LSD GAČR project

Hidden in the Layers

Interpretation of Neural Networks for Natural Language Processing

David Mareček, Jindřich Libovický, Rudolf Rosa, Tomáš Musil, Tomasz Limisiewicz

November 30, 2020

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

LSD team

Linguistic **S**tructure representation in **D**eep networks

(GAČR 2018 – 2020)

David Mareček

Jindřich Libovický

Rudolf Rosa

Tomáš Musil

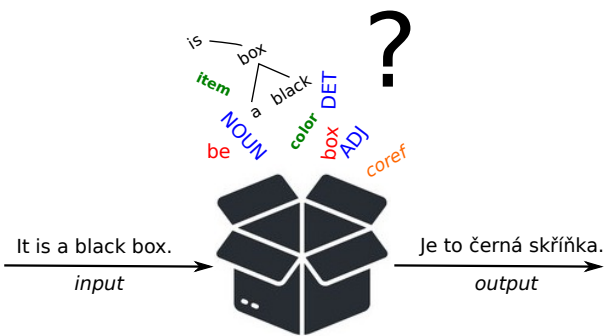
Tomasz Limisiewicz

Introduction and Motivation

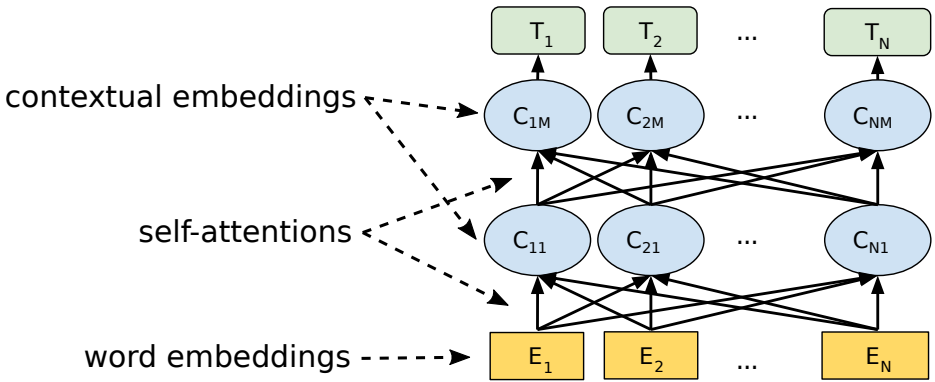
- Deep neural networks have rapidly become a central component in many NLP systems.
- They do not have any explicit knowledge of linguistic abstractions.
- End-to-end-trained models are black boxes that are very hard to interpret.

Introduction and Motivation

- How do they work? What emergent abstractions can we observe in them?
- How can we interpret their internal representations?
- Are the emergent structures and abstractions similar to classical linguistic structures and abstractions?



What We Analyze

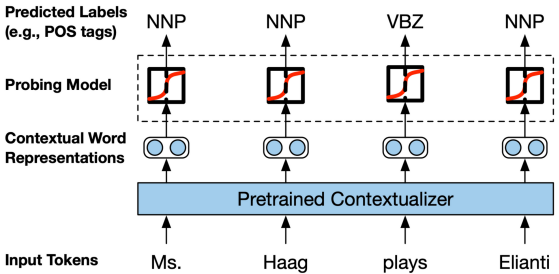


Outline

Methods

Supervised Methods: Probing

- Training supervised classifiers predicting linguistic features (e.g. POS tagger) on top of the internal representations.
- We assume that when probing classifier accuracy is high the networks encodes linguistic abstraction well.



Liu et al. (2019): "Linguistic Knowledge and Transferability of Contextual Representations"

Supervised vs. Unsupervised Methods

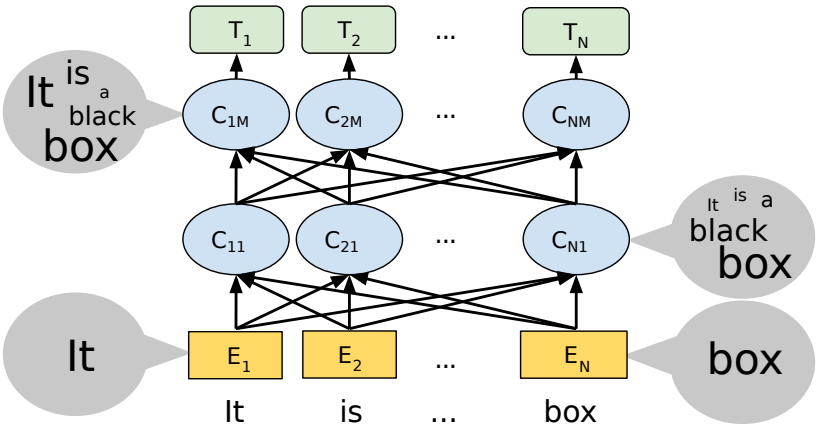
Supervised:

- Requires data annotated for the studied property.
- We can only reveal the kind of information that we have previously decided to look for.
- Retroactively affirms the correctness of the conceptualization and design decisions.

Unsupervised:

- Clustering, Component analysis, Structural induction from attentions
- We analyze the features that emerge in the representations, and only then we try to map them to existing linguistic abstractions.
- Complicated evaluation.

Words versus States



Words versus Subwords

The models typically operate on subword units.

Many linguistic formalisms are based on words (POS tags, dependency trees, ...)

- Words to Subwords – modify the linguistic abstraction to apply to subwords
- Subwords to Words – reconstruct word representations from the subword representations
- Fully Word-Based Approach – train the model on words

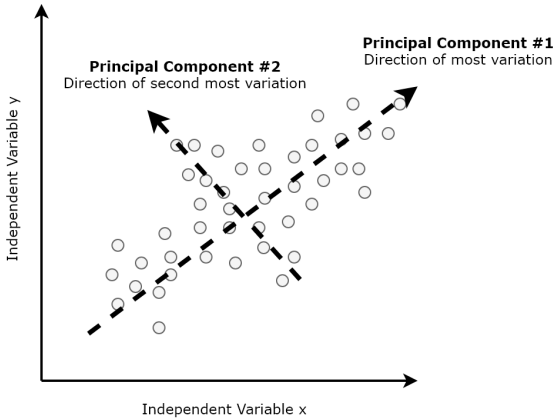
Word Embeddings

Word Embeddings

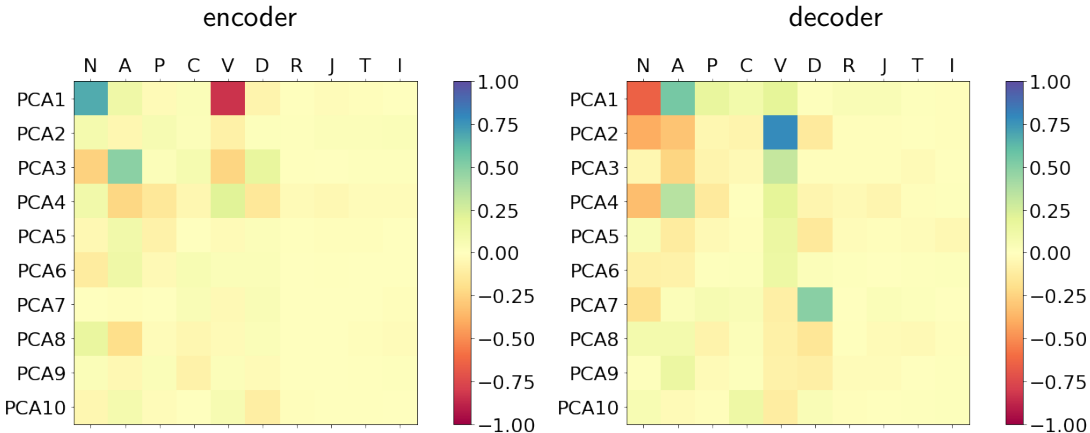
- A vector for each word (e.g. 100 dimensional, i.e. each word associated with a list of 100 real numbers)
- Learned in an unsupervised way from large plaintext corpora
- Observes the distributional hypothesis: words that appear in similar context have similar embeddings
- word2vec $>$ RNN $>$ attention $>$ Transformer

Principal Component Analysis (PCA)

- Transformation to another orthogonal basis set
- 1st principal component has the largest possible variance across the data
- Each other principal component is orthogonal to all preceding components and has the largest possible variance.
- If something correlates with the highest principal components its possibly very important for the NLP task.

