Exploring Thematic Fit in Language with Neural Models

By

Anjani Prasad Atluri , Samrat Halder, Mughilan Muthupari, Priyadharshini Rajbabu, Jake Stamell

Mentors: Yuval Marton, Asad Sayeed, Smaranda Muresan

Table of Contents

Introduction: Thematic Fit

Our goal is to use neural models for **thematic fit**. This aims to identify how well a given word or concept fits a into a role of an event

Example sentence with roles				
Sentence	I	cut	the cake	with a knife
Roles	agent	action	patient	instrument

How would a human interpret potential role-fillers for the following sentence?

Sentence: The cake was cut with the *[instrument]* by me

Role-fillers: knife, scissors, floss, brick

Introduction: Language Models vs. Thematic Fit

Given the promising developments in pre-trained language models (e.g. BERT), one might ask whether these can be used directly for this task

- When do language models fail?
- Why we need extra information about roles?

Sentence:

I cut the cake with the [MASK].

Mask 1 Predictions:
18.0% **knife**
11.1% **fork**
5.0% **spoon**
4.4% **butter**
3.8% **bread**

Sentence:

The cake was cut with the [MASK] by me.

Mask 1 Predictions:
6.8% **knife**
4.1% **bread**
2.9% **cake**
2.8% **wine**
2.6% **butter**

Previous Work: Event Representation Models

- Goal is to predict the appropriate word in a sentence given both the role of that supposed word and the surrounding context in the form of word-role pairs.

- Non-incremental role-filler (NNRF) model
  - fails to distinguish two similarly worded sentences with different meanings
  - e.g. Kid watches TV and TV watches kid

- NNRF model is extended in three iterations:
  - NNRF-MT (multi-tasking objective)
  - RoFa-MT (Role-Filler Averaged model)
  - ResRoFa-MT (Residual connections to solve vanishing gradient issues)

Embedding Approaches

- Random - currently used in ResRoFa-MT
- Non-contextual embeddings
  - Word2vec
  - GloVe
  - FastText
- Contextual embeddings
  - RoBERTa
  - XLNet
  - ERNIE 2.0

Our Contributions

- ❑ Do non-contextual embeddings outperform random embeddings?
- ❑ How important is tuning embeddings for this specific task?
- ❑ How does the role embedding size (input style) affect model performance?
- ❑ Do contextual embedding outperform non-contextual embedding?
- ❑ Rapid optimization of training speed and codebase
- ❑ Other explorations with ResRoFa-MT model architecture.

Experiments

- ❑ Separate vs. Shared Embedding Layers
- ❑ Fixed vs. Tuned Random Embeddings
- ❑ Fixed vs. Tuned Non-contextual Embeddings
- ❑ Shrinking Role Embeddings
- ❑ Orthogonal Role Embeddings

Baseline : We use random embeddings as baseline of our experiments

Dataset: We use a 10% sample of the RW-Eng-v2 corpus for training (see appendix for additional details on the corpus)

Experiments: Separate vs. Shared Embedding Layers

Description

- ResRoFa has a pair of embedding layers: 1) input words and roles; 2) target word and role
- Second set is used for the prediction task
- Goal is to test whether a single set of embeddings can be used for both purposes (new model named RRF-Shared)

Results

- Test performance for loss, word prediction accuracy, and role prediction accuracy are nearly identical; similar performance on thematic fit tasks
- RRF-Shared has ~50% fewer parameters than ResRoFa(~30M vs. ~67M)

Model	Initial Embedding	Fixed/Tuned	Test Loss	Test Role Accuracy	Test Word Accuracy	PADO-all Correlation	McRae-all Correlation
ResRoFa	Random	Tuned	5.49	94.00%	29.66%	0.26	**0.30**
RRF-Shared	**Random**	**Tuned**	**5.49**	**94.10%**	**29.66%**	**0.30**	**0.28**

Experiments: Fixed vs. Tuned Random Embeddings

Description

- ❑ Using RRF-Shared, we compare two random initializations of the model
- ❑ One where embeddings are held fixed and another where they are tuned

Results

- ❑ Significant drop-off in role prediction accuracy when holding embeddings fixed (tuned-96%, fixed-68%)
- ❑ Drastic performance difference on thematic fit tests (PADO, McRae)

