

Interpretability of sentence embeddings in low-resource languages

Mid-term presentation; July 2, 2019



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Agenda

- 1 Introduction
- 2 Embedding Algorithms
- 3 Probing Tasks
- 4 Downstream Applications
- 5 Results
- 6 Summary

Section:
Introduction



TECHNISCHE
UNIVERSITÄT
DARMSTADT

- ▶ Sentence embeddings are vectorial representations of sentences.
- ▶ They can be fed into ML classifiers.
- ▶ They have become an **integral part** in NLP applications.
 - ▶ Boosted downstream task performance.
 - ▶ Plethora of algorithms like *InferSent*, *sent2vec*, *Quickthought*, and many more!
- ▶ Neural networks are often used to obtain such embeddings.
- ▶ **Problem:** Neural networks are black-box models!
- ▶ Workshops like **BlackBoxNLP** attempt to shed light onto this and open the black-box.





- ▶ It is advantageous to know what is captured by an embedding.
 - ▶ Relevant for the choice of an embedding technique for a specific task.
 - ▶ Allows for a quality assessment of sentence embeddings.
- ▶ Many authors (Shi.2016, Adi.2017, Conneau.2018; inter alia) introduce the notion of **probing tasks**.
- ▶ What are probing tasks?



- ▶ Simple classification tasks that probe an embedding for **linguistic properties**.
- ▶ Mostly restricted to the English language in the literature.

Scope of this Thesis

- ▶ The focus in this thesis is on **lower-resource languages**.

- ▶ Languages considered:

English (EN)

German (DE)

Russian (RU)

Turkish (TR)

Georgian (KA)

Deutsch

русский язык

Türkçe

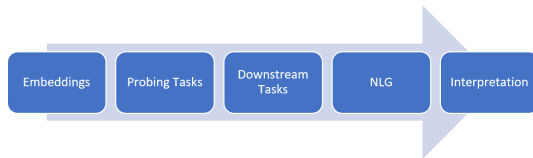
ქართული ენა

low-resource

low-resource

low-resource

- ▶ Process:



Current Status



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Item	Status
Literature review	●
Implementation of...	
...word embeddings	✓
...sentence embeddings	✓
...probing tasks	✓
...downstream tasks	✓
Training of...	
...word embeddings	✓
...sentence embeddings	✓
Translation of resources	●
Natural language generation	✗
Interpretation of results	●
Writing thesis	●

Legend: ✓ done ● in progress ✗ not started

Section:
Embedding Algorithms



TECHNISCHE
UNIVERSITÄT
DARMSTADT



- ▶ Two types (Yang.2018):
 - ▶ **Non-parametric:** No training required, average of word embeddings.
 - ▶ **Parametric:** Sentence embeddings are trained from scratch.
- ▶ Selected algorithms for the experiments:
 - ▶ Bag of vectors (vanilla average; 300 d)
 - ▶ p -means ($p \in \{1, 2, 3, -\infty, +\infty\}$; 1,500 d)
 - ▶ Geometric embeddings (GEM; 300 d)
 - ▶ Smooth inverse frequency (SIF; 300 d)
 - ▶ Hierarchical pooling (300 d)
 - ▶ Random embeddings (4,096 / 8,192 d)
 - ▶ InferSent (2,048 d)
 - ▶ Quickthought (2,400 d)
 - ▶ sent2vec (700 d)
 - ▶ LASER (1,024 d)
 - ▶ BERT (768 d)

Non-parametric

Parametric

- ▶ English pre-trained models were already available for download.
- ▶ Other languages:
 - ▶ Trained word2vec on respective Wikipedia dumps (CBow, window size 10).
 - ▶ Fasttext and attract-repel embeddings available for all languages.
 - ▶ InferSent requires translation of SNLI for all languages.
 - ▶ Quickthought / sent2vec are trained on sentences extracted from Wikipedia.
 - ▶ LASER / BERT: Multi-lingual models are already available.
- ▶ Google translate API limits the number of translations per day.
⇒ **The translation process is slow!**

≈ 10,000 sentences / day.



Section:
Probing Tasks



TECHNISCHE
UNIVERSITÄT
DARMSTADT

- ▶ Conneau.2018 distinguishes between three types of probing tasks:
Surface, **Syntax** and **Semantics**
- ▶ Possible probing tasks (many more in the literature):

① Surface

Sentence length*
Word content*
...

