Linguistic Dependencies and Statistical Dependence

Jacob Louis Hoover[1,3], **Alessandro Sordoni**[2], **Wenyu Du**[4], and **Timothy J. O'Donnell**[1,3]

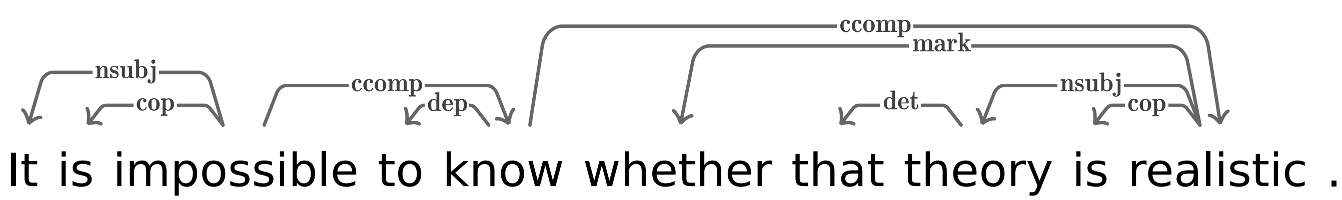
https://arxiv.org/abs/2104.08685

2021-10-11

1 McGill
2 Microsoft Research
3 Mila
4 香 港 大 學 THE UNIVERSITY OF HONG KONG

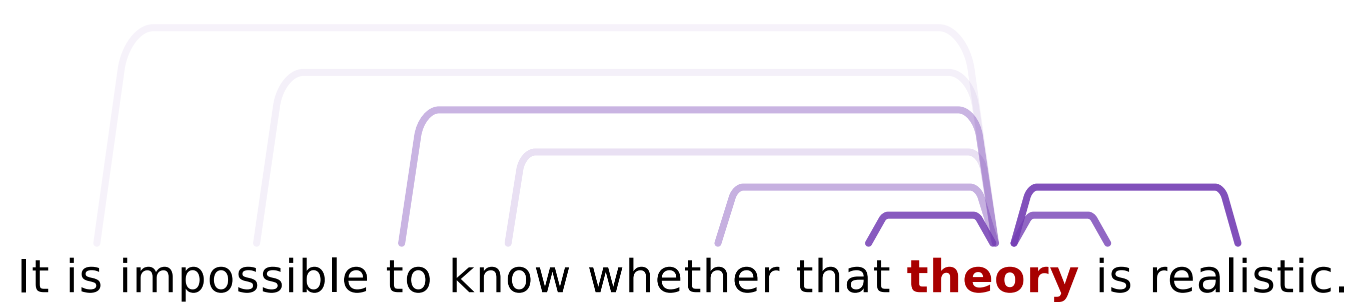
linguistic dependency

how are words combined to make a sentence?



statistical dependence

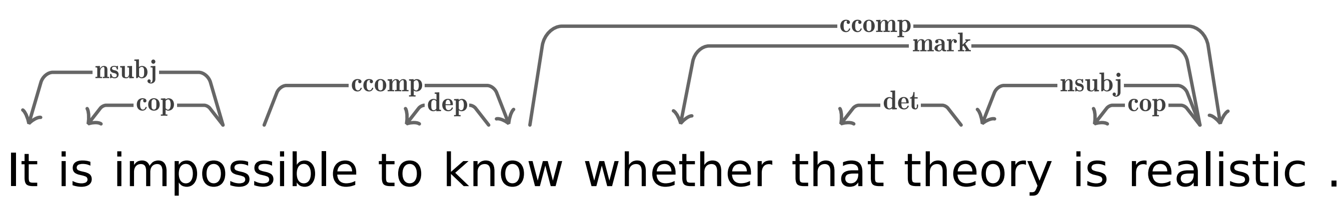
how do words inform the probability of other words?



linguistic dependency

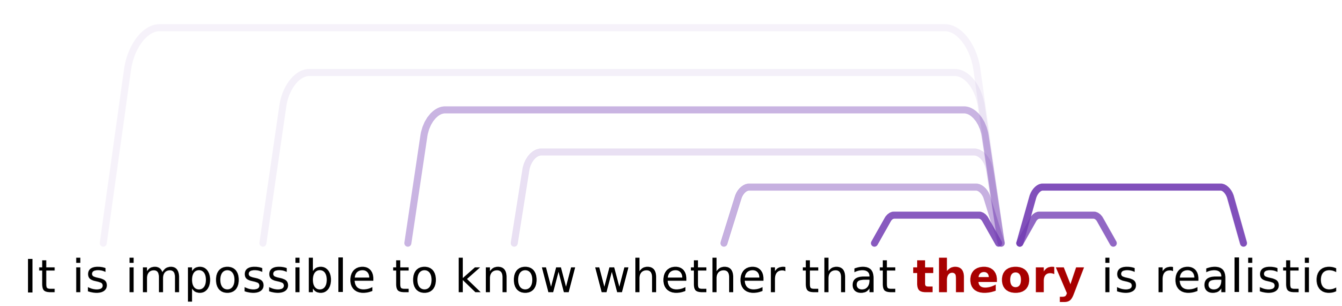
how are words combined to make a sentence?

- tree structure of word-to-word links
- representing compositional structure / trace of the computation to build sentence



statistical dependence

how do words inform the probability of other words?



linguistic dependency

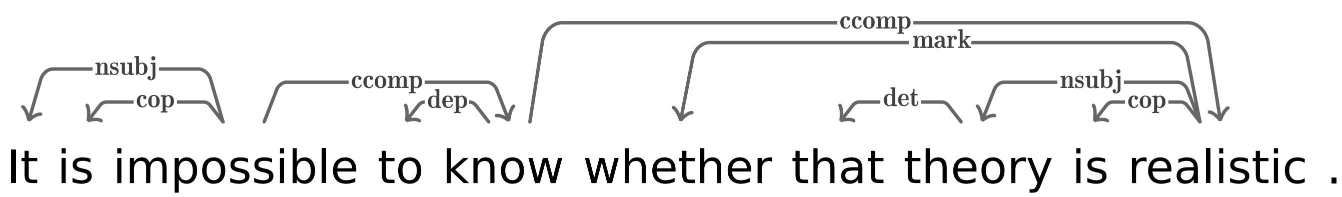
how are words combined to make a sentence?