Model	Initial Embedding	Fixed/Tuned	Test Loss	Test Role Accuracy	Test Word Accuracy	PADO-all Correlation	McRae-all Correlation
RRF-Shared	Random	Fixed	6.08	75.79%	29.66%	-0.05	-0.01
RRF-Shared	Random	Tuned	5.49	94.10%	29.66%	0.30	0.28

Experiments: Fixed vs. Tuned Non-Contextual

Description

- We initialize embeddings with non-contextual word embeddings: Word2Vec, FastText, GloVe
- We compare fixing vs. tuning embeddings
- Non-contextual embeddings can have out-of-vocabulary (OOV) words
  - Additional experiments test how OOV embeddings are initialized

Results

- The role and word prediction accuracy improved when we fine tuned the embeddings as expected
- Also we observed performance improvement in the thematic fit tests with fine tuning

Model	Initial Embedding	Fixed/Tuned	Test Loss	Test Role Accuracy	Test Word Accuracy	PADO-all Correlation	McRae-all Correlation
RRF-Shared	GloVe	Fixed	5.34	93.58%	29.66%	0.05	-0.09
RRF-Shared	GloVe	Tuned	5.34	94.15%	29.66%	0.37	0.23
RRF-Shared	FastText	Fixed	5.35	93.74%	29.66%	0.22	0.24
RRF-Shared	**FastText**	**Tuned**	**5.34**	**94.10%**	**29.66%**	**0.32**	**0.31**

Experiments: Fixed vs. Tuned Non-Contextual

Embedding Source	OOV initialization	Fixed/Tuned	Validation Loss	Validation Role Accuracy	Validation Word Accuracy
Word2Vec	Avg	Fixed	5.98	96.22%	13.65%
Word2Vec	Null	Fixed	5.98	96.26%	13.61%
Word2Vec	**Avg**	**Tuned**	**5.97**	**96.68%**	**13.87%**
Word2Vec	Null	Tuned	5.98	96.67%	13.86%
FastText	Avg	Fixed	5.99	96.36%	13.44%
FastText	Null	Fixed	6.00	96.35%	13.43%
FastText	**Avg**	**Tuned**	**5.98**	**96.64%**	**13.83%**
FastText	Null	Tuned	5.98	97.64%	13.79%
GloVe	Avg	Fixed	5.99	96.03%	13.51%
GloVe	Null	Fixed	5.99	95.98%	13.51%
GloVe	**Avg**	**Tuned**	**5.97**	**96.62%**	**13.87%**
GloVe	Null	Tuned	5.98	96.65%	13.87%

Experiments: Shrinking Role Embeddings

Description

- Prior models use size 300 embeddings
- Experiment a new model (RRF-Small Role or RRF-SR) that uses randomly initialized role embeddings of size 3, 30, and 300.
  - Size 300 role embeddings in RRF-SR corresponds to the same size as the baseline; however, the method of composition is different.
  - Rather than using the Hadamard product, we now concatenate the embeddings

Results

- While role accuracy is similar across the runs, shrinking role embeddings sees a deterioration in performance on loss and word accuracy

Embedding Size	Validation Loss	Validation Role Accuracy	Validation Word Accuracy
3	6.13	96.56%	12.93%
30	**6.10**	**96.57%**	**12.98%**
300	6.20	96.50%	12.64%

Experiments: Orthogonal Role Embeddings

Description

- ❑ Extend low dimensional roles to one hot orthogonal vectors instead of random
- ❑ Randomly initialized embeddings can place some roles closer to each other in the vector space

Results

- ❑ Orthogonally initialized embeddings perform similar to the other lower dimensional role embedding experiments
- ❑ Experimented on 10% data.

Key Takeaways

- ❑ Shared embeddings for input and target words/roles performs well with a drastic reduction in model size
- ❑ Strong performance on the validation/test sets does not necessarily equate to strong performance on external thematic fit benchmarks
- ❑ Tuning embeddings is vital for performance on thematic fit evaluation tasks
- ❑ RRF-Shared with non-contextual initialized embeddings does not significantly outperform randomly initialized embeddings
- ❑ Based on initial results, smaller role embeddings introduced in RRF-SR do not improve performance

Ethical Considerations

- Data is sourced entirely from UK web sources/proceedings from the 20th century
  - This is not a heterogeneous dataset and could have biases embedded in it
- Automatically parsing the corpus could introduce additional biases from the parsing algorithms
  - At the very least, there is minimal validation of the data, which could hurt model performance
- Pre-trained language models have their own set of issues as well
  - They are known to encode biases and can be affected by toxic data
  - Environmental footprint of training models is enormous