② Syntax

Bi-gram shift*
Subject-verb agreement*
Subject-verb distance*
Top constituent*
Tree depth
Voice*
Word order*
...

③ Semantics

End of sentence*
Grammaticality
Object number
Subject number
Tense
...

* $\hat{=}$ implemented

- ▶ Limited the number of instances to 10,000.

Probing Task: Sentence Length



TECHNISCHE
UNIVERSITÄT
DARMSTADT



- ▶ **Category:**
Surface probing task.
- ▶ **Description:**
Given a sentence embedding predict the lengths of the original sentence.
- ▶ **Implementation:**
The probing task is phrased as a 10-way classification task. The bins are as follows:
[1;4], [5;8], [9;12], [13-16], [17;20], [21;25], [26;29], [30;33], [34;55], [56;)
- ▶ **Languages:** EN ✓ DE ✓ RU ✓ TR ✓ KA ✓
- ▶ **Example:**
The fox jumped over the lazy dog . \Rightarrow Class 1 [5;8]

Probing Task: Subject-Verb Agreement



TECHNISCHE
UNIVERSITÄT
DARMSTADT



- ▶ **Category:**

Syntactic probing task.

- ▶ **Description:**

A classifier has to predict whether the verb form agrees with the form required by the subject of the sentence.

- ▶ **Implementation:**

The task is restricted to present tense. For each language the most common verbs and their conjugations were acquired. Each sentence from the corpus was tested if it contains one of the words from the list. If it contains one such word it is replaced by different word from the list.

- ▶ **Languages:** EN ✓ DE ✓ RU ✓ TR ✓ KA ✓

- ▶ **Example:**

The plane_{subj} from London to Paris take_{verb} off without any delay . ⇒ Class 'Disagree'

Probing Task: End of Sentence



TECHNISCHE
UNIVERSITÄT
DARMSTADT



- ▶ **Category:**
Semantic probing task.
- ▶ **Description:**
Decide where a sequence of words has to be split (end of one sentence, begin of another).
- ▶ **Implementation:**
Two sentences were concatenated with punctuation removed. Also, the words were lower-cased (except for German) in order not to provide any hints. The possible split indices were again binned.
- ▶ **Languages:** EN ✓ DE ✓ RU ✓ TR ✓ KA ✓
- ▶ **Example:**
this is great || where are you

Section:

Downstream Applications



TECHNISCHE
UNIVERSITÄT
DARMSTADT



- ▶ Downstream applications considered:

- ▶ **Argumentation mining** (Labels: FOR, AGAINST, NONE)
- ▶ **Sentiment analysis** (Labels: POS, NEG, NEU)
- ▶ (Natural language inference)



- ▶ Data sets:

- ▶ Argumentation mining:
 - ▶ Data set contains $\approx 25,000$ sentences.
 - ▶ Various topics: abortion, nuclear energy, gun control, marijuana legalization, ...
 - ▶ Has to be translated into all target languages. ✓
- ▶ Sentiment analysis:
 - ▶ Different data set for each language.
 - ▶ Created own data set for Georgian (since none could be found).



Section:
Results



TECHNISCHE
UNIVERSITÄT
DARMSTADT

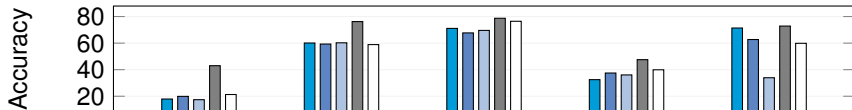
- ▶ The classifier is a Multi-Layer-Perceptron (MLP).

- ▶ MLP architecture:

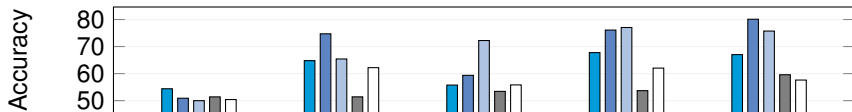
# hidden layers:	1
# hidden units:	300
Hidden layer activation func.:	ReLU
Dropout rate:	0.50
# epochs:	30
Optimizer:	RmsProp
Loss function:	Categorical cross-entropy

- ▶ Later experiments will also be conducted without the hidden layer.