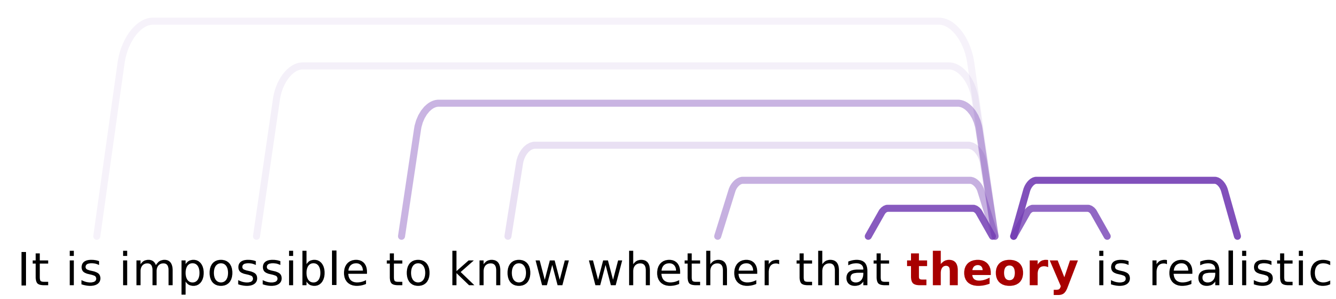
- tree structure of word-to-word links
- representing compositional structure / trace of the computation to build sentence

statistical dependence

how do words inform the probability of other words?

- plays important role in theories of processing and acquisition
- in NLP, successful models use *language modelling loss* learning patterns of statistical dependence





linguistic dependency

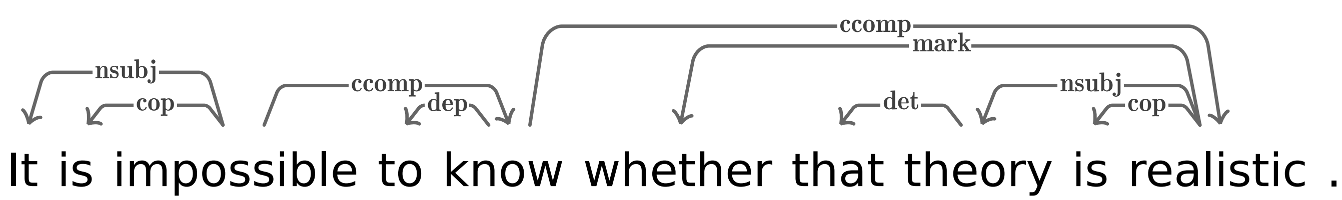
how are words combined to make a sentence?

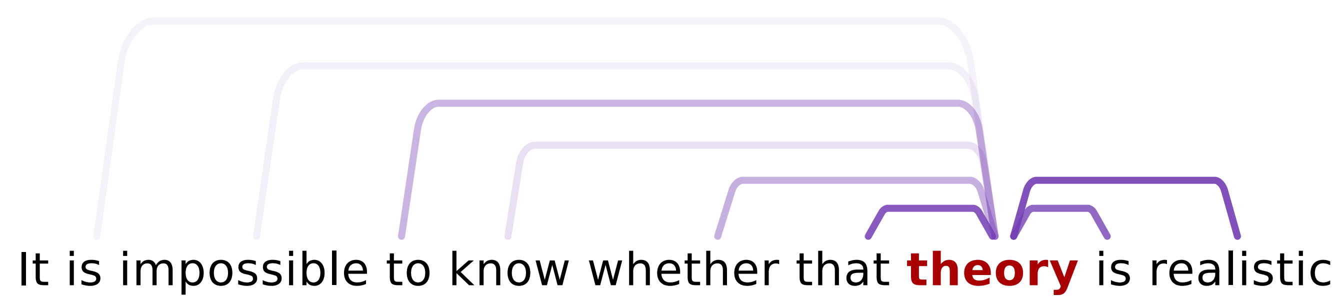
- tree structure of word-to-word links
- representing compositional structure / trace of the computation to build sentence

statistical dependence

how do words inform the probability of other words?

- plays important role in theories of processing and acquisition
- in NLP, successful models use *language modelling loss* learning patterns of statistical dependence





how are they connected?

linguistic dependency & statistical dependence

how are they connected?

linguistic dependency & statistical dependence

how are they connected?

- Long tradition of unsupervised dependency parsing assumes a connection. Also explored in earlier statistical studies
    - Magerman and Marcus (1990), de Paiva Alves (1996) …

linguistic dependency & statistical dependence

how are they connected?

- Long tradition of unsupervised dependency parsing assumes a connection. Also explored in earlier statistical studies
  - Magerman and Marcus (1990), de Paiva Alves (1996) …
- Recently some work has explicitly proposed that linguistic dependencies connect words that are statistically dependent
  - Futrell et al. (2019): *Syntactic dependencies correspond to word pairs with high mutual information.*

linguistic dependency & statistical dependence

how are they connected?

- Long tradition of unsupervised dependency parsing assumes a connection. Also explored in earlier statistical studies
  - Magerman and Marcus (1990), de Paiva Alves (1996) …
- Recently some work has explicitly proposed that linguistic dependencies connect words that are statistically dependent
  - Futrell et al. (2019): *Syntactic dependencies correspond to word pairs with high mutual information.*
  - very recently, Zhang & Hashimoto (2021): *On the Inductive Bias of Masked Language Modeling: From statistical to syntactic dependencies.* [Closely related study, simultaneous to ours. I'll return to this]

linguistic dependency & statistical dependence

our investigation

We set out to answer the question: Are words that are *statistically* dependent likely to be in *linguistic* dependencies?

- Estimate statistical dependence between words **using modern pretrained contextualized language models** (e.g. BERT, XLNet)—our current best estimators of probability of words in context—rather than earlier statistical techniques

We find that connecting words which are statistically dependent and comparing with linguistic dependency yields **accuracy only as high as simple baseline connecting adjacent words**.