Future Work

- Incorporate a more efficient way to make it feasible to run ResRoFa model with contextual embeddings
- Integrate team 2's work on the model architecture side

Acknowledgement

We would like to thank Dr. Yuval Marton, and Dr. Asad Sayeed for giving us the opportunity to work on this project, discussing different ideas, and guiding us throughout the course of this project. We are grateful to Dr. Smaranda Muresan and Dr. Eleni Drinea for helping us with the logistics and providing us required computing resources. Lastly, we are thankful to Bloomberg, DSI and CS Department at the Columbia University for for providing us with all kinds of support throughout the semester and our graduate program.

THANK YOU!

References

Appendix

NNRF-MT

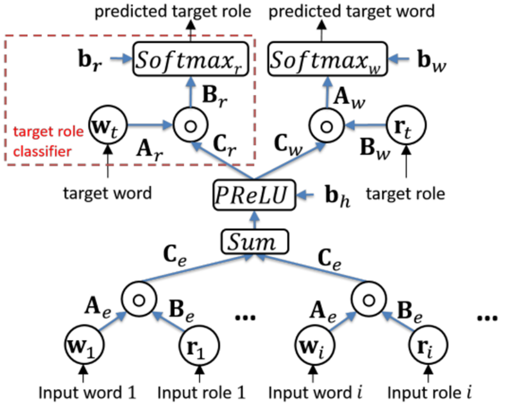

Figure 1: Architecture of multi-task role-filler model.

RoFa-MT

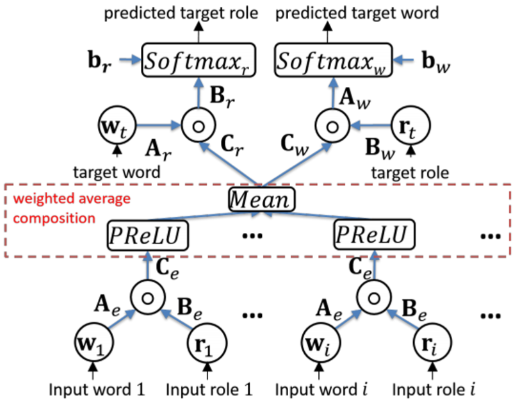

Figure 2: Architecture of role-filler averaging model.

ResRoFa-MT

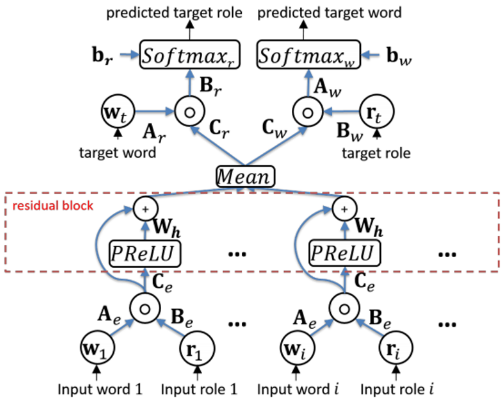

Figure 3: Architecture of residual role-filler averaging model.

Corpus

- Rollenwechsel-English (**RW-eng**)
  - ***Propbank*** approach to Semantic Role Labelling (SRL)
  - Corpus is referenced from ***ukWaC*** and ***British National Corpus***
  - Two versions of the corpus (*RW-eng-v1* and ***RW-eng-v2***)

- This corpus applies dependency parsing algorithms combined with heuristics in order to perform the SRL task
  - While this allows the creation of a large dataset, it also means the samples can be very messy

Experiment Results (models trained on 10% data)

Model	Initial Embedding	Fixed/Tuned	Validation Loss	Validation Role Accuracy	Validation Word Accuracy
RRF-Shared	Random	Fixed	6.09	75.93%	29.63%
RRF-Shared	Random	Tuned	5.50	94.14%	29.63%
RRF-Shared	GloVe (Avg.)	Fixed	5.52	93.59%	29.63%
RRF-Shared	GloVe (Avg.)	Tuned	5.50	94.13%	29.63%
RRF-Shared	FastText (Avg.)	Fixed	5.51	93.86%	29.63%
RRF-Shared	FastText (Avg.)	Tuned	5.50	94.14%	29.63%
RRF-SR	Orthogonal Role	Tuned	5.59	93.45%	29.63%
RRF-SR	30-dimensional Role	Tuned	5.64	93.26%	29.63%