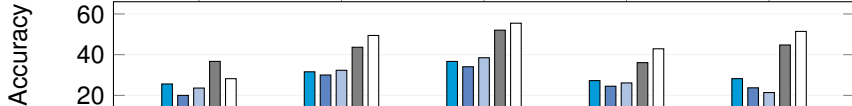
Results Sentence Length Task



Results Subject-Verb Agreement Task



Results End of Sentence Task



EN DE RU TR KA

Probing Task Results for Georgian



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Language: Georgian								
Embedding	Sentence length	End of sentence	Word content	Word order	Bi-gram shift	Voice	SV agreement	SV distance
Size of the data set								
# instances	10,000	10,000	8,674	262	10,000	10,000	7,786	n. i.
Majority class (baseline)								
Majority	21.300	28.200	13.000	33.500	50.100	69.100	50.400	n. i.
Random embeddings (baseline)								
BOREP	47.640	46.000	57.157	33.084	50.570	79.390	61.847	n. i.
ESN	46.439	43.129	30.956	32.309	50.220	76.800	53.178	n. i.
Random LSTM	54.450	49.590	45.348	35.748	50.460	79.850	59.895	n. i.
Average embeddings (fasttext)								
Vanilla average	33.400	39.200	64.786	33.075	49.630	79.040	64.415	n. i.
p-Means	58.890	49.430	64.829	33.084	50.260	81.460	62.206	n. i.
SIF	30.780	37.610	64.485	32.701	48.990	78.260	64.056	n. i.
GEM	39.660	43.420	83.815	32.692	49.830	79.040	63.799	n. i.
hier. pooling	54.870	49.400	55.816	32.683	49.710	80.580	62.271	n. i.
Trained embeddings								
InferSent	△	△	△	△	△	△	△	n. i.
QuickThought	59.860	51.420	73.707	84.403	54.300	76.440	57.622	n. i.
sent2vec	39.920	42.880	91.711	33.467	50.450	75.820	62.040	n. i.
BERT	59.789	50.170	54.005	46.382	57.090	78.940	54.604	n. i.
LASER	76.450	55.450	40.705	45.293	55.630	79.440	55.824	n. i.

Training Set Size matters!

Example for EN:

Word content task / p-means embedding				
p	hidden	size of data set	accuracy	majority
{1,2,3,min,max}	TRUE	1.028	20,81	20,00
		10.000	45,80	13,00
		55.884	67,12	11,00
		140.451	74,96	11,00
	FALSE	140.451	81,87	11,00
		448.418	88,86	11,00
{1, min, max}	FALSE	448.418	88,00	11,00



Language	Accuracy	Embedding
-	56.200	Majority
EN	65.407	sent2vec
DE	65.080	LASER
RU	63.873	LASER
TR	64.106	LASER
KA	61.867	SIF

- ▶ For downstream tasks it would be better to compute F1-scores.
- ▶ It is not necessary to achieve state-of-the-art performance.
It is more important to be able to compare embeddings given an architecture.



- ▶ Trained models usually perform best (except for Georgian; models might need more training data).
- ▶ All embeddings are above the majority baseline.
- ▶ High accuracy for random embeddings probably due to high dimensionality.
- ▶ LASER performs well on sentence length, end of sentence, voice and subject-verb distance.
- ▶ s2v shows strong performance on word content, otherwise bad performance.
- ▶ Quickthought performs well on word order task.
- ▶ **Training data set size matters!**

Section:
Summary



TECHNISCHE
UNIVERSITÄT
DARMSTADT

- ▶ It is **not clear** what linguistic properties are captured by sentence embeddings.
- ▶ **Probing tasks** are an attempt to shed light onto this.
- ▶ Downstream applications evaluate embeddings in **real applications**.
- ▶ **Outlook:**
 - ▶ Acquire **larger probing data sets** (especially for word content and word order task).
 - ▶ **Interpretation of results** and comparison to literature.
 - ▶ Implementation of **NLG**.

Thank you very much for your attention!

Presenter:

Daniel Wehner

Date:

July 2, 2019

Topic:

Interpretability of sentence embeddings
in low-resource languages



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Section:
Appendix



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Section:
Further Probing Tasks



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Probing Task: Word Content



TECHNISCHE
UNIVERSITÄT
DARMSTADT



- ▶ **Category:**
Surface probing task.
- ▶ **Description:**
Decide which of 30 possible words is contained in the sentence.
- ▶ **Implementation:**
30 mid-frequency nouns were chosen for each language. Each sentence in the data set is chosen such that **exactly one** noun from the list is contained in the sentence.
- ▶ **Languages:** EN ✓ DE ✓ RU ✓ TR ✓ KA ✓
- ▶ **Example:**
Federal elections took place last year. \Rightarrow Class 'Year'