- true across languages,
- true for syntactically-aware LMs,
- true statistical dependencies between POS tags too

Contextualized Pointwise Mutual Information

our measure of statistical dependence between words

- Pointwise mutual information (PMI) between x and y, in context c, is

$$\operatorname{pmi}(x; y \mid c) \equiv \log \frac{p(x, y \mid c)}{p(x \mid c) p(y \mid c)} = \log \frac{p(x \mid y, c)}{p(x \mid c)} .$$

Contextualized Pointwise Mutual Information

our measure of statistical dependence between words

- Pointwise mutual information (PMI) between x and y, in context c, is

$$\mathrm{pmi}(x; y \mid c) \equiv \log \frac{p(x, y \mid c)}{p(x \mid c) p(y \mid c)} = \log \frac{p(x \mid y, c)}{p(x \mid c)}.$$

- We define **contextualized pointwise mutual information** (**CPMI**) between words estimated using language model M, as

Contextualized Pointwise Mutual Information

our measure of statistical dependence between words

- Pointwise mutual information (PMI) between x and y, in context c, is

$$\mathrm{pmi}(x; y \mid c) \equiv \log \frac{p(x, y \mid c)}{p(x \mid c) p(y \mid c)} = \log \frac{p(x \mid y, c)}{p(x \mid c)}.$$

- We define **contextualized pointwise mutual information** (**CPMI**) between words estimated using language model M, as

$$\mathrm{CPMI}_M(w_i; w_j) \equiv \log \frac{p_M(w_i \mid W_{-i})}{p_M(w_i \mid W_{-i,j})}$$

Contextualized Pointwise Mutual Information

our measure of statistical dependence between words

- Pointwise mutual information (PMI) between x and y, in context c, is

$$\mathrm{pmi}(x; y \mid c) \equiv \log \frac{p(x, y \mid c)}{p(x \mid c) p(y \mid c)} = \log \frac{p(x \mid y, c)}{p(x \mid c)}.$$

- We define **contextualized pointwise mutual information (CPMI)** between words estimated using language model M, as

$$\mathrm{CPMI}_M(w_i; w_j) \equiv \log \frac{p_M(w_i \mid W_{-i})}{p_M(w_i \mid W_{-i,j})}$$

- for example, $s =$ That $\overset{w_j}{\text{theory}}$ is $\overset{w_i}{\text{realistic}}$.

Contextualized Pointwise Mutual Information

our measure of statistical dependence between words

- Pointwise mutual information (PMI) between x and y, in context c, is

$$\mathrm{pmi}(x; y \mid c) \equiv \log \frac{p(x, y \mid c)}{p(x \mid c)p(y \mid c)} = \log \frac{p(x \mid y, c)}{p(x \mid c)} .$$

- We define **contextualized pointwise mutual information (CPMI)** between words estimated using language model M, as

$$\mathrm{CPMI}_M(w_i; w_j) \equiv \log \frac{p_M(w_i \mid W_{-i})}{p_M(w_i \mid W_{-i,j})}$$

- for example, $s = \text{That}\ \underbrace{\overset{w_j}{\text{theory}}\ \text{is}\ \overset{w_i}{\text{realistic}}}_{\mathrm{CPMI}_M(w_i;\, w_j)}\ .$