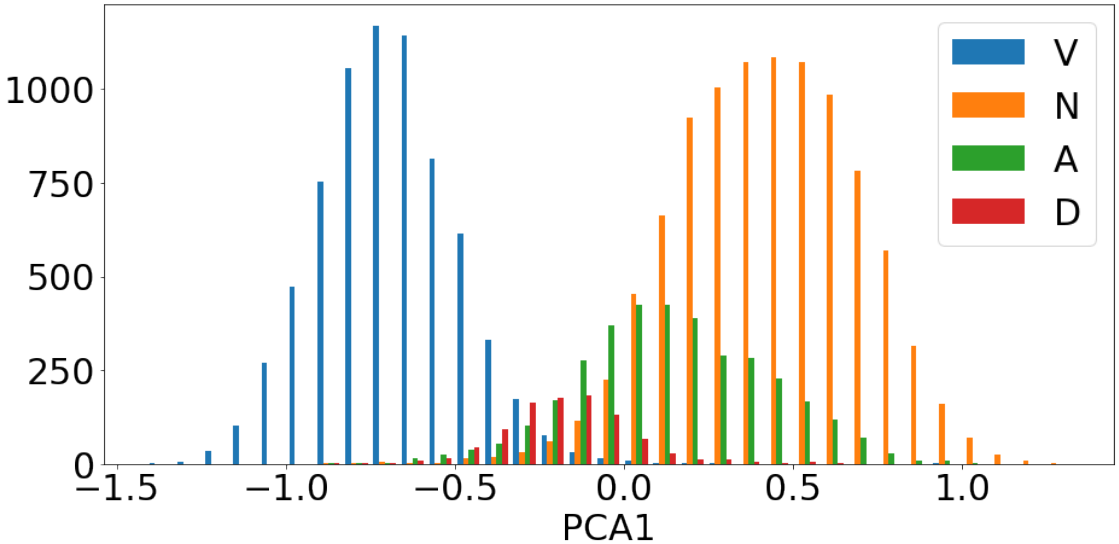


Word-embeddings learned by NMT, correlation with POS tags

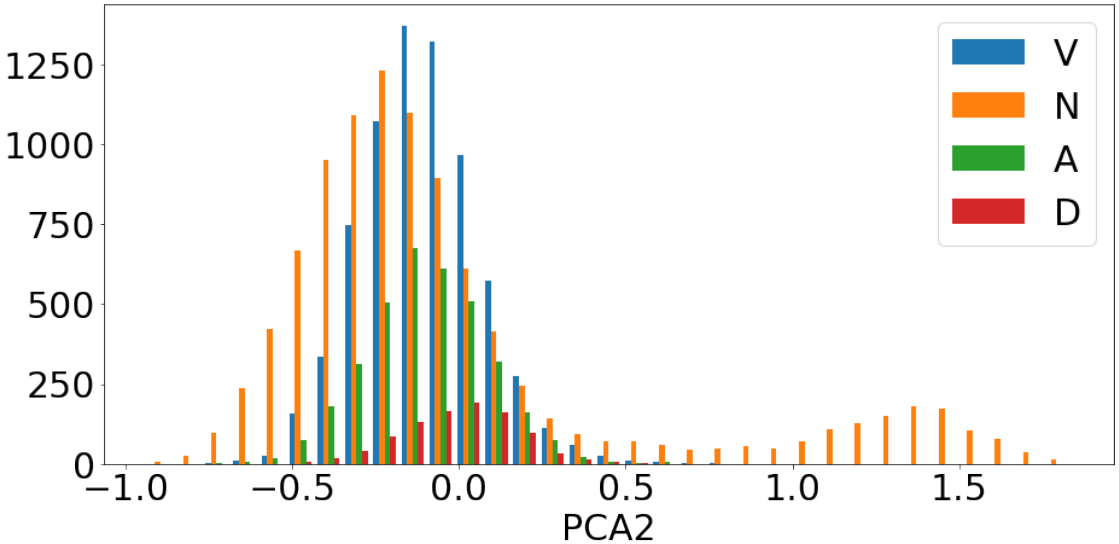


N = Nouns, A = Adjectives, P = Pronouns, C = Numerals, V = Verbs,
D = Adverbs, R = Prepositions, J = Conjunctions, T = Particles, I = Interjections

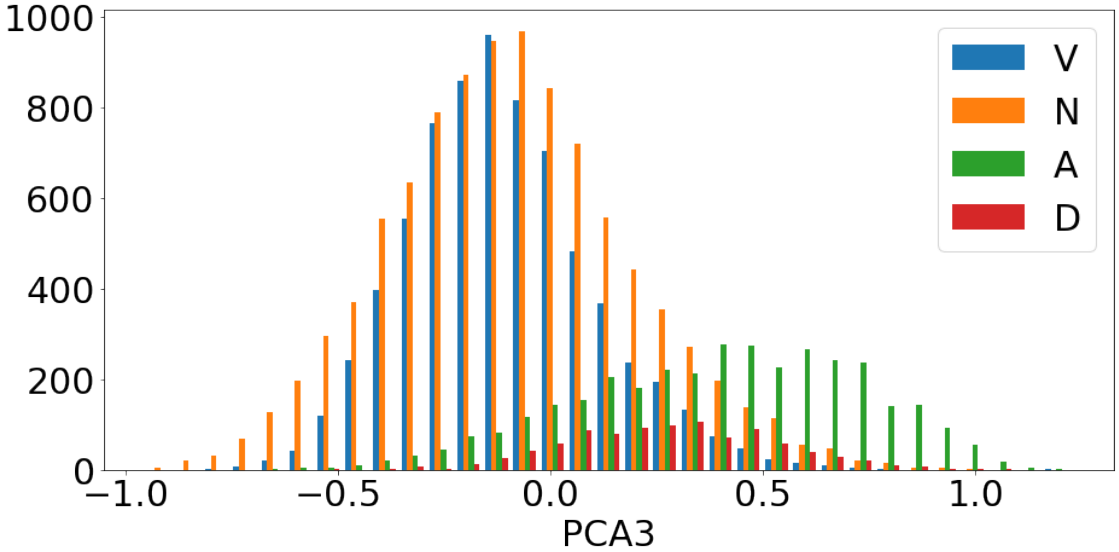
Word-embedding space learnt by NMT encoder



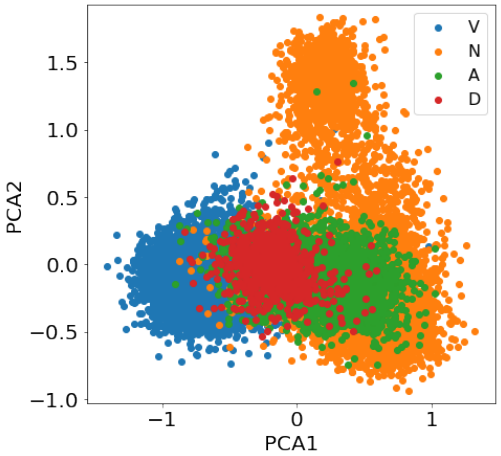
Word-embedding space learnt by NMT encoder

