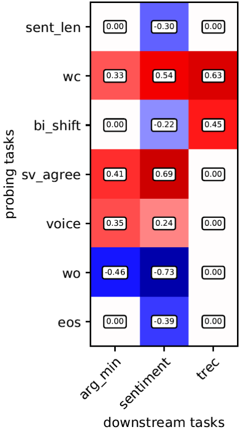
English language

- WC has high positive correlations *(intuitive)*
- TREC is correlated positively with **almost all probing tasks** *(found by Conneau.2018)*, also ARGMIN
- SENTI is less connected to probing tasks
- No negative correlations

positive / negative correlations

Correlations of Probing and Downstream Tasks

TECHNISCHE UNIVERSITÄT DARMSTADT



Georgian language

- WC has high positive correlations
- Many correlations **below an absolute value of 0.20 or negative**
- WO is negatively correlated
  (flexible word order in KA)
- **Correlations are language-dependent!**

positive / negative correlations

UKP

Section:

Stability Analysis

TECHNISCHE UNIVERSITÄT DARMSTADT

Stability Analysis

- Discrepancies with the literature were found
- Different evaluation setups:

`Size`	10k	$\Leftrightarrow$	90k+
`Class balance`	imbalanced	$\Leftrightarrow$	(im)balanced
`Classifier`	MLP	$\Leftrightarrow$	MLP / Logistic regression
`HP tuning`	no	$\Leftrightarrow$	yes (sometimes no)

- **A stability analysis is performed in order to investigate the effects of these factors**
- The word content task (English) is used as an example

Stability across Classifiers

10,000 instances

	LR	NN	NN_H	RF	SVM	NB
LR	1,00					
NN	0,91	1,00				
NN_H	0,83	0,94	1,00			
RF	0,41	0,66	0,81	1,00		
SVM	0,85	0,90	0,94	0,76	1,00	
NB	0,84	0,83	0,76	0,34	0,78	1,00

- Rankings are quite unstable, especially for `RF` classifier
- `NN` and `LR` are similar, also `NN` and `NN_H`
- **Recommendation: Use a neural architecture** (outperforms other classifiers)

Stability across Data Set Sizes

TECHNISCHE UNIVERSITÄT DARMSTADT

NN

	5,000	10,000	30,000	60,000
5,000	1,00			
10,000	0,98	1,00		
30,000	0,90	0,96	1,00	
60,000	0,88	0,94	0,99	1,00

NN_H

	5,000	10,000	30,000	60,000
5,000	1,00			
10,000	0,95	1,00		
30,000	0,87	0,94	1,00	
60,000	0,84	0,92	0,99	1,00

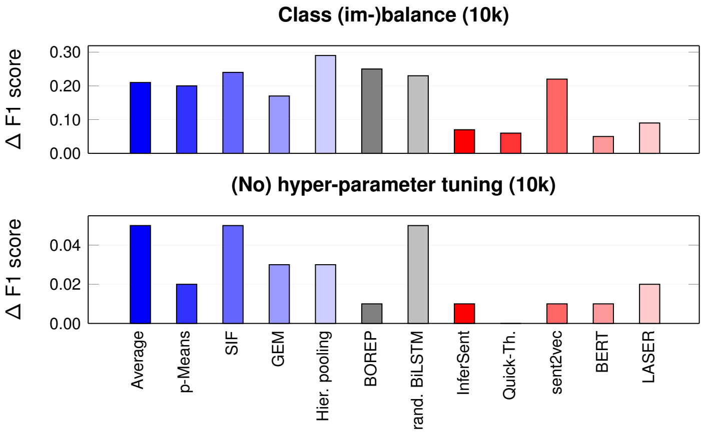
RF

	5,000	10,000	30,000	60,000
5,000	1,00			
10,000	0,99	1,00		
30,000	0,97	0,97	1,00	
60,000	0,97	0,97	1,00	1,00

- Correlations between 5k ⇔ { 10k, 30k, 60k } decrease
- However, high correlations for `RF` (less data sufficient for stable ranking)
- Correlations between 30k ⇔ 60k close to 1.0
- **Recommendation: Use at least 30k instances**

Effects of Class Balance and HP Tuning

TECHNISCHE UNIVERSITÄT DARMSTADT



Section:

Summary

TECHNISCHE UNIVERSITÄT DARMSTADT

Summary

TECHNISCHE
UNIVERSITÄT
DARMSTADT

- The gap between trained encoders and compositional models **vanishes in low-resource languages**
- Correlations in English and Georgian differ (e. g. no word order in Georgian)
- Nevertheless, the **results should be treated with caution**:
  - Use balanced data sets *(considerable impact on ranking)*
  - Use at least 30k instances
  - Use an MLP with hyper-parameter tuning
- **The evaluation should be agnostic to factors like class balance or data set size** *(probing tasks suboptimal?)* $\rightarrow$ **future research**

Thank you very much for your attention!