- ▶ **Category:**
Syntactic probing task.
- ▶ **Description:**
Decide if a word is located at the beginning, the end or in the middle of a sentence (3-way classification).
- ▶ **Implementation:**
The most common noun in the corpus was chosen. The data set comprises sentences which contain this word **exactly once**. The word considered to be at the beginning/end if it is among the first/last 5 words in the sentence (4 for Turkish and Georgian due to agglutinative property).
- ▶ **Languages:** EN ✓ DE ✓ RU ✓ TR ✓ KA ✓
- ▶ **Example:**
This year a new software || update was released which patched || most of the vulnerabilities .
⇒ Class '*Beginning*'

Probing Task: Bi-Gram Shift

- ▶ **Category:**
Syntactic probing task.
- ▶ **Description:**
Tests for a legal word order. The classifier has to decide if a bi-gram was switched or not.
- ▶ **Implementation:**
In 50 % of the time a word in the sentence was picked randomly which subsequently switched positions with its right neighbor. In the other cases the sentences remained unaltered.
- ▶ **Languages:** EN ✓ DE ✓ RU ✓ TR ✓ KA ✓
- ▶ **Example:**
This is awesome . (Class 0) \Rightarrow This awesome is . (Class 1)



- ▶ **Category:**
Syntactic probing task.
- ▶ **Description:**
Probes a sentence embedding whether it encodes information about the voice of the sentence (active / passive).
- ▶ **Implementation:**
Sentences from the corpus containing a passive construct were labeled accordingly.
- ▶ **Languages:** EN ✓ DE ✓ RU ✓ TR ✓ KA ✓
- ▶ **Example:**
This picture was painted by Leonardo Da Vinci. ⇒ Class '*Passive*'

Probing Task: Subject-Verb Distance



TECHNISCHE
UNIVERSITÄT
DARMSTADT

- ▶ **Category:**

Syntactic probing task.

- ▶ **Description:**

A classifier has to predict how far subject and verb are apart.

- ▶ **Implementation:**

Similar to the sentence length task the number of words is binned. The bins were chosen as follows: [1], [2;4], [5;7], [8;12], [13;). Given a sentence representation the classifier has to decide for a bin. The task is not implemented for Georgian since the corpus does not include dependency parses.

- ▶ **Languages:** EN ✓ DE ✓ RU ✓ TR ✓ KA ✗

- ▶ **Example:**













The plane_{subj} from London to Paris takes_{verb} off without any delay . ⇒ Class 1 [2;4]

Probing Task: Top Constituent

- ▶ **Category:**
Syntactic probing task.
- ▶ **Description:**
Sentences are labeled with their top constituent. The task is formulated as a 10-way classification task.
- ▶ **Implementation:**
The Stanford Parser was used for English and German. For Georgian a small data set was made available. The 9 most frequent top constituents were taken as a label. The rest was labeled with 'Other'.
- ▶ **Languages:** EN ✓ DE ✓ RU ✗ TR ✗ KA ✓
- ▶ **Example:**
The children went to school. \Rightarrow NP VP .

Sentiment Data Set for Georgian

- ▶ Collected tweets from Georgian Twitter and labeled tweets with sentiment based on emojis used (Choudhary.2018).

Positive sentiment		Neutral sentiment		Negative sentiment	
Emoji	Short name	Emoji	Short name	Emoji	Short name
	:heart_eyes:		:no_mouth:		:cry:
	:grinning:		:thinking_face:		:angry:
	:grin:		:neutral_face:		:rage:
	:joy:		:smirk:		:sob:

- ▶ Procedure:
 1. Got a list of most frequent words.
 2. Search term: One word from the list and an emoji.
 3. Non-Georgian text was filtered / hashing to prevent duplicate tweets.
 4. The data set contains around 13,000 tweets.