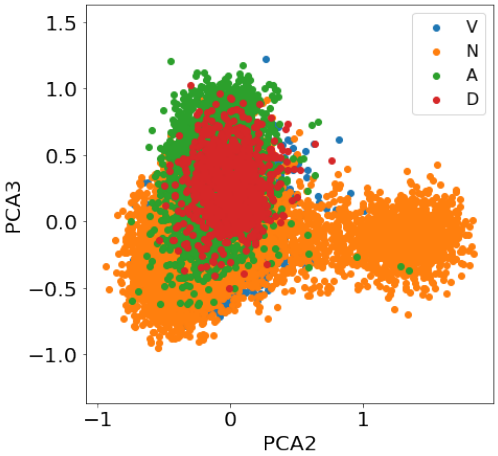


Word-embedding space learnt by NMT encoder

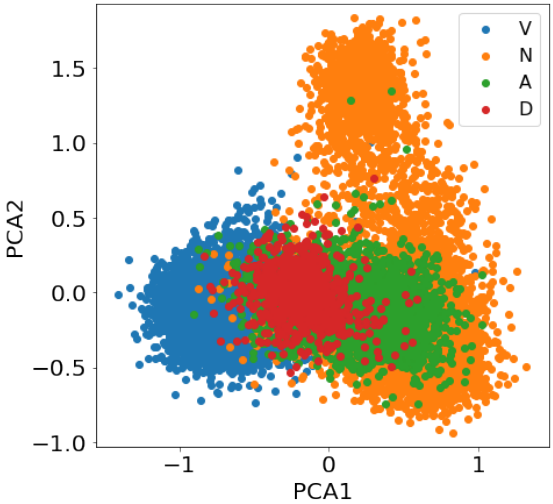


Word-embedding space learnt by NMT encoder





Word-embedding space learnt by NMT encoder



What is the separated island of Nouns visible in PCA2?

When we take a sample of words from this cluster, it contains almost exclusively named entities:

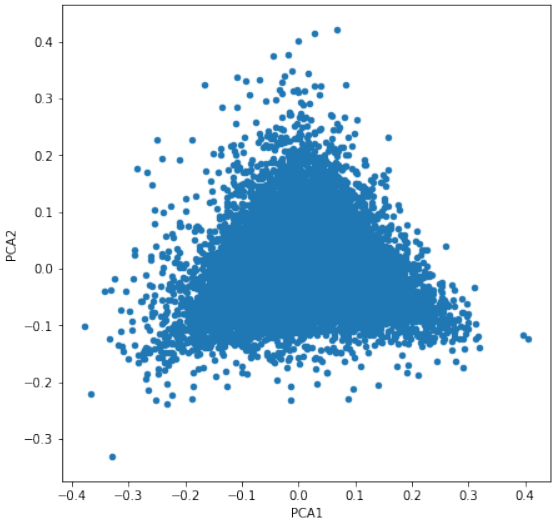
Fang, Eliáš, Jož, Aenea, Bush, Eddie, Zlatoluna, Gordon, Bellondová, Hermiona

Word-embedding space learnt by Sentiment Analysis

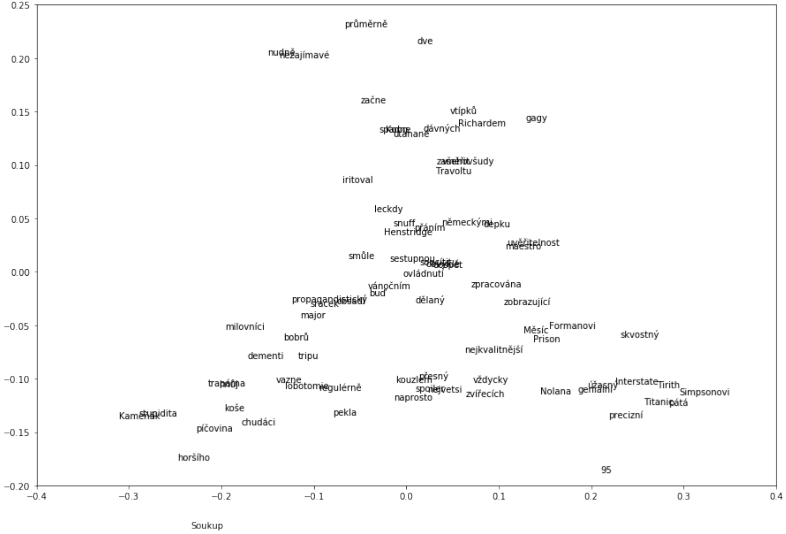
- Task: deciding whether a given text is emotionally positive, negative, or neutral.
- Trained on Czech ČSFD database (`https://www.csfd.cz/`), data were obtained from user comments and rankings of movies.
- Architecture: Convolutional neural network based on Kim 2014.

Neg: *"Very boring. I felt asleep."*

Pos: *"Great movie with super effects!!!"*

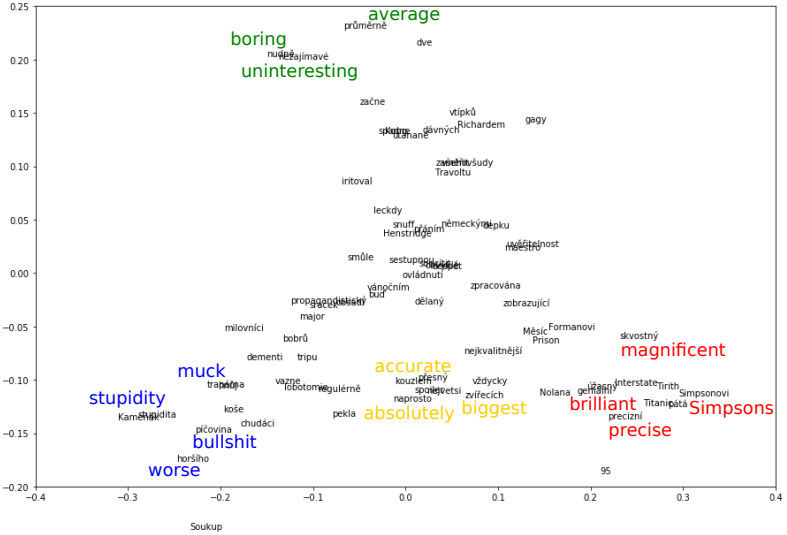


Word-embedding space learnt by Sentiment Analysis



We sampled some words from the vector space...

Word-embedding space learnt by Sentiment Analysis



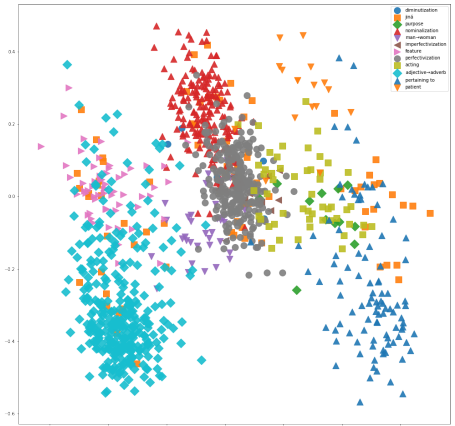
↔ . . . polarity of the word

↕ . . . intensity of the word

Word embedding space is shaped by the task for which it is trained.

