ETH zürich | DISTributed COmputing

Natural Language Multitasking

Analyzing and Improving Syntactic Saliency of Hidden Representations

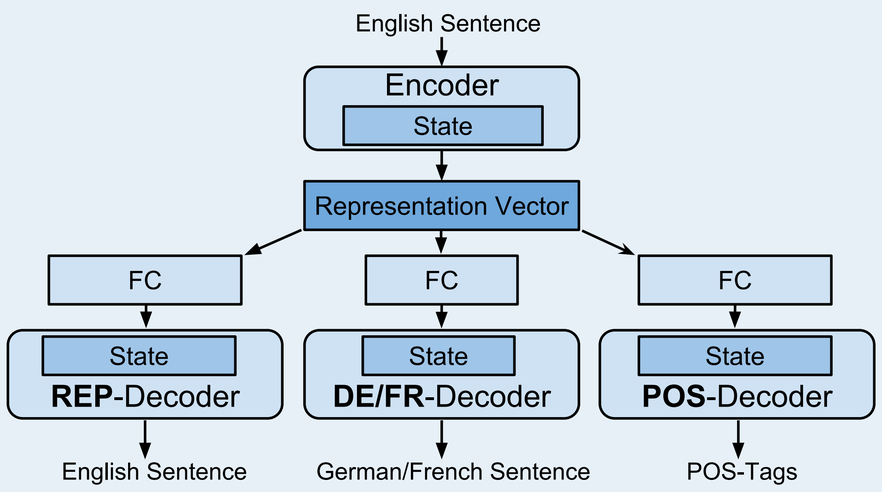
Gino Brunner, Yuyi Wang, Roger Wattenhofer, Michael Weigelt
Distributed Computing Group, ETH Zurich

Introduction

Humans are very good at disentangling different aspects of a natural language, such as syntax and semantics. While most languages were not built around a set of rules, we managed to distil such rules into books that we can then use to learn languages more efficiently. Humans can describe the syntax of a sentence regardless of its meaning, and vice versa. Having such clear representations helps us to understand natural languages and enables efficient communication. Machine learning tasks related to natural language processing would likely also benefit from having "human-like" language representations. **In this work we investigate how hidden representations of generative language models can be analyzed and manipulated to improve the performance of NLP tasks.**

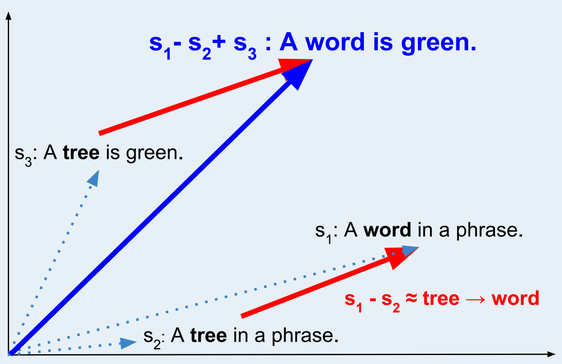
Method Overview

- Multi-task autoencoder with character-based LSTMs. We attach different tasks (decoders) to the representation layer:
  - REP: Reconstruct input
  - DE/FR: Translate to German/French
  - POS: Perform part-of-speech-tagging
- Dataset: 1.7M sentences aligned across English, German, French
- We use different numbers of encoders and representation vector sizes.



A high-level overview of our models. The state of the encoder is essentially copied into the state of the decoders. We use intermediate fully connected (FC) layers to allow for arbitrary model sizes.

Sentence Arithmetic



We perform vector operations on sentence representations in latent space. The difference between s_1 and s_2 is basically the change from "tree" to "word". When adding this difference to s_3, "tree" is replaced with "word" and the rest of the sentence is left intact.