Presenter:	Daniel Wehner
Date:	October 15, 2019
Topic:	Interpretability of sentence embeddings in low-resource languages

TECHNISCHE UNIVERSITÄT DARMSTADT

Universal Dependencies - Example

1	But	but	CC	_	8:cc
2	in	in	IN	_	4:case
3	my	my	PRP$	Number=Sing|Person=1|Poss=Yes	4:nmod:poss
4	view	view	NN	Number=Sing	8:obl
5	it	it	PRP	Case=Nom|Gender=Neut|Number=Sing	8:nsubj
6	is	be	VBZ	Mood=Ind|Number=Sing|Person=3	8:cop
7	highly	highly	RB	_	8:advmod
8	significant	significant	JJ	Degree=Pos	0:root
9	.	.	.	_	8:punct

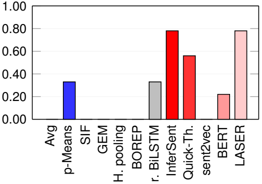
Winner Statistics (Probing Tasks)

Top-three counts					
Embedding	**EN**	**DE**	**RU**	**TR**	**KA**
Vanilla Average	**0** (0.00)	**1** (0.11)	**3** (0.33)	**1** (0.11)	**0** (0.00)
p-Means	**3** (0.33)	**1** (0.11)	**3** (0.33)	**3** (0.33)	**4** (0.57)
SIF	**0** (0.00)	**0** (0.00)	**1** (0.11)	**1** (0.11)	**0** (0.00)
GEM	**0** (0.00)	**1** (0.11)	**1** (0.11)	**2** (0.22)	**0** (0.00)
hier. Pooling	**0** (0.00)	**0** (0.00)	**0** (0.00)	**0** (0.00)	**0** (0.00)
BOREP	**0** (0.00)	**1** (0.11)	**0** (0.00)	**0** (0.00)	**1** (0.14)
Random BiLSTM	**3** (0.33)	**2** (0.22)	**3** (0.33)	**2** (0.22)	**2** (0.29)
InferSent	**7** (0.78)	**2** (0.22)	**1** (0.11)	**1** (0.11)	**3** (0.43)
Quick-Thought	**5** (0.56)	**5** (0.56)	**5** (0.56)	**4** (0.44)	**2** (0.29)
sent2vec	**0** (0.00)	**4** (0.44)	**2** (0.22)	**1** (0.11)	**2** (0.29)
BERT	**2** (0.22)	**3** (0.33)	**3** (0.33)	**5** (0.56)	**3** (0.43)
LASER	**7** (0.78)	**7** (0.78)	**5** (0.56)	**8** (0.89)	**4** (0.57)

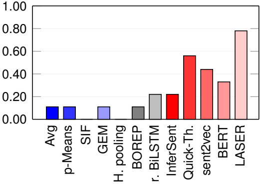
Winner Statistics (Probing Tasks)

TECHNISCHE UNIVERSITÄT DARMSTADT

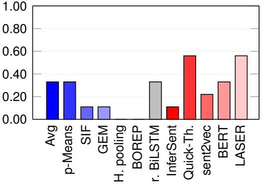
Top-three scores (EN)



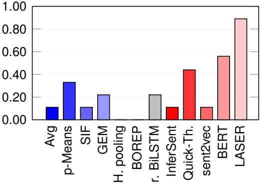
Top-three scores (DE)



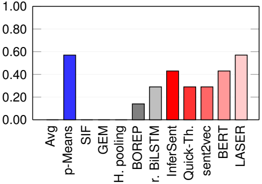
Top-three scores (RU)



Top-three scores (TR)



Top-three scores (KA)



Effect of Hyper-Parameter Tuning (other Tasks)

TECHNISCHE UNIVERSITÄT DARMSTADT

Effect of hyper-parameter tuning on other tasks			
	Δ SENTI	Δ VOICE	Δ SUBJNUM
Vanilla average	0.02	0.00	0.00
p-Means	0.00	0.02	0.03
BOREP	0.00	0.02	0.03
Random BiLSTM	0.00	0.01	0.00
InferSent	0.05	0.03	0.03
Quick-Thought	0.03	0.02	0.04
sent2vec	0.00	0.00	0.01
BERT	0.01	0.02	0.02
LASER	0.01	0.03	0.02

Stability Analysis: Effects on Ranking

TECHNISCHE UNIVERSITÄT DARMSTADT

	Vanilla average	p-Means	SIF	GEM	hier. pooling	BOREP	Random BiLSTM	InferSent	Quick-Thought	sent2vec	BERT	LASER
❶ class (im-)balance												
Ranking *(imbalanced, 10k)*	8	6	7	5	12	9	11	1	3	4	9	2
Ranking *(balanced, 10k)*	8	7	6	5	11	9	10	2	4	1	12	3
				Spearman correlation: 0.91								
❷ (no) hyper-parameter tuning												
Ranking *(balanced, no optimization, 10k)*	8	7	6	5	11	9	10	2	4	1	12	3
Ranking *(balanced, optimization, 10k)*	7	8	4	4	11	10	9	2	6	1	12	3
				Spearman correlation: 0.95								
❹ size (30k ↔ 60k)												
Ranking *(balanced, 30k)*	6	6	5	8	11	10	9	2	4	1	12	3
Ranking *(balanced, 60k)*	7	5	5	9	11	10	8	1	4	1	12	3
				Spearman correlation: 0.98								