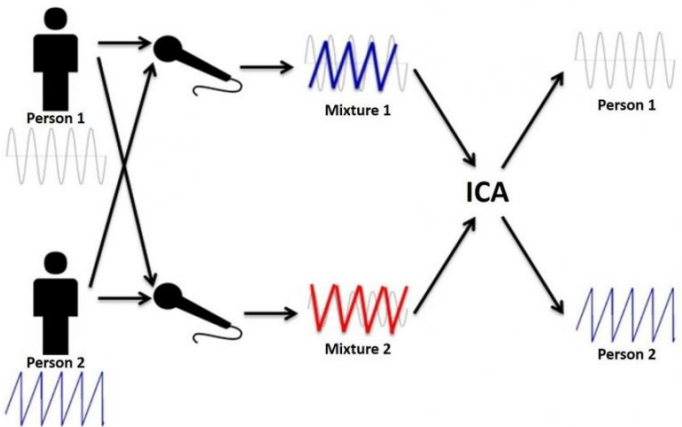
Looking for derivational relations

e.g. kompenzovat – kompenzace (compensate – compensation)



smutný – smutně letní – letně

Independent Component Analysis



- Iteratively find components that are as non-gaussian as possible

Independent Component Analysis

Semantic category: words with similar semantic content (e.g., law and justice) from various syntactic categories (in this case predominantly nouns in nominative and genitive morphological case):

zákona, Unie, členských, zákon, stanoví, Komise, zákony, soud, zákonů, zákonem, Evropské, práva, práv, ustanovení, nařízení, porušení, soudu, tj, souladu, podmínek

Glosses: *law*$_{\text{noun gen. sg.}}$, *union*$_{\text{noun nom. sg.}}$, *member*$_{\text{adj. gen. masc.}}$, *law*$_{\text{noun nom. sg.}}$, *determines*$_{\text{verb}}$, *comittee*$_{\text{noun nom. sg.}}$, *laws*$_{\text{noun nom. pl.}}$, *court*$_{\text{noun nom. sg.}}$, *laws*$_{\text{noun gen. pl.}}$, *law*$_{\text{noun inst. sg.}}$, *european*$_{\text{adj. gen. fem. sg.}}$, *rights*$_{\text{noun nom. pl.}}$, *rights*$_{\text{noun gen. pl.}}$, *provision*$_{\text{noun sg.}}$, *regulation*$_{\text{noun sg.}}$, *violation*$_{\text{noun sg.}}$, *court*$_{\text{noun gen. sg.}}$, *ie*$_{\text{shortcut}}$, *compliance*$_{\text{noun gen. sg.}}$, *conditions*$_{\text{noun gen. pl.}}$

Independent Component Analysis

Semantic and syntactic category: words that are defined both semantically and syntactically, in this case, predominantly verbs associated with *going somewhere* in the past tense masculine:

šel, zašel, zajít, jít, spěchal, šla, zavedl, vešel, dopravit, nešel, vrátil, poslal, vydal, šli, poslat, přišel, odjel, přijel, jel, dorazil

Glosses: *went*$_{\text{verb masc.}}$, *went down*$_{\text{verb masc.}}$, *go down*$_{\text{verb inf.}}$, *go*$_{\text{verb inf.}}$, *hurried*$_{\text{verb masc.}}$, *went*$_{\text{verb fem.}}$, *led*$_{\text{verb masc.}}$, *entered*$_{\text{verb masc.}}$, *transport*$_{\text{verb inf.}}$, *didn't go*$_{\text{verb masc.}}$, *returned*$_{\text{verb masc.}}$, *sent*$_{\text{verb masc.}}$, *issued*$_{\text{verb masc.}}$, *went*$_{\text{verb masc. pl.}}$, *send*$_{\text{verb inf.}}$, *came*$_{\text{verb masc.}}$, *left*$_{\text{verb masc.}}$, *came*$_{\text{verb masc.}}$, *went*$_{\text{verb masc.}}$, *arrived*$_{\text{verb masc.}}$

Independent Component Analysis

Syntactic subcategory: words with specific syntactic features, but semantically diverse (in this case, adjectives in feminine singular form):

Velká, moudrá, občanská, dlouhá, slabá, čestná, železná, překrásná, hladká, určitá, marná, tmavá, hrubá, příjemná, bezpečná, měkká, svatá, nutná, volná, zajímavá

Glosses: *big*$_{\text{adj. fem.}}$, *wise*$_{\text{adj. fem.}}$, *citizen*$_{\text{adj. fem.}}$, *long*$_{\text{adj. fem.}}$, *weak*$_{\text{adj. fem.}}$, *honest*$_{\text{adj. fem.}}$, *iron*$_{\text{adj. fem.}}$, *beautiful*$_{\text{adj. fem.}}$, *smooth*$_{\text{adj. fem.}}$, *certain*$_{\text{adj. fem.}}$, *in vain*$_{\text{adj. fem.}}$, *dark*$_{\text{adj. fem.}}$, *gross*$_{\text{adj. fem.}}$, *pleasant*$_{\text{adj. fem.}}$, *safe*$_{\text{adj. fem.}}$, *soft*$_{\text{adj. fem.}}$, *holy*$_{\text{adj. fem.}}$, *necessary*$_{\text{adj. fem.}}$, *free*$_{\text{adj. fem.}}$, *interesting*$_{\text{adj. fem.}}$

Independent Component Analysis

Feature across POS categories: e.g., feminine plural form for adjectives, pronouns and verbs:

tyto, tyhle, neměly, byly, mohly, začaly, vynořily, zmizely, měly, objevily, všechny, vypadaly, nebyly, zdály, změnily, staly, takové, podobné, jiné, tytéž

Glosses: *these*$_{\text{pron. fem. pl.}}$, *those*$_{\text{pron. fem. pl.}}$, *didn't have*$_{\text{verb fem. pl.}}$, *were*$_{\text{verb fem. pl.}}$, *could*$_{\text{verb fem. pl.}}$, *began*$_{\text{verb fem. pl.}}$, *emerged*$_{\text{verb fem. pl.}}$, *disappeared*$_{\text{verb fem. pl.}}$, *had*$_{\text{verb fem. pl.}}$, *discovered*$_{\text{verb fem. pl.}}$, *all*$_{\text{pron. fem. pl.}}$, *looked*$_{\text{verb fem. pl.}}$, *weren't*$_{\text{verb fem. pl.}}$, *seemed*$_{\text{verb fem. pl.}}$, *changed*$_{\text{verb fem. pl.}}$, *happened*$_{\text{verb fem. pl.}}$, *such*$_{\text{pron. fem. pl.}}$, *similar*$_{\text{adj. fem. pl.}}$, *other*$_{\text{adj. fem. pl.}}$, *same*$_{\text{pron. fem. pl.}}$

Independent Component Analysis

Stylistic: in this case, words that often appear in informal spoken language (often second person verbs and colloquial forms):

máš, bys, tý, nemáš, seš, ses, víš, Hele, kterej, sis, jseš, bejt, vo, svýho, celej, děláš, chceš, teda, každej, velkej

Glosses: *have*$_{\text{verb 2nd}}$, *would*$_{\text{verb 2nd}}$, *the*$_{\text{pron. fem. gen. coll.}}$, *don't have*$_{\text{verb 2nd}}$, *are*$_{\text{verb 2nd coll.}}$, *have*$_{\text{verb 2nd refl.}}$, *know*$_{\text{verb 2nd}}$, *Hey*$_{\text{intj. coll.}}$, *which*$_{\text{pron. masc. coll.}}$, *have*$_{\text{verb 2nd refl.}}$, *are*$_{\text{verb 2nd coll.}}$, *be*$_{\text{verb inf. coll.}}$, *about*$_{\text{prep. coll.}}$, *your*$_{\text{pron. masc. gen. coll.}}$, *whole*$_{\text{adj. masc. coll.}}$, *do*$_{\text{verb 2nd}}$, *want*$_{\text{verb 2nd}}$, *well*$_{\text{part. coll.}}$, *each*$_{\text{pron. masc. coll.}}$, *big*$_{\text{adj. masc. coll.}}$