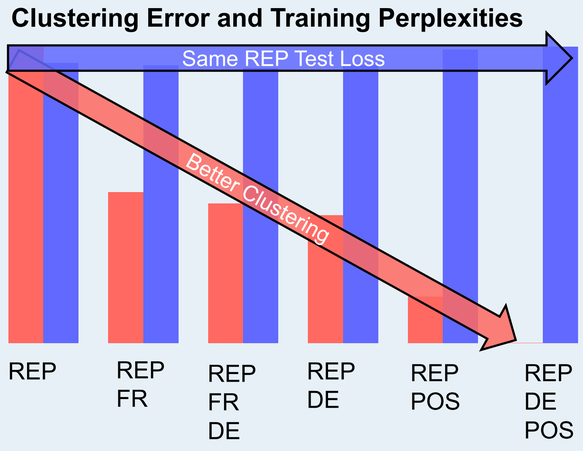
						REP	REP-DE-POS
s_1	-	s_2	+	s_3	=	s_1-s_2+s_3	s_1-s_2+s_3
I am one.	-	I am two.	+	You are two.	=	You exame on.	You are one.
This example works.	-	This example fails.	+	Another attempt fails.	=	Another attempt: someo han	Another attempts work.
A word in a phrase.	-	A tree in a phrase.	+	A tree is green.	=	A word is green.	A word is green.
The end is easier.	-	The start is easier.	+	A start is next!	=	A year nid and!	A negation is need!
A large number of people want to work.	-	A small number of people want to work.	+	A small sentence is enough.	=	A larges is economic any out.	A large sector for challenge.

Syntax Cluster Separation

- 14 sentence prototypes with differing syntax
- N (noun), V (verb), A (adjective), D (adverb)
- Prototypes randomly populated with English words → Syntax is different, but no semantic meaning.
- We cluster hidden representations of each sentence
- The models have never seen these sentences during training.

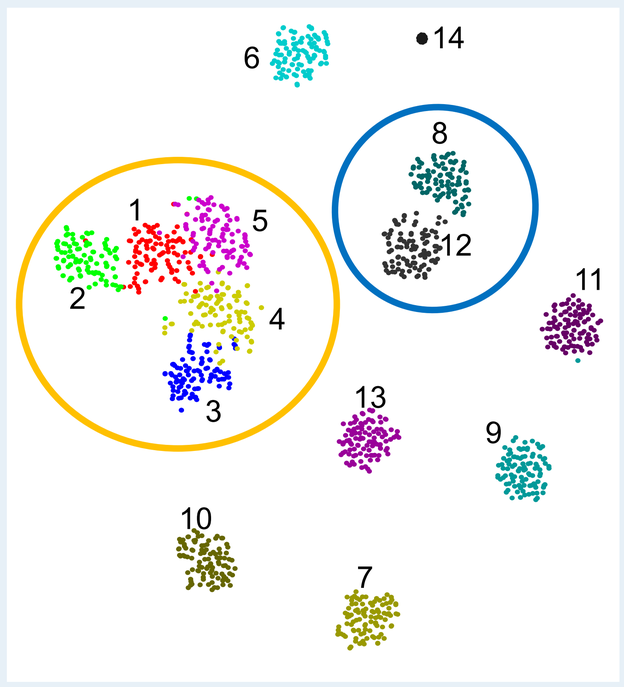
1: The +**N** is +**A** **2:** The +**N** +**V**s **3:** The +**N** has a +**N** **4:** The +**N** +**V**s a +**N**
5: The +**N** +**V**s +**D** **8:** +**A** +**N**s often +**V** like +**N**s. **12:** +**N** +**V**s in order to +**V** on a +**N** …