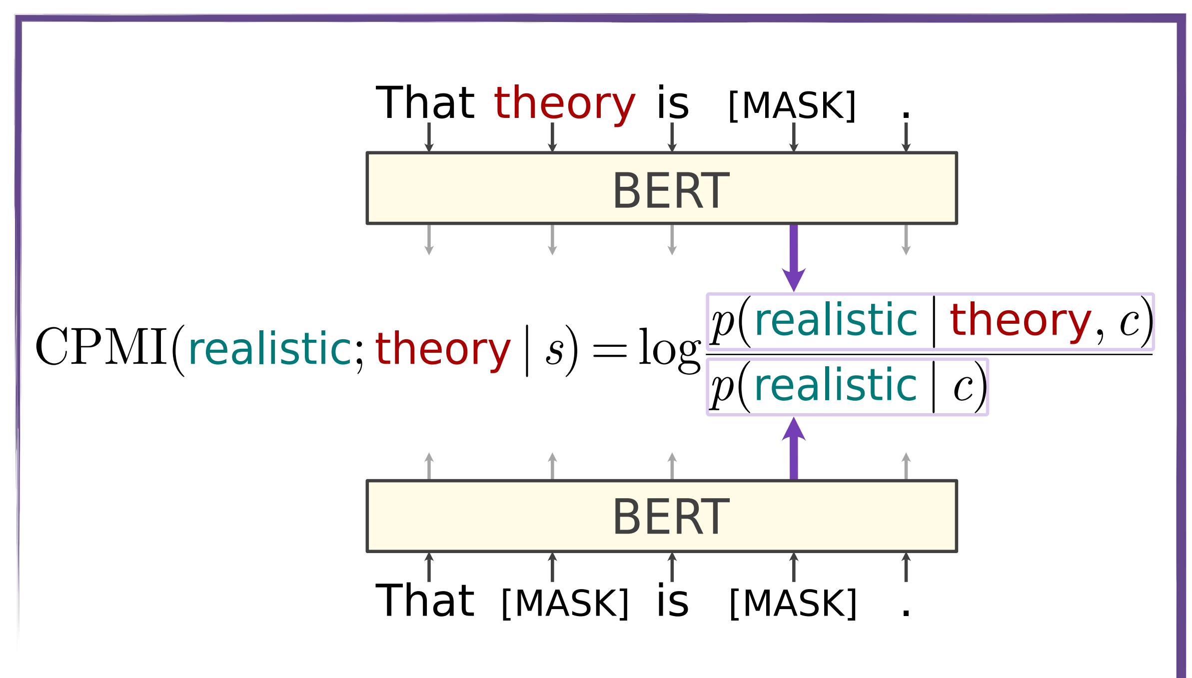
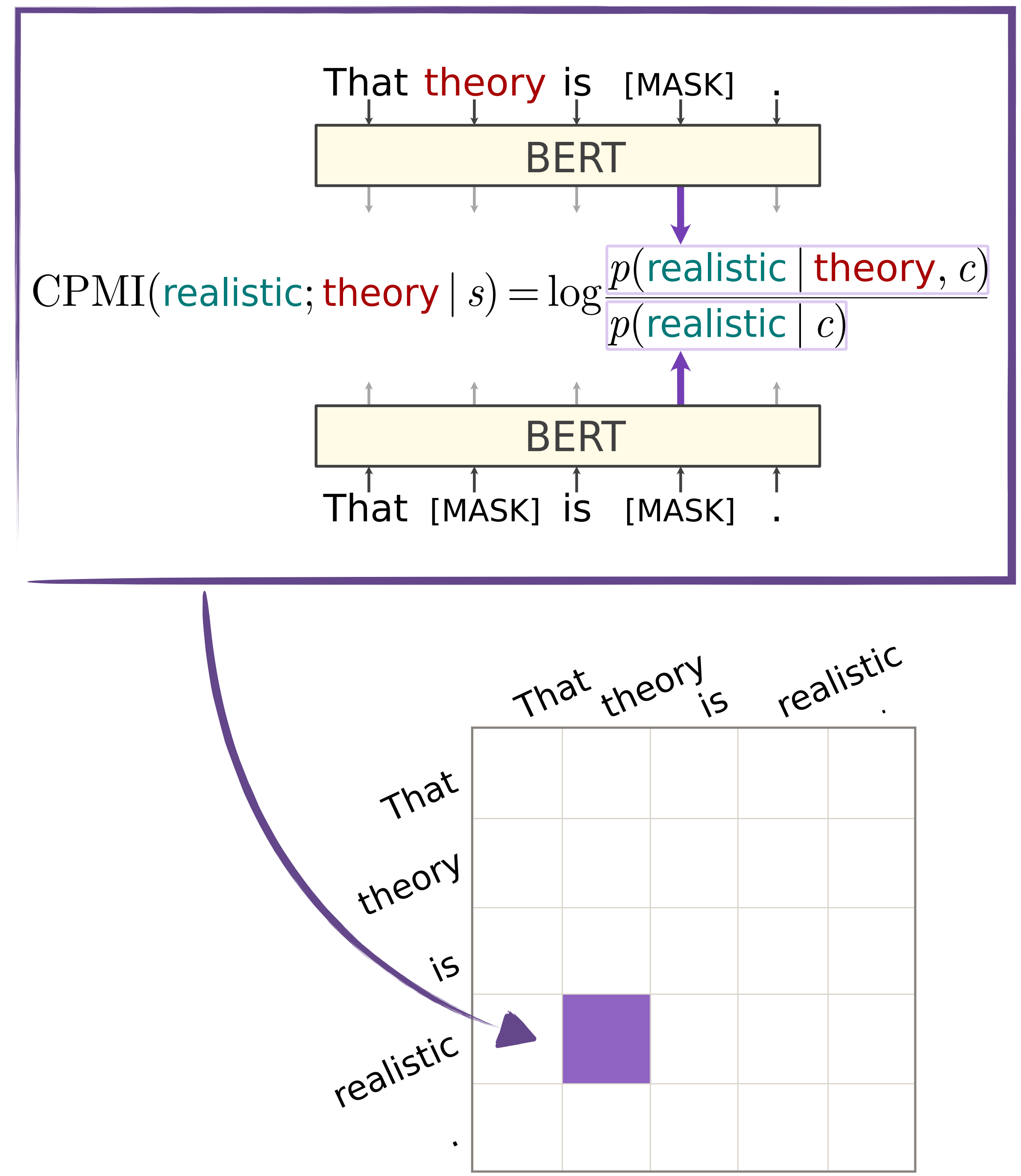


Figure 2 in paper. using BERT to compute the probability of realistic with and without masking theory.

CPMI-dependency parsing

method

1. compute of CPMI values

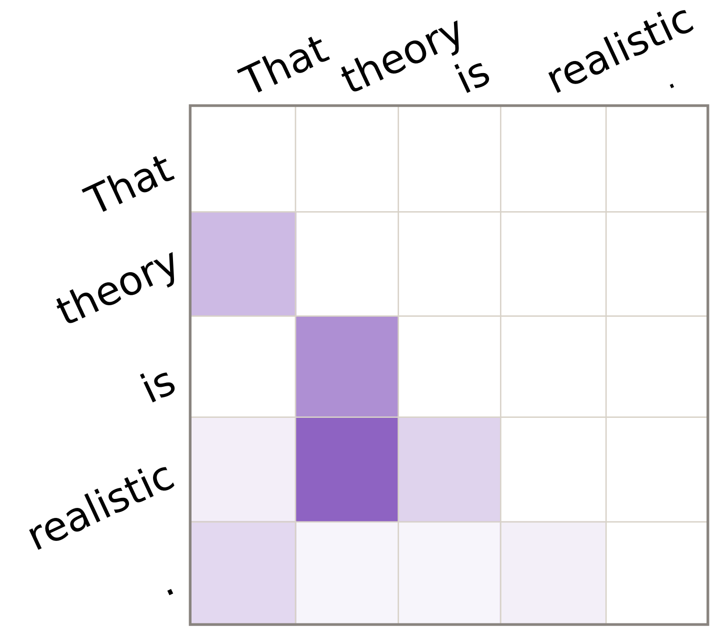


$$s = \text{That } \underbrace{\overset{w_j}{\text{theory}} \text{ is } \overset{w_i}{\text{realistic}}}_{\mathrm{CPMI}_M(w_i; w_j)} \text{ .}$$

CPMI-dependency parsing

method

That theory is realistic .