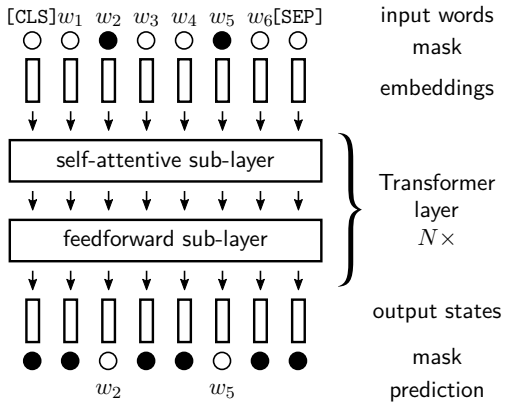
Contextual Embeddings

Contextual Embeddings

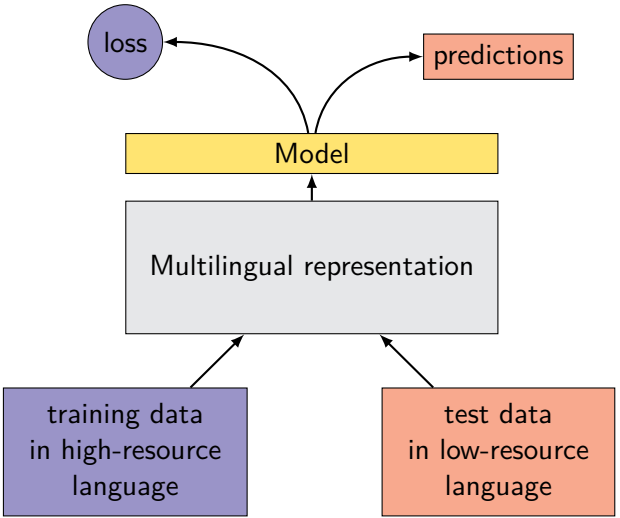
Multilingual Properties of the Multilingual BERT

Multilingual Pre-trained Representations

- Trained as standard BERT, but with 100 languages
- No information about language identity provided during training
- Surprisingly succesful in zero-short model transfer



Zero-Shot Evaluation



- Literature presents inconsistent results
- Methodological issue: How can we know the model does not overfit to the parent language
- If it works well, we cannot distinguish the role of the model and pre-trained representation

Our evaluation

Avoid zero-shot transfer, use the representation directly.

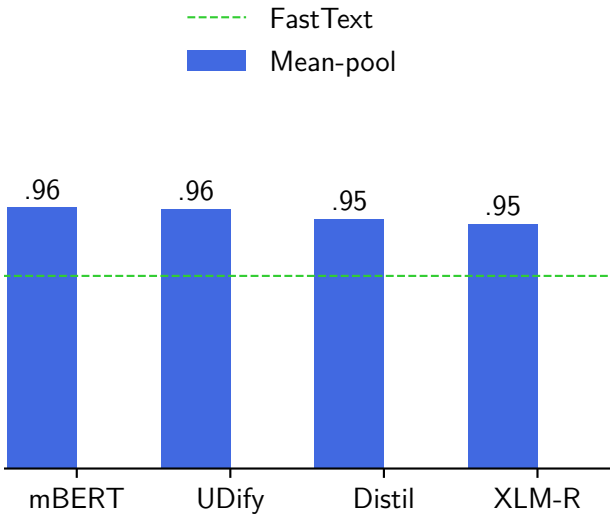
Probing tasks:

- Language identification
- Parallel sentence retrieval
- Word alignment
- MT quality estimation (skipped)

Probed models:

- Multilingual BERT
- DistilBERT
- XLM-RoBERTa
- Finetuned mBERT from UDify

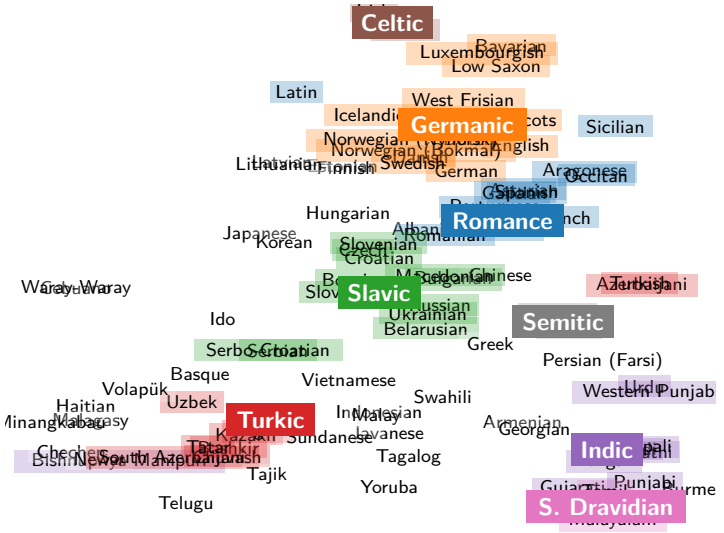
Language Identity



- Probing classifier trained 100 languages, 50k sent. / language
- Accuracy is higher than SoTA classifier (FastText)

How can the representation be language-neutral if language is so well-represented?

Language Clustering



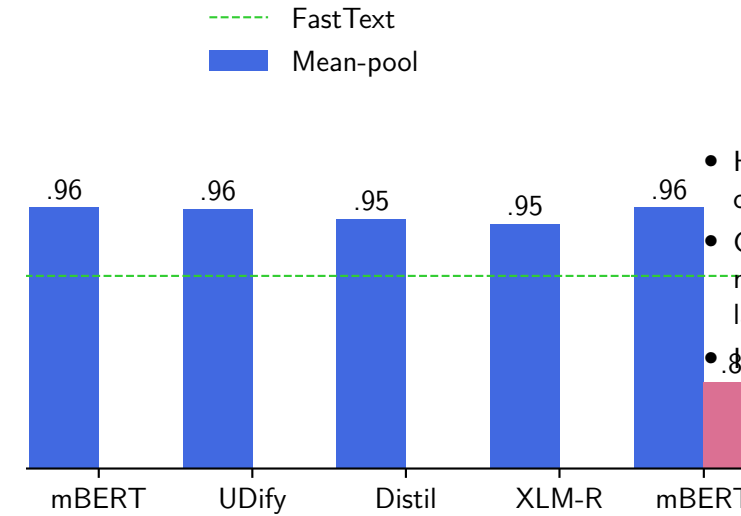
Average sentence vectors tend to cluster by language families.

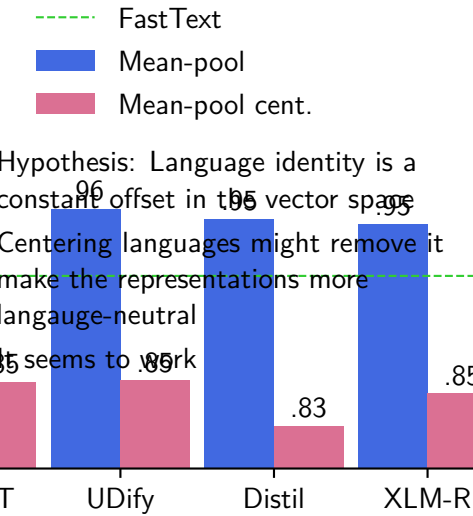
Hierarchical clustering vs. families from WALS:

	H	C	V
mBERT	82.0	82.9	82.4
UDify	80.5	79.7	80.0
XLM-R	69.7	69.1	69.3
Distil	81.6	81.1	81.3
random	60.2	64.3	62.1

H homogenity
C comleteness
V V-measure (harm. avg. of H and C)