Clustering Error and Training Perplexities

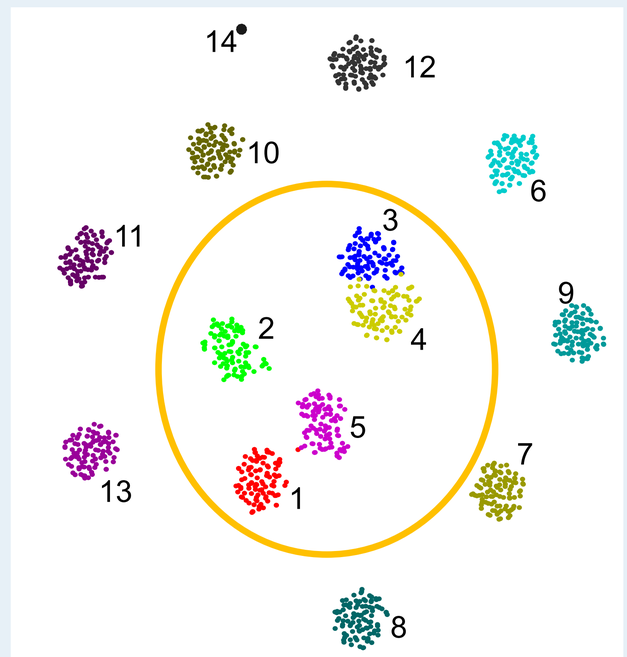


- Clustering error: Number of sentences assigned to a wrong cluster
- Test loss for all models the same, but clustering errors very different!
- Adding more linguistic tasks generally leads to better separated syntax clusters
- Translation helps, but part-of-speech tagging helps more
- French seems less helpful than German
- POS is closely related to syntax. Translation might help more for "semantic clustering"
- Clustering error was computed using k-means

REP



REP-DE-POS



- t-SNE to visualize clusters
- Comparison between a model with a single REP decoder and one with a REP, DE and POS decoder.
- The REP-DE-POS model's sentence representation clusters are more clearly separated, indicating it has a better "understanding" of syntax

Sentence Interpolation

Another way of exploring the latent space is by interpolating between sentence representations. We pick two sentences (**S**tart and **E**nd) from the test set and linearly interpolate between them. The resulting intermediate sentences can give us a clue about what the models have learned. Clearly, the model on the right (REP-DE-POS) produces more plausible sentences and fewer non-words or gibberish. This is consistent with the results from the sentence arithmetic and syntax clustering experiments. More work is needed to investigate other ways of interpolating and come up with more quantitative measures of quality.

REP

S: I think that is unfair.
I think hit item wen it.
I citible a staly by Tis.
We it hweman at you nowars.
We will ask Eur to by stalf.
We will avoid u networ for that.
E: We will make a note of this.

S: We are again faced with a long-running scandal.
We are agri-far you wait lack one sationslluz:
We argue as old issue - will renewingly, sirh leavin.
The macret and house if ACP tell-ownaires stune.
The just are lomping way I sell long-situates'.
The surl laved metho with I aloperouclous Itycs day.
E: They have dealt well with a long-running scandal.

REP-DE-POS

S: I think that is unfair.
I think that is unfair.
I think what is an often.
I who came wit anoth on I.
We will amoun this under it.
We will make a note of this.
E: We will make a note of this.

S: We are again faced with a long-running scandal.
We are again edited in a vivious' nurorial scan.
We are having failed way with round-and scandal.
They are weapen deal with a long-running scandal.
They have added dettine with a long-running scandal.
They have danged well with a long-running scandal.
E: They have dealt well with a long-running scandal.

Conclusion

We used a simple sequence to sequence autoencoder model based on vanilla LSTMs. We first trained the model to reproduce English sentences. We were interested in understanding the hidden representations learned by the models. We found that adding linguistic "auxiliary" tasks in the form of additional decoders changes the learned representations significantly. In order to analyze the changes, we fed instances of syntactic sentence prototypes to each model and extracted the resulting hidden sentence representations. The models using additional linguistic auxiliary tasks produce sentence representations that are more separated in the latent space. In the case of the best model (REP-DE-POS) and using k-means with $k = 14$, the clusters perfectly correspond to the 14 sentence prototypes. The simpler models perform significantly worse in this task. This experiment also shows that, not too surprisingly, the perplexity loss does not reflect the improvement in the syntax clustering. This highlights the importance of finding new ways of assessing the performance of generative models, beyond manual inspection by humans. We performed further experiments in the latent space such as adding difference vectors of two representations to a third one, and interpolating between representations. In the future we plan on formalizing our findings, including semantics and combining our models with state-of-the-art downstream tasks such as neural machine translation. We also want to find and disentangle other features of natural languages, such as artistic style and content.