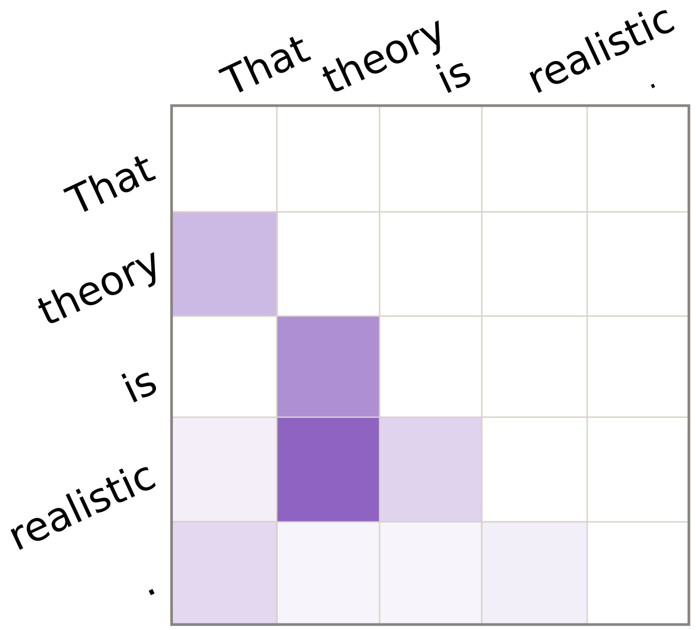


CPMI-dependency parsing

method

1. compute of CPMI values for each pair of words in sentence

That theory is realistic .



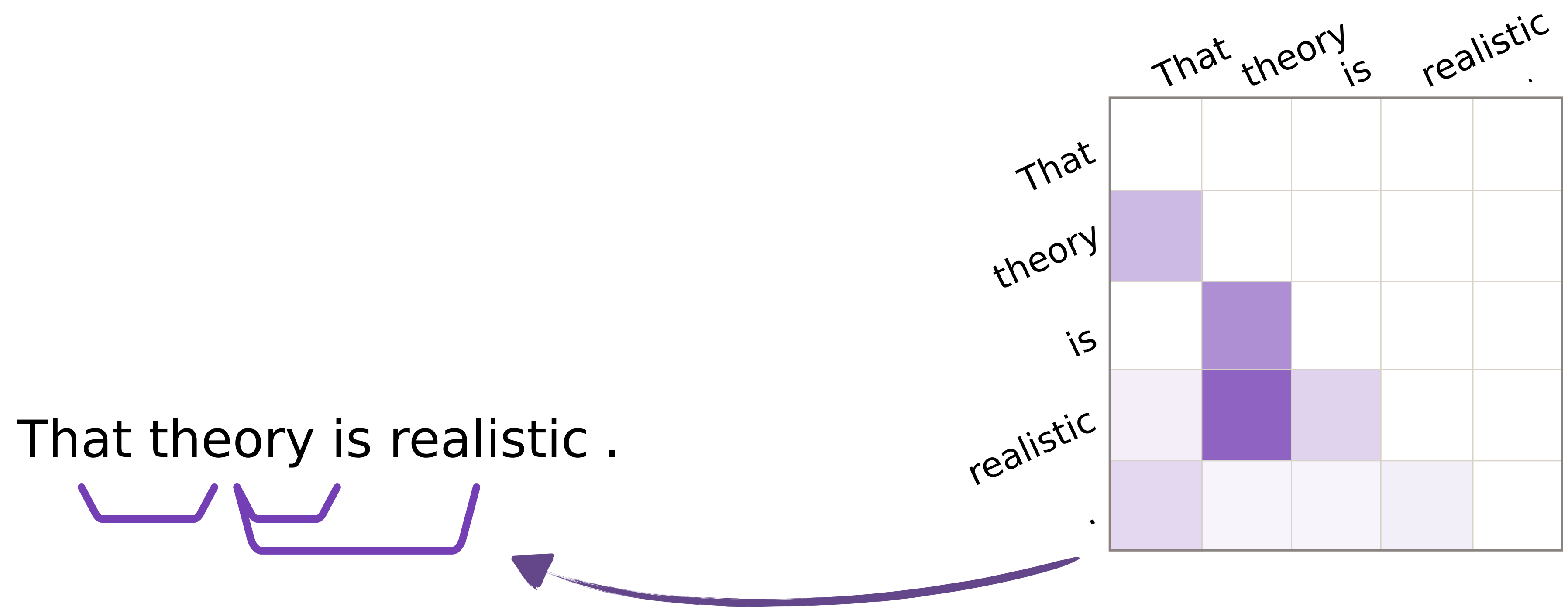
CPMI-dependency parsing

method

1. compute of CPMI values for each pair of words in sentence

   - extract the maximum-CPMI spanning tree

   Note: PMI is symmetric, but LM's estimates may not be. We symmetrize the matrix first.



CPMI-dependency parsing

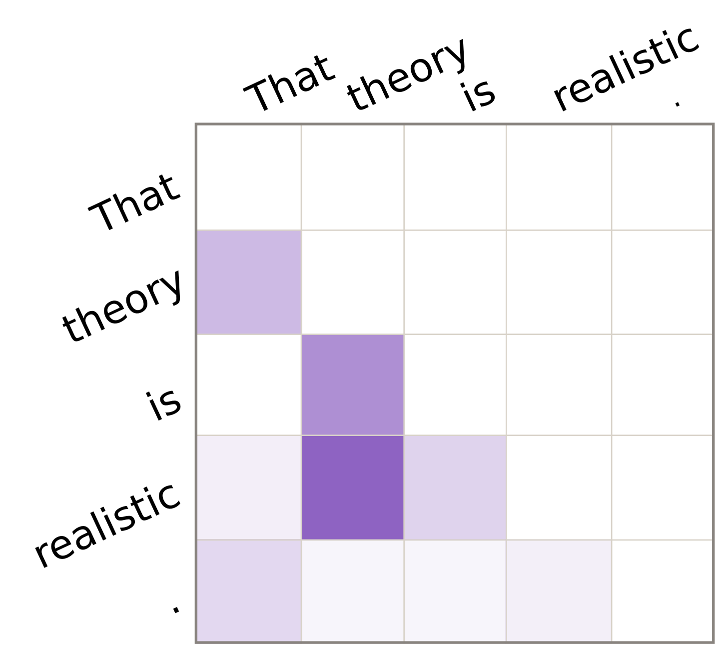
method

1. compute of CPMI values for each pair of words in sentence

   - extract the maximum-CPMI spanning tree

   Note: PMI is symmetric, but LM's estimates may not be. We symmetrize the matrix first.