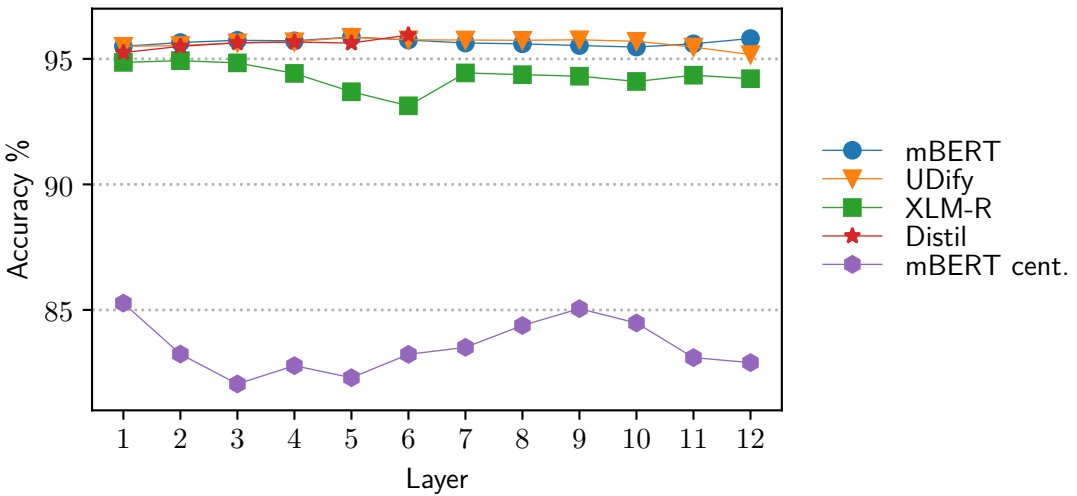
Language Identity (2)





- Hypothesis: Language identity is a constant offset in the vector space
- Centering languages might remove it make the representations more langauge-neutral
- It seems to work

Language Identity in Layers

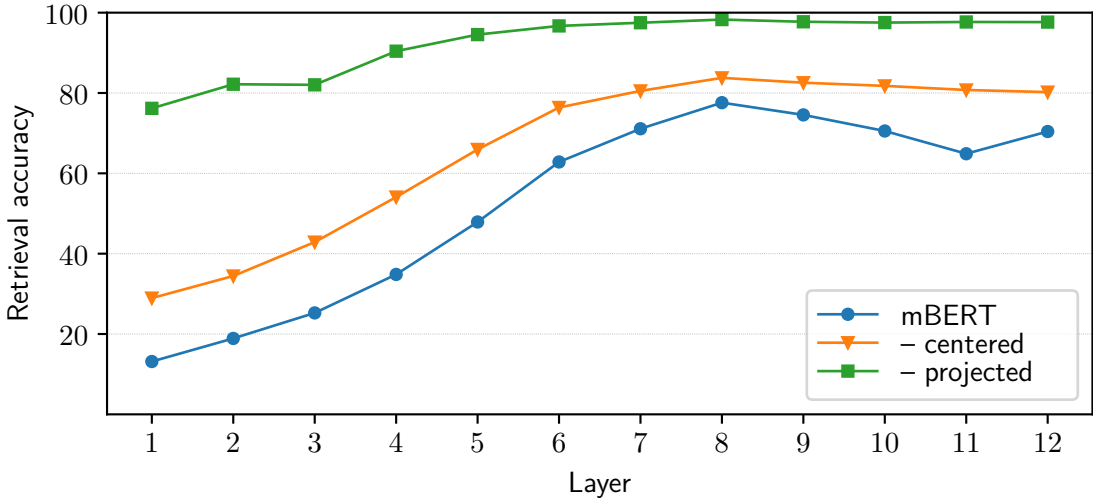


Sentence Retrieval (1)

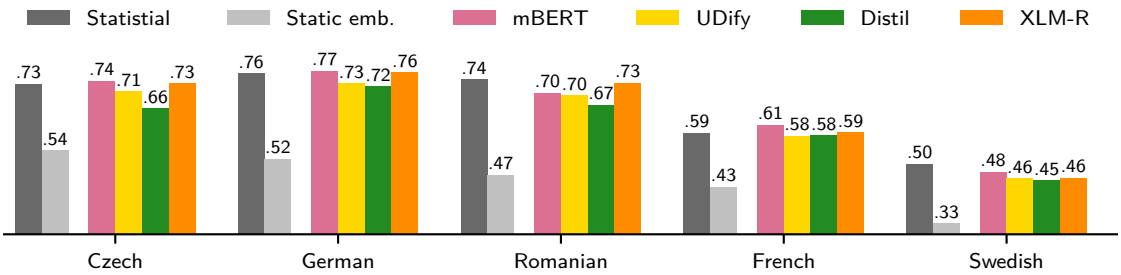
Accuracy of parallel sentence retrieval on WMT14 data.
Based on cosine similarity of the sentence vectors.

	Static	mBERT	UDify	Distil	XLM-R
mean	.113	.776	.314	.600	.883
mean, cent.	.496	.838	.564	.770	.923
mean, proj.	.650	.983	.906	.980	.996

Sentence Retrieval (2)



Word Alignment



- Minimum weighted edge cover in a bipartite graph
- Matches statistical aligners trained on 1M parallel sentences
- Indirectly confirm the constant shift hypothesis

Language Neutrality: Summary

1. Pre-trained multilignual represetnation are not much language neutral langauge ID is useful during pre-training
2. Language-specific representation centering is an unsupervised way of improving lanauge neutrality
3. Training explicit projection is better, but requires parallel data

A nice side-effect: SoTA results on language ID and word alignment

Remaing questions about language neutrality

- Can we match projection trained on parallel data in an unsupervised way?
- How make the representation langauage-neutral by default without post-processing?

Contextual Embeddings

Memorization in Probing

Risk of Memorization in Probing

- A typical observation
  - Probing classifier can predict a label (e.g. POS) from contextual embeddings
- Possible explanations
  - \+ The *probed model captures* POS in contextual embeddings
  - − The *probing classifier memorizes* POS for individual words
- Expected solution
  - Use disjoint sets of training words and testing words
- Problem
  - Representations are *contextual*, POS tags are determined *by context*
  - Need to use full sentences, and sentences overlap in words

Possible Solutions

- Rosa, Musil, and Mareček (2020)
  - Split train sentences into *seen* and *unseen*
  - Train on contextual embeddings of words in seen sentences
  - Test on contextual embeddings of unseen words in test sentences
  - Compare to performance on seen words in test sentences
- Bisazza and Tump (2018)
  - Split vocabulary into train words and test words
  - Train only on contextual embeddings of train words
  - Test only on contextual embeddings of test words
- Hewitt and Liang (2019)
  - Compare to randomly assigned labels (probing classifier must memorize)

Memorization Findings

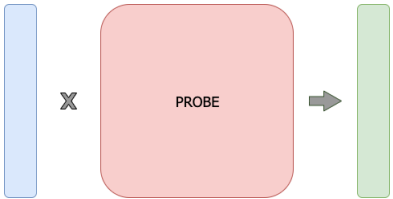
- Memorization of word identities by the probing classifier **does** occur
- Effect is stronger with: static embeddings, small training data, stronger classifier
- Case study on predicting POS from NMT encoder representations
  - Word embeddings, MLP trained on 50 sentences
    - Accuracy 98.5% on seen words versus 74.3% on unseen words
  - Output states, MLP trained on 1,000 sentences
    - Accuracy 96.8% on seen words versus 94.9% on unseen words
  - Output states, linear classifier trained on 10,000 sentences
    - Accuracy 95.7% on seen words versus 95.5% on unseen words

Contextual Embeddings

Separating Lexical and Syntactic Features

Syntactic Structure Probing (Hewitt's approach)

Hewitt and Manning (2019) took the BERT contextual vectors and trained a projection matrix to obtain another vectors whose differences would approximate distances between tokens in dependency trees.

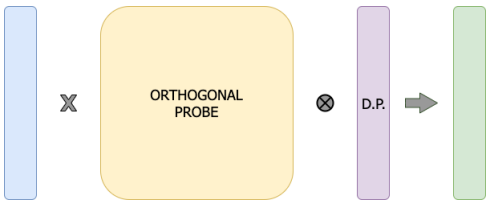


Dependency trees obtained by Minimum spanning tree gained 82.5% UAS on English PTB.