ETH zürich | DISTributed COmputing

Natural Language Multitasking

Analyzing and Improving Syntactic Saliency of Hidden Representations

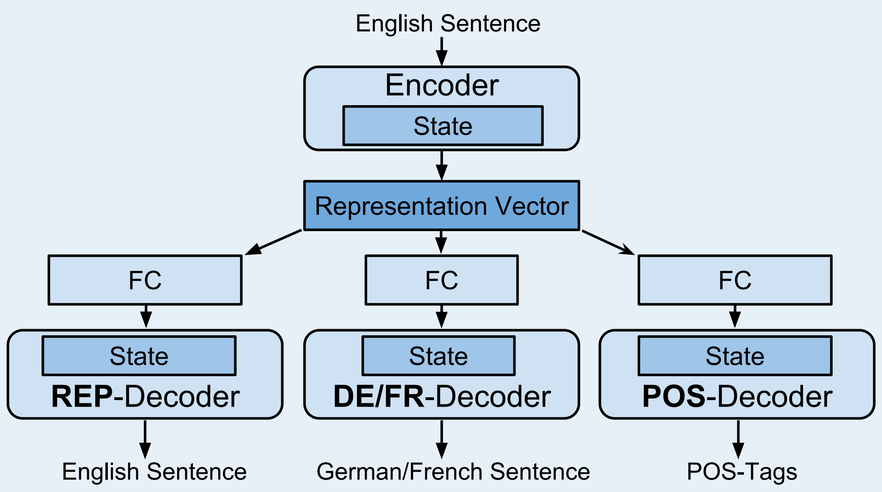
Gino Brunner, Yuyi Wang, Roger Wattenhofer, Michael Weigelt
Distributed Computing Group, ETH Zurich

Introduction

Humans are very good at disentangling different aspects of a natural language, such as syntax and semantics. While most languages were not built around a set of rules, we managed to distil such rules into books that we can then use to learn languages more efficiently. Humans can describe the syntax of a sentence regardless of its meaning, and vice versa. Having such clear representations helps us to understand natural languages and enables efficient communication. Machine learning tasks related to natural language processing would likely also benefit from having "human-like" language representations. **In this work we investigate how hidden representations of generative language models can be analyzed and manipulated to improve the performance of NLP tasks.**

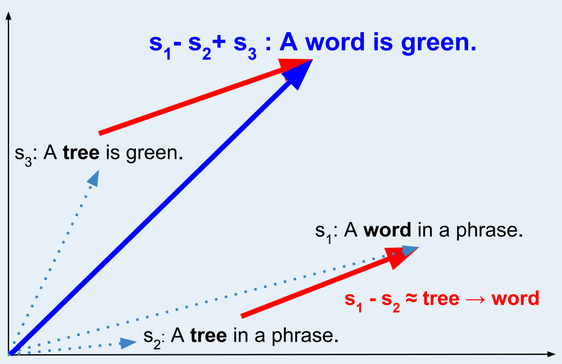
Method Overview

- Multi-task autoencoder with character-based LSTMs. We attach different tasks (decoders) to the representation layer:
  - REP: Reconstruct input
  - DE/FR: Translate to German/French
  - POS: Perform part-of-speech-tagging
- Dataset: 1.7M sentences aligned across English, German, French
- We use different numbers of encoders and representation vector sizes.



A high-level overview of our models. The state of the encoder is essentially copied into the state of the decoders. We use intermediate fully connected (FC) layers to allow for arbitrary model sizes.

Sentence Arithmetic



We perform vector operations on sentence representations in latent space. The difference between s_1 and s_2 is basically the change from "tree" to "word". When adding this difference to s_3, "tree" is replaced with "word" and the rest of the sentence is left intact.