2. compare max-CPMI tree to gold tree



CPMI-dependency parsing

using large pretrained LMs

random	.22
connect-adjacent	**.49**
Word2Vec	.39
BERT base	.46
BERT large	.47
DistilBERT	.48
Bart large	.38
XLM	.42
XLNet base	.45
XLNet large	.41

Table 1 in paper.
Unlabeled undirected attachement score (UUAS) for max-CPMI trees pretrained language models on PTB dev split (sec 22).

CPMI-dependency parsing

using large pretrained LMs

Here are results (accuracy scores as UUAS = *unlabeled undirected attachment score*: the number of edges in common with gold dependencies)

random	.22
connect-adjacent	**.49**
Word2Vec	.39
BERT base	.46
BERT large	.47
DistilBERT	.48
Bart large	.38
XLM	.42
XLNet base	.45
XLNet large	.41

Table 1 in paper.
Unlabeled undirected attachement score (UUAS) for max-CPMI trees pretrained language models on PTB dev split (sec 22).

CPMI-dependency parsing

using large pretrained LMs

Here are results (accuracy scores as UUAS = *unlabeled undirected attachment score*: the number of edges in common with gold dependencies)

baselines

- random
- connect-adjacent-words
- Word2Vec (noncontextual)

random	.22
connect-adjacent	**.49**
Word2Vec	.39
BERT base	.46
BERT large	.47
DistilBERT	.48
Bart large	.38
XLM	.42
XLNet base	.45
XLNet large	.41

Table 1 in paper.
Unlabeled undirected attachement score (UUAS) for max-CPMI trees pretrained language models on PTB dev split (sec 22).

CPMI-dependency parsing

using large pretrained LMs

Here are results (accuracy scores as UUAS = *unlabeled undirected attachment score*: the number of edges in common with gold dependencies)

baselines

- random
- connect-adjacent-words
- Word2Vec (noncontextual)

For all the large pretrained LMs, overall attachment score is **no higher than the connect-adjacent baseline**.

random	.22
connect-adjacent	**.49**
Word2Vec	.39
BERT base	.46
BERT large	.47
DistilBERT	.48
Bart large	.38
XLM	.42
XLNet base	.45
XLNet large	.41

Table 1 in paper.
Unlabeled undirected attachement score (UUAS) for max-CPMI trees pretrained language models on PTB dev split (sec 22).

CPMI-dependency parsing

using large pretrained LMs

Here are results (accuracy scores as UUAS = *unlabeled undirected attachment score*: the number of edges in common with gold dependencies)

baselines

- random
- connect-adjacent-words
- Word2Vec (noncontextual)

For all the large pretrained LMs, overall attachment score is **no higher than the connect-adjacent baseline**.

random	.22
connect-adjacent	**.49**
Word2Vec	.39
BERT base	.46
BERT large	.47
DistilBERT	.48
Bart large	.38
XLM	.42
XLNet base	.45
XLNet large	.41

Table 1 in paper.
Unlabeled undirected attachement score (UUAS) for max-CPMI trees pretrained language models on PTB dev split (sec 22).

CPMI-dependency parsing

comparison with Zhang & Hashimoto (2021)

Method	UUAS
RANDOM	9.14 ± 0.42
LINEARCHAIN	47.69
Klein and Manning (2004)	48.76 ± 0.24
PMI	28.05
CONDITIONAL PMI	44.75 ± 0.09
CONDITIONAL MI	$\mathbf{50.62} \pm 0.38$

Table 4 in Zhang and Hashimoto (2021).
Unlabeled undirected attachement score (UUAS) using **BERT** base on subsampled PTB test split (sec 23).

random	.22
connect-adjacent	**.49**
Word2Vec	.39
BERT base	.46
BERT large	.47
DistilBERT	.48
Bart large	.38
XLM	.42
XLNet base	.45
XLNet large	.41

Table 1 in paper.
Unlabeled undirected attachement score (UUAS) for max-CPMI trees pretrained language models on PTB dev split (sec 22).

Their method is slightly different, but their results are very similar (though their interpretation is different).

For their study as for ours, attachment score is **about as high as the connect-adjacent baseline.**

CPMI-dependency parsing

using large pretrained LM (multilingual)

CPMI-dependency parsing

using large pretrained LM (multilingual)

Q: Is the similarity in accuracy to the attach-adjacent baseline particular to English?

CPMI-dependency parsing

using large pretrained LM (multilingual)

Q: Is the similarity in accuracy to the attach-adjacent baseline particular to English?

A: No.