Orthogonal Probing

We decompose the trained projection matrix into two matrices:

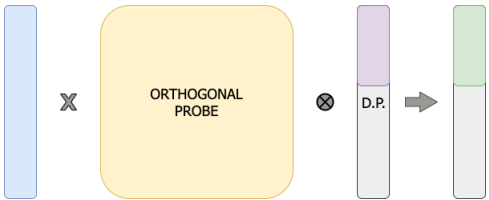
- orthogonal matrix: only rotates the vector space
- diagonal matrix: assigns weights to individual dimensions - how important they are for the probing task



Orthogonal Probing

Observation:

- Many weights trained in the diagonal matrix are close to zero.
- This method shows us which dimensions of the rotated space are useful for the probing task, e.g., syntax



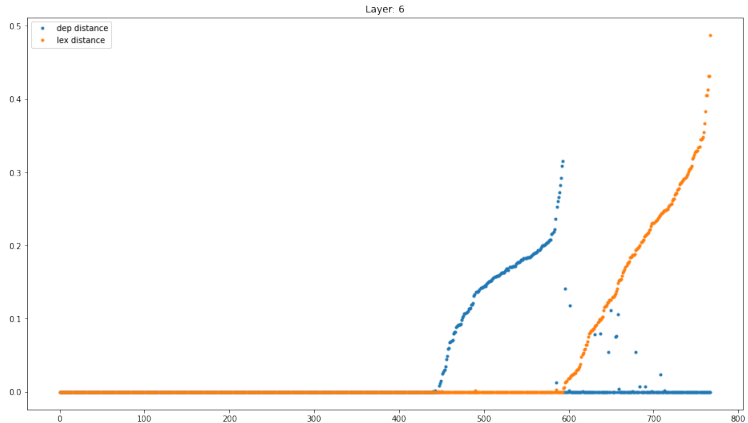
Idea:

- Probing for more tasks at once with shared orthogonal matrix.
- Could we separate the dimensions needed for specific tasks?

Separating Lexical and Syntactic Features

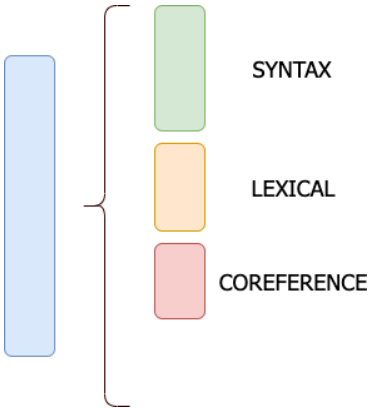
We are probing for distances between two words

- in dependency tree (syntactic features)
- in the WordNet hyperonymic tree structure (lexical features)



Ortogonal Probing - Conclusions

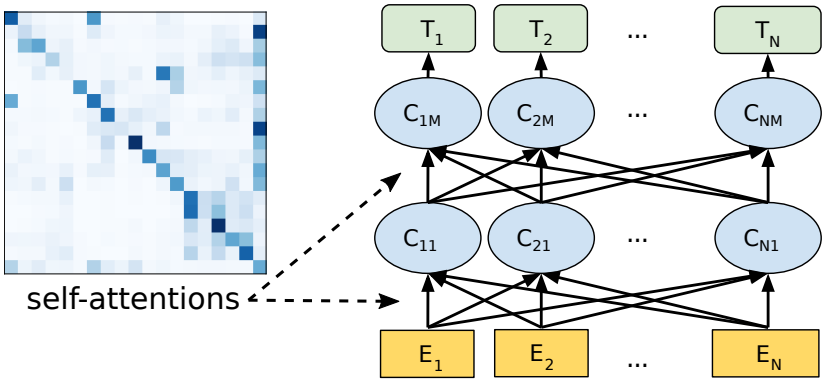
- We are able to identify syntactic and lexical dimensions of the BERT representations.
- We only need to rotate the vector space to transform the features hidden in linear combinations of dimensions into single dimensions.
- What is hidden in the rest of the dimensions?



Self-Attentions and Syntax

Self-Attentions in Transformer

Self-attentions: Weighted connections between word representations showing how much a word representation in one layer contributes to another word representation in the following layer.



Self-Attentions in Transformer – The Goals

We analyze self-attention weight matrices in:

- Neural Machine Translation Transformer Encoder (16 heads x 6 layers)
- BERTbase pre-trained model (12 heads x 12 layers)

To what extent attentions between individual (sub)word representations correspond to syntactic features?

Is it possible to extract syntactic structures from them?

Self-Attentions and Syntax

Transformer NMT Encoder and Phrase Trees

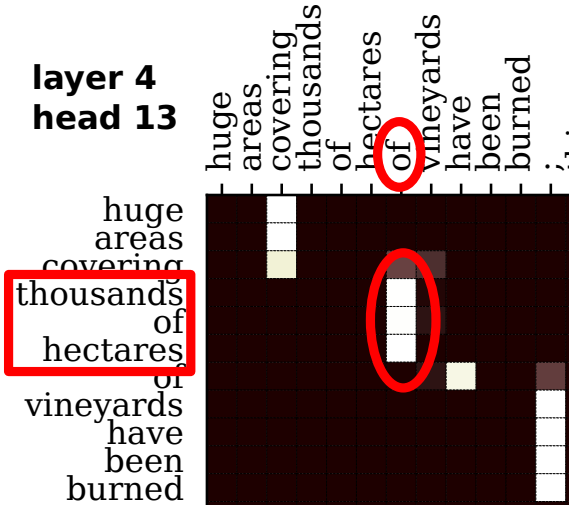
Self-Attentions in NMT Encoder and Phrase Trees

Observation: Common pattern in cca 70% of self-attention heads: "balustrades"

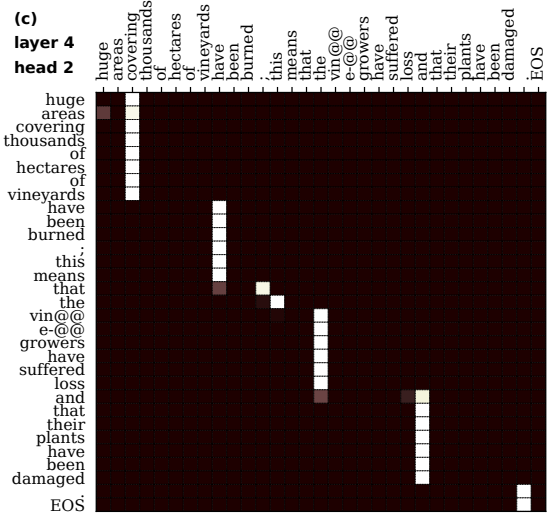
- Baluster: continuous sequence of words attending to the same position
- Looks like a syntactic phrase
- Usually attends to phrase boundary

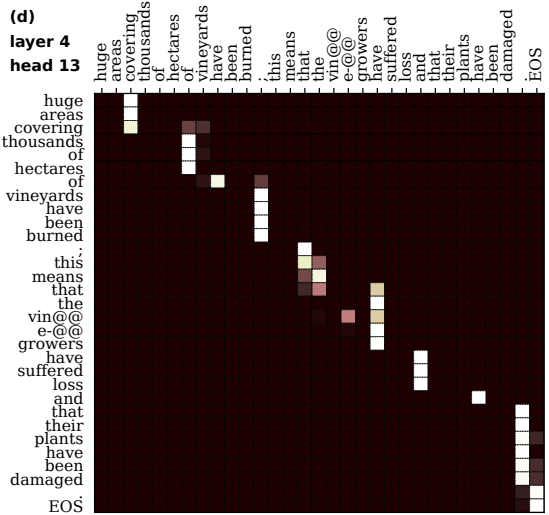
Research questions:

- Is that syntactic? To what extent?
- Could we extract phrase trees from attentions?
- How they differ from manually annotated phrase-trees?