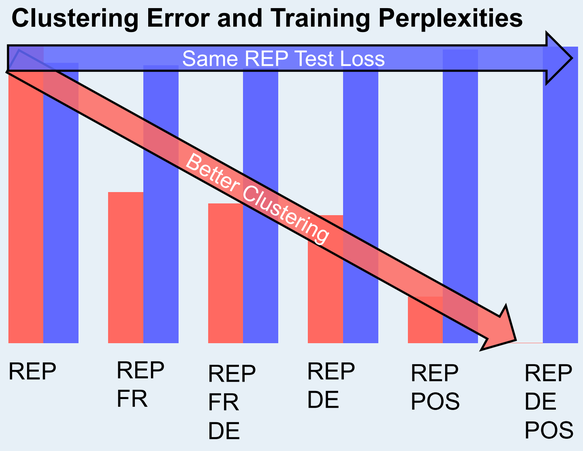
						REP	REP-DE-POS
s_1	-	s_2	+	s_3	=	s_1-s_2+s_3	s_1-s_2+s_3
I am one.	-	I am two.	+	You are two.	=	You exame on.	You are one.
This example works.	-	This example fails.	+	Another attempt fails.	=	Another attempt: someo han	Another attempts work.
A word in a phrase.	-	A tree in a phrase.	+	A tree is green.	=	A word is green.	A word is green.
The end is easier.	-	The start is easier.	+	A start is next!	=	A year nid and!	A negation is need!
A large number of people want to work.	-	A small number of people want to work.	+	A small sentence is enough.	=	A larges is economic any out.	A large sector for challenge.

Syntax Cluster Separation

- 14 sentence prototypes with differing syntax
- N (noun), V (verb), A (adjective), D (adverb)
- Prototypes randomly populated with English words → Syntax is different, but no semantic meaning.
- We cluster hidden representations of each sentence
- The models have never seen these sentences during training.

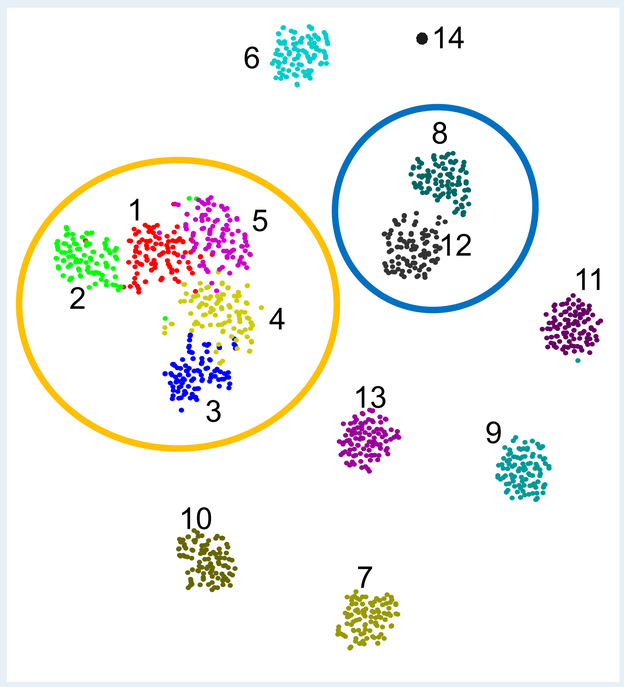
1: The +**N** is +**A** **2:** The +**N** +**V**s **3:** The +**N** has a +**N** **4:** The +**N** +**V**s a +**N**
5: The +**N** +**V**s +**D** **8:** +**A** +**N**s often +**V** like +**N**s. **12:** +**N** +**V**s in order to +**V** on a +**N** …