Section:
Detailed Results



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Probing Task Results for English



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Language: English								
Embedding	Sentence length	End of sentence	Word content	Word order	Bi-gram shift	Voice	SV agreement	SV distance
Size of the data set								
# instances	9,997	5,283	1,028	557	9,997	9,997	5,485	6,045
Majority class (baseline)								
Majority	17.900	25.600	20.500	33.333	50.500	90.200	54.400	49.600
Random embeddings (baseline)								
BOREP	44.865	28.065	21.004	33.333	50.025	90.187	64.036	53.458
ESN	42.189	25.625	20.512	33.333	50.255	90.187	50.747	50.249
Random LSTM	55.229	31.224	20.512	33.333	50.615	90.427	63.926	57.924
Average embeddings (fasttext)								
Vanilla average	31.600	26.796	22.372	33.333	51.510	91.200	67.408	59.726
p-Means	60.091	31.584	20.805	33.333	49.800	90.480	64.801	57.130
SIF	28.300	26.476	21.101	33.333	50.620	90.800	65.585	59.230
GEM	36.530	27.403	59.881	32.796	50.550	90.600	60.171	56.747
hier. pooling	54.787	30.202	20.608	34.229	51.485	89.937	64.072	60.669
Trained embeddings								
InferSent	△	△	△	△	△	△	△	△
QuickThought	71.379	28.219	30.110	69.713	52.660	90.100	67.026	63.066
sent2vec	32.490	27.252	62.631	38.172	51.570	91.210	67.773	59.047
BERT	53.989	33.006	40.812	50.179	63.920	90.500	62.650	64.637
LASER	71.090	36.675	21.099	65.233	58.040	93.650	55.760	68.790

Probing Task Results for German

Language: German								
Embedding	Sentence length	End of sentence	Word content	Word order	Bi-gram shift	Voice	SV agreement	SV distance
Size of the data set								
# instances	10,000	6,708	977	766	10,000	10,000	3,432	10,000
Majority class (baseline)								
Majority	19.900	20.000	04.200	33.400	50.200	80.400	50.900	43.600
Random embeddings (baseline)								
BOREP	41.372	23.244	06.056	33.377	49.660	82.240	78.491	46.320
ESN	26.845	18.126	04.175	33.377	50.060	80.430	50.596	43.600
Random LSTM	56.030	27.183	09.699	37.133	50.570	85.360	75.357	48.379
Average embeddings (fasttext)								
Vanilla average	30.220	23.462	20.825	30.530	49.070	93.230	77.504	49.610
p-Means	59.311	30.004	13.302	33.247	49.520	90.020	74.719	46.940
SIF	27.120	22.517	26.061	27.945	49.490	87.500	76.139	47.110
GEM	34.340	21.515	56.152	23.416	49.640	94.450	75.704	47.060
hier. pooling	52.150	27.402	16.618	32.598	49.890	87.390	74.659	47.880
Trained embeddings								
InferSent	△	△	△	△	△	△	△	△
QuickThought	62.721	23.715	22.566	39.198	58.840	94.570	80.117	58.341
sent2vec	37.480	24.495	88.854	24.711	51.290	93.630	76.109	50.320
BERT	49.849	29.641	39.288	42.945	63.910	88.780	72.830	52.280
LASER	67.690	34.060	28.300	41.145	58.460	90.660	59.362	55.750

Probing Task Results for Russian

Language: Russian								
Embedding	Sentence length	End of sentence	Word content	Word order	Bi-gram shift	Voice	SV agreement	SV distance
Size of the data set								
# instances	10,000	10,000	3,213	2,123	10,000	10,000	4,017	10,000
Majority class (baseline)								
Majority	17.400	23.600	04.200	33.400	51.300	81.300	50.000	42.300
Random embeddings (baseline)								
BOREP	40.389	26.121	08.534	33.318	49.870	81.230	58.074	43.901
ESN	37.241	23.570	04.147	33.318	50.180	81.230	50.484	42.260
Random LSTM	56.390	29.220	14.485	35.045	51.280	81.430	65.840	48.210
Average embeddings (fasttext)								
Vanilla average	31.580	25.888	32.809	26.598	50.660	83.050	72.637	49.590
p-Means	60.240	32.324	19.179	33.318	50.030	81.260	65.416	48.010
SIF	28.870	24.353	39.641	23.566	50.230	84.000	73.307	48.820
GEM	36.401	28.142	68.829	19.132	50.630	84.270	73.158	45.720
hier. pooling	54.610	00.000	20.573	33.878	50.400	81.580	67.551	48.950
Trained embeddings								
InferSent	△	△	△	△	△	△	△	△
QuickThought	33.937	21.397	04.147	70.787	52.630	81.980	75.738	44.879
sent2vec	36.050	26.121	94.986	23.565	50.130	82.860	77.053	48.680
BERT	56.400	32.662	34.136	47.783	59.970	82.410	63.855	49.810
LASER	69.580	38.489	41.130	67.010	58.760	87.440	72.240	57.740