Examples of Heads with Balustrades





Approach and Results

1. Transformer NMT: French $\leftrightarrow$ English, German $\leftrightarrow$ English, French $\leftrightarrow$ German
2. Phrase Scores:
   - based on the attention weights of "balusters"
   - collected and averaged over all heads and layers in the Encoder
3. Binary constituency trees:
   - linguistically uninformed algorithm
   - tree score = sum of phrase scores
   - CKY: find tree with maximal score

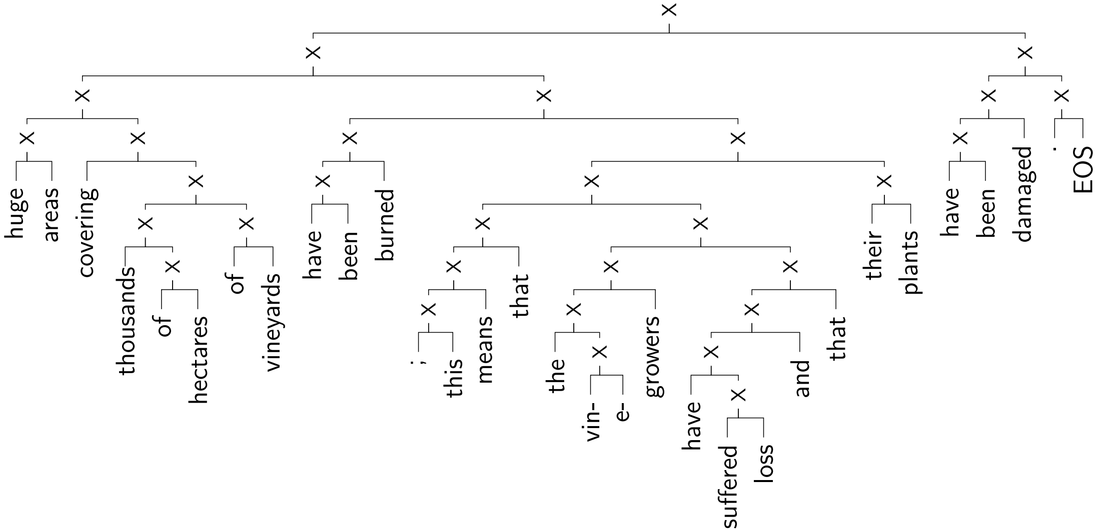
Comparison to standard constituency syntactic trees:

- we observe a 40% match, baseline has a 30% match (right-aligned balanced binary tree)

Analysis:

- The emergent structures can be seen as syntactic to some extent
- Shorter phrases are often captured
- Sentence clauses are often captured

Example of Tree



Self-Attentions and Syntax

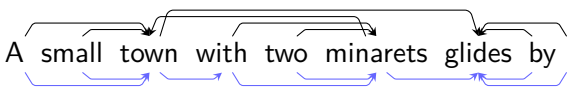
BERT Model and Dependency Relations

BERT Model and Dependency Relations

Many previous work showed that individual BERT attention heads tend to encode particular dependency relations.

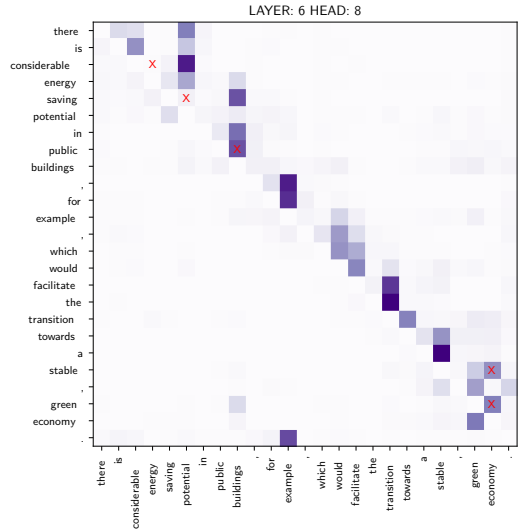
Our contributions:

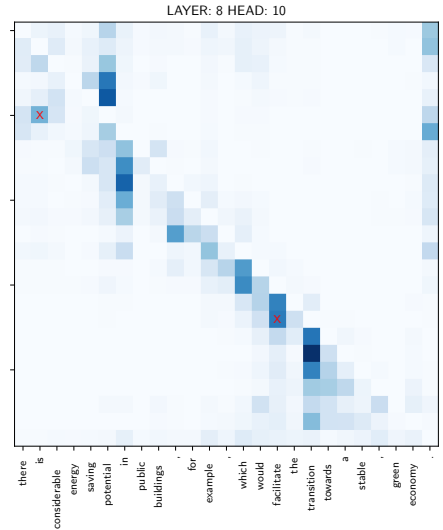
- Some heads are more abstract (include more dependency relations)
- Some heads are more specific (separate one relation type into more subtypes)
- We show a method how to extract labeled dependency trees (52% UAS, 22% LAS on English UD).



BERT model and Dependency Relations

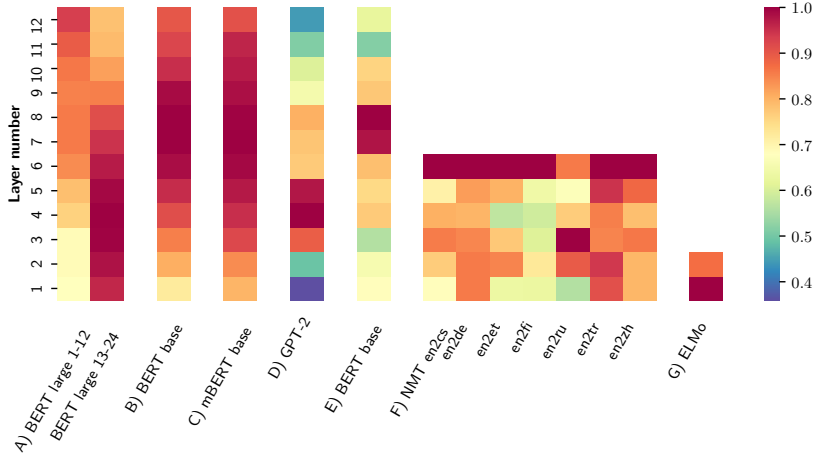
Self-attention in a particular heads of a language model aligns with dependency relations





Syntactic Accuracy Across Layers

Relative syntactic information across attention models and layers



Summary

1. Advantages of unsupervised methods.
2. Word embeddings capture morphological features.
3. In contextual embeddings, the lexical and syntactic information can be separated.
4. Language information in mBERT.
5. Constituency phrases captured in NMT.
6. Dependency relations captured by individual self-attention heads.
7. Today, we are finishing the book “Hidden in the Layers”.

References

Arianna Bisazza and Clara Tump. "The Lazy Encoder: A Fine-Grained Analysis of the Role of Morphology in Neural Machine Translation". In: *Proc. EMNLP*. Brussels, Belgium: ACL, 2018, pp. 2871–2876. URL: https://www.aclweb.org/anthology/D18-1313.

John Hewitt and Percy Liang. "Designing and Interpreting Probes with Control Tasks". In: *Proc. EMNLP-IJCNLP*. Hong Kong, China: ACL, 2019, pp. 2733–2743. URL: https://www.aclweb.org/anthology/D19-1275.

Yoon Kim. "Convolutional Neural Networks for Sentence Classification". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014, pp. 1746–1751. DOI: 10.3115/v1/d14-1181. URL: http://dx.doi.org/10.3115/v1/d14-1181.

Nelson F. Liu et al. "Linguistic Knowledge and Transferability of Contextual Representations". In: *NAACL-HLT*. 2019.

Rudolf Rosa, Tomáš Musil, and David Mareček. "Measuring Memorization Effect in Word-Level Neural Networks Probing". In: *International Conference on Text, Speech, and Dialogue*. Springer. 2020, pp. 180–188.