Clustering Error and Training Perplexities

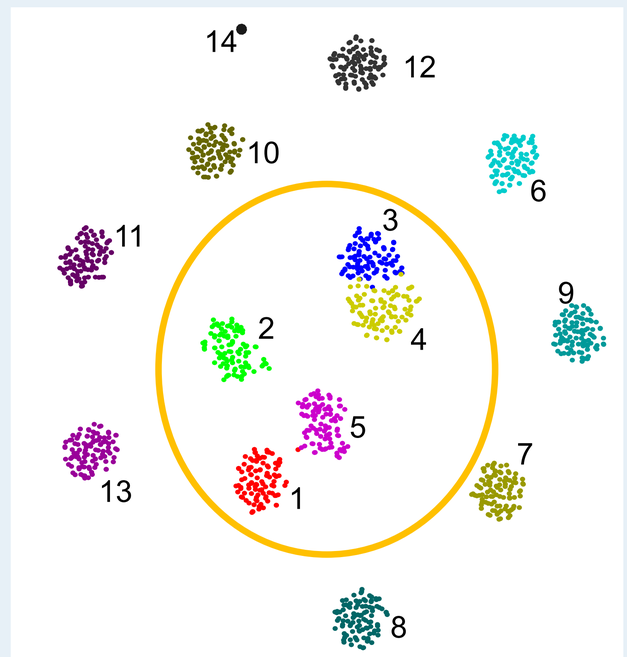


- Clustering error: Number of sentences assigned to a wrong cluster
- Test loss for all models the same, but clustering errors very different!
- Adding more linguistic tasks generally leads to better separated syntax clusters
- Translation helps, but part-of-speech tagging helps more
- French seems less helpful than German
- POS is closely related to syntax. Translation might help more for "semantic clustering"
- Clustering error was computed using k-means

REP



REP-DE-POS



- t-SNE to visualize clusters
- Comparison between a model with a single REP decoder and one with a REP, DE and POS decoder.
- The REP-DE-POS model's sentence representation clusters are more clearly separated, indicating it has a better "understanding" of syntax

Sentence Interpolation

Another way of exploring the latent space is by interpolating between sentence representations. We pick two sentences (**S**tart and **E**nd) from the test set and linearly interpolate between them. The resulting intermediate sentences can give us a clue about what the models have learned. Clearly, the model on the right (REP-DE-POS) produces more plausible sentences and fewer non-words or gibberish. This is consistent with the results from the sentence arithmetic and syntax clustering experiments. More work is needed to investigate other ways of interpolating and come up with more quantitative measures of quality.

REP

S: I think that is unfair.
I think hit item wen it.
I citible a staly by Tis.
We it hweman at you nowars.
We will ask Eur to by stalf.
We will avoid u networ for that.
E: We will make a note of this.

S: We are again faced with a long-running scandal.
We are agri-far you wait lack one sationslluz:
We argue as old issue - will renewingly, sirh leavin.
The macret and house if ACP tell-ownaires stune.
The just are lomping way I sell long-situates'.
The surl laved metho with I aloperouclous Itycs day.
E: They have dealt well with a long-running scandal.

REP-DE-POS

S: I think that is unfair.
I think that is unfair.
I think what is an often.
I who came wit anoth on I.
We will amoun this under it.
We will make a note of this.
E: We will make a note of this.

S: We are again faced with a long-running scandal.
We are again edited in a vivious' nurorial scan.
We are having failed way with round-and scandal.
They are weapen deal with a long-running scandal.
They have added dettine with a long-running scandal.
They have danged well with a long-running scandal.
E: They have dealt well with a long-running scandal.

Conclusion

We used a simple sequence to sequence autoencoder model based on vanilla LSTMs. We first trained the model to reproduce English sentences. We were interested in understanding the hidden representations learned by the models. We found that adding linguistic "auxiliary" tasks in the form of additional decoders changes the learned representations significantly. In order to analyze the changes, we fed instances of syntactic sentence prototypes to each model and extracted the resulting hidden sentence representations. The models using additional linguistic auxiliary tasks produce sentence representations that are more separated in the latent space. In the case of the best model (REP-DE-POS) and using k-means with $k = 14$, the clusters perfectly correspond to the 14 sentence prototypes. The simpler models perform significantly worse in this task. This experiment also shows that, not too surprisingly, the perplexity loss does not reflect the improvement in the syntax clustering. This highlights the importance of finding new ways of assessing the performance of generative models, beyond manual inspection by humans. We performed further experiments in the latent space such as adding difference vectors of two representations to a third one, and interpolating between representations. In the future we plan on formalizing our findings, including semantics and combining our models with state-of-the-art downstream tasks such as neural machine translation. We also want to find and disentangle other features of natural languages, such as artistic style and content.