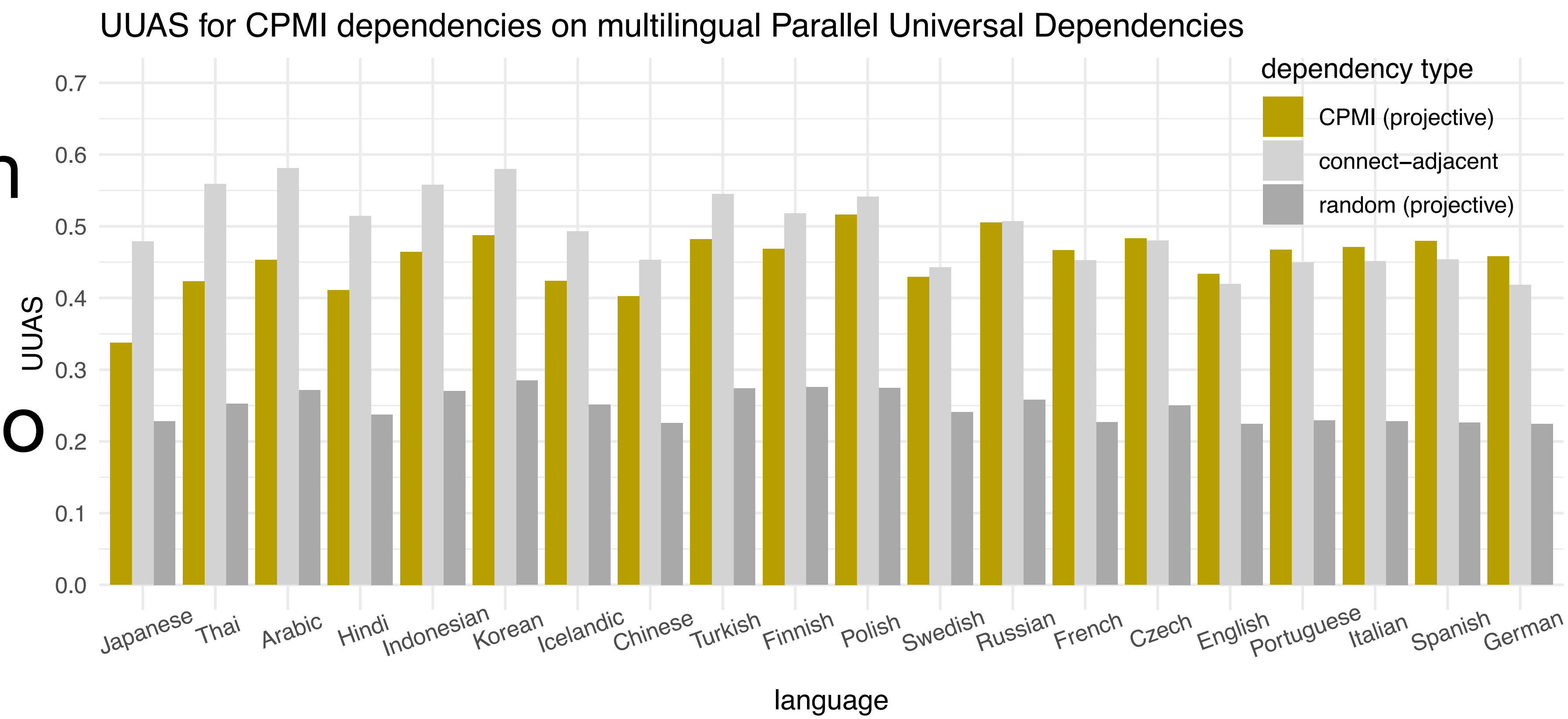


Figure 13 in paper. Unlabeled undirected attachement score (UUAS) for max-CPMI trees from BERT-multilingual.

CPMI-dependency parsing

using large pretrained LM (multilingual)

Q: Is the similarity in accuracy to the attach-adjacent baseline particular to English?

A: No.

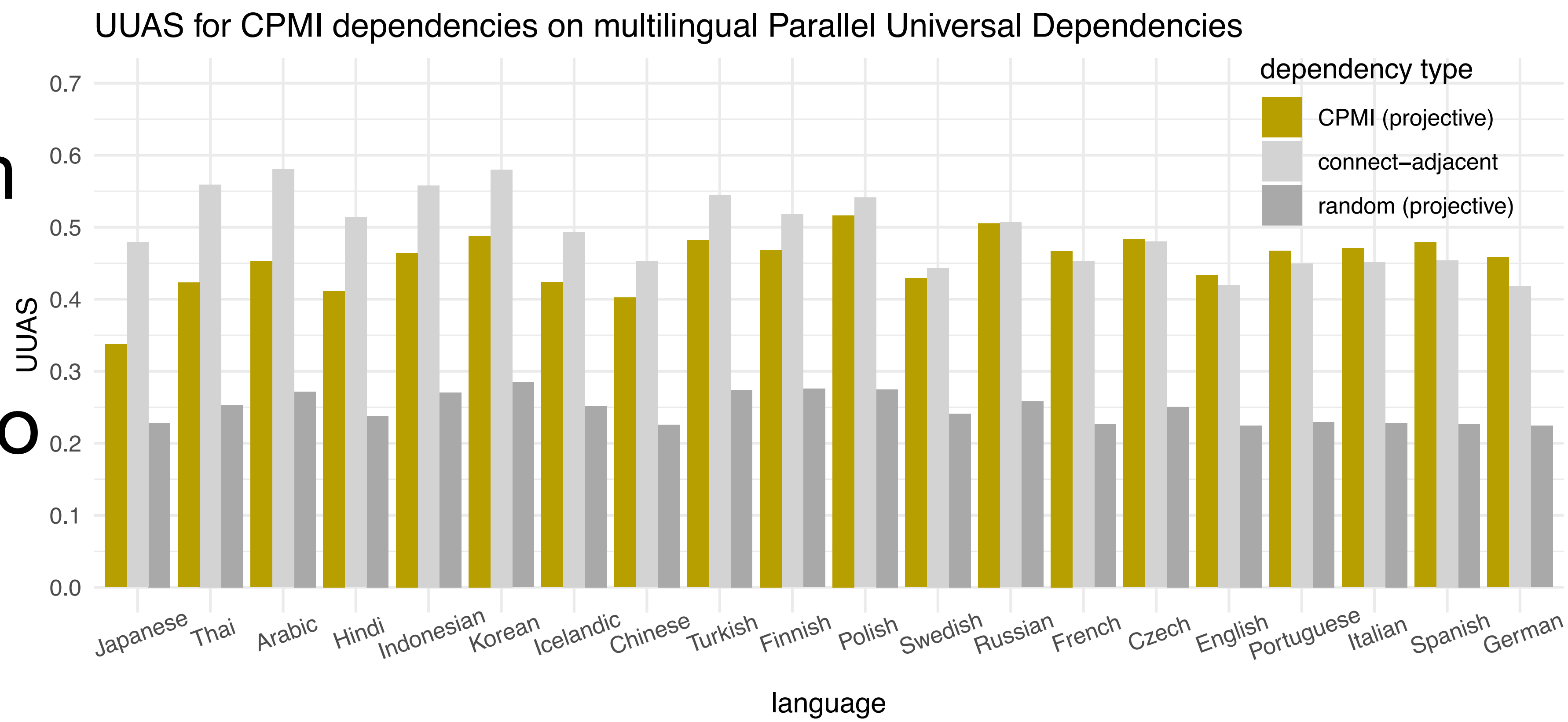


Figure 13 in paper. Unlabeled undirected attachement score (UUAS) for max-CPMI trees from BERT-multilingual.