Probing Task Results for Turkish



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Language: Turkish								
Embedding	Sentence length	End of sentence	Word content	Word order	Bi-gram shift	Voice	SV agreement	SV distance
Size of the data set								
# instances	8,416	4,242	425	101	8,416	8,416	390	2,750
Majority class (baseline)								
Majority	43.000	36.700	00.800	33.333	50.200	86.200	51.400	39.000
Random embeddings (baseline)								
BOREP	57.225	38.242	11.446	33.333	51.128	86.352	50.135	44.418
ESN	49.512	32.587	07.993	33.333	50.446	86.352	50.382	39.004
Random LSTM	70.636	45.032	10.328	31.313	53.864	86.834	51.407	46.710
Average embeddings (fasttext)								
Vanilla average	46.711	37.491	24.021	33.333	50.975	87.937	52.946	45.728
p-Means	76.227	43.646	19.187	33.333	50.094	87.104	51.407	50.091
SIF	45.572	37.373	24.722	33.333	50.752	88.114	49.115	46.529
GEM	50.282	37.539	57.556	34.259	50.317	88.313	50.644	46.310
hier. pooling	69.744	44.021	16.777	30.471	50.893	86.869	49.117	47.764
Trained embeddings								
InferSent	△	△	△	△	△	△	△	△
QuickThought	72.845	44.749	29.995	47.054	60.395	87.515	59.583	50.200
sent2vec	47.545	36.081	64.263	35.354	50.012	86.939	53.703	46.091
BERT	70.038	48.036	21.560	42.929	58.233	87.456	50.893	50.016
LASER	78.729	52.031	20.240	34.259	59.467	89.746	53.439	54.382

Results for Argumentation Mining

Task: Argumentation Mining					
Embedding	English	German	Russian	Turkish	Georgian
Size of the data set					
# instances	15,000				
Majority class (baseline)					
Majority	56.200				
Random embeddings (baseline)					
BOREP	56.240	56.180	56.140	55.907	57.087
ESN	56.240	56.180	56.140	55.907	55.987
Random LSTM	56.767	56.600	57.147	57.427	58.333
Average embeddings (fasttext)					
Vanilla average	62.873	60.560	61.887	62.187	61.027
p-Means	58.013	57.887	56.913	58.526	57.607
SIF	64.827	62.880	63.260	62.967	61.867
GEM	63.840	62.493	62.940	61.947	61.173
hier. pooling	58.533	58.560	59.053	58.826	59.293
Trained embeddings					
InferSent	△	△	△	△	△
QuickThought	57.987	57.453	61.067	58.693	56.433
sent2vec	65.407	61.600	61.160	58.847	58.873
BERT	62.513	59.027	60.447	58.293	57.393
LASER	64.507	65.080	63.873	64.106	57.900

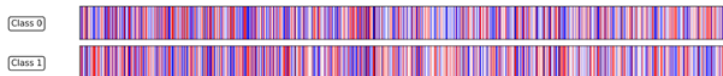
Section:
Visualizations



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Visualization of Classifier Weights

- ▶ Visualization of probing-task-classifier weights.
- ▶ Answers the question which dimensions capture the information.
- ▶ Example: Task *'voice'*, *'Quickthought'* embeddings and language *'de'*:

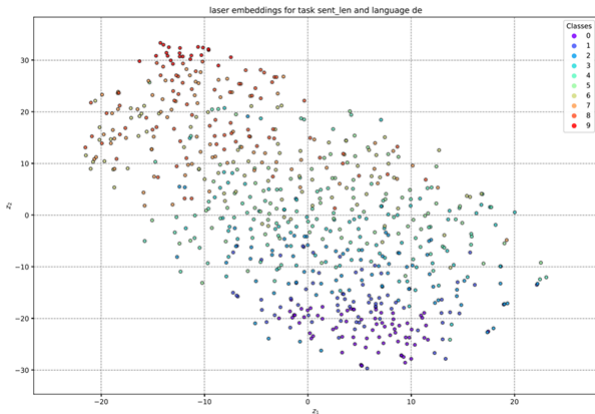


- ▶ Red bars indicate high weights, blue bars indicate low values, white / light bars represent weights around zero.
- ▶ \Rightarrow It might be useful to add a threshold.

Embedding Projections



TECHNISCHE
UNIVERSITÄT
DARMSTADT



Confusion Matrix

- Confusion matrix for task '*sentence length*', '*LASER*' embeddings and language '*de*':