Across 20 languages (from multiple language families), the overall attachment score is still **only about as high as the connect-adjacent baseline.**

CPMI-dependency parsing

using large pretrained LM (multilingual)

Q: Is the similarity in accuracy to the attach-adjacent baseline particular to English?

A: No.

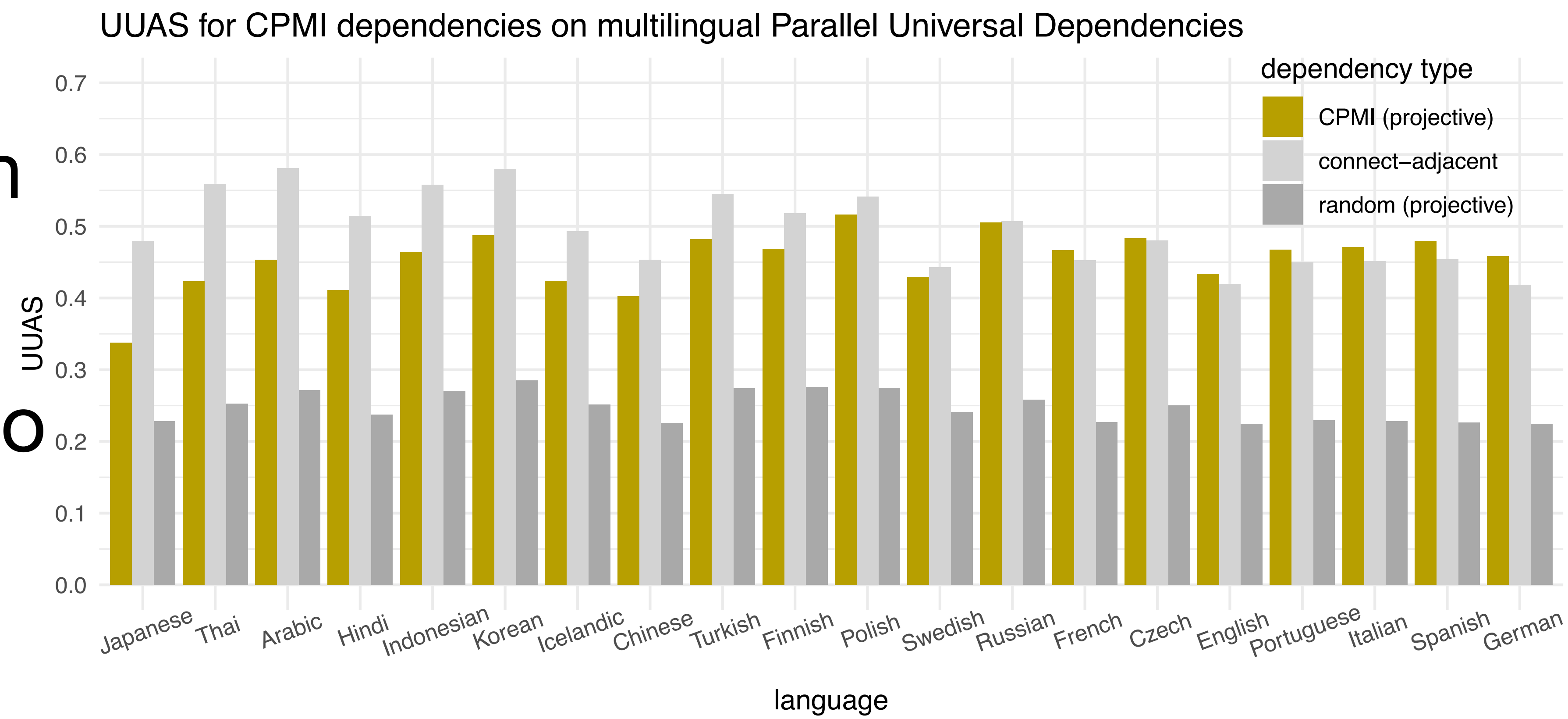


Figure 13 in paper. Unlabeled undirected attachement score (UUAS) for max-CPMI trees from BERT-multilingual.

Across 20 languages (from multiple language families), the overall attachment score is still **only about as high as the connect-adjacent baseline.**

CPMI-dependency parsing

using syntactically-aware models

CPMI-dependency parsing

using syntactically-aware models

Q: Is accuracy higher using models designed to have linguistically-oriented structural bias?

CPMI-dependency parsing

using syntactically-aware models

Q: Is accuracy higher using models designed to have linguistically-oriented structural bias?

A: No.

random	.22
connect-adjacent	**.49**
vanilla LSTM	.44
ONLSTM	.44
ONLSTM-SYD	.45

Table 1 in paper.
Unlabeled undirected attachment score (UUAS) from syntactically-aware LSTM models on PTB dev split (sec 22).

- **ONLSTM**: LSTM-based language model with inductive bias to model hierarchical structures
- **ONLSTM-SYD**: ONLSTM with additional auxiliary task to predict syntactic parses

CPMI-dependency parsing

using syntactically-aware models

Q: Is accuracy higher using models designed to have linguistically-oriented structural bias?

A: No.

baseline

- vanilla LSTM

random	.22
connect-adjacent	**.49**
vanilla LSTM	.44
ONLSTM	.44
ONLSTM-SYD	.45

Table 1 in paper.
Unlabeled undirected attachment score (UUAS) from syntactically-aware LSTM models on PTB dev split (sec 22).

- **ONLSTM**: LSTM-based language model with inductive bias to model hierarchical structures
- **ONLSTM-SYD**: ONLSTM with additional auxiliary task to predict syntactic parses

CPMI-dependency parsing

using syntactically-aware models

Q: Is accuracy higher using models designed to have linguistically-oriented structural bias?

A: No.

baseline

- vanilla LSTM

Also for syntactically aware models, overall attachment score is **no higher than the connect-adjacent baseline**.

random	.22
connect-adjacent	**.49**
vanilla LSTM	.44
ONLSTM	.44
ONLSTM-SYD	.45

Table 1 in paper.
Unlabeled undirected attachment score (UUAS) from syntactically-aware LSTM models on PTB dev split (sec 22).

- **ONLSTM**: LSTM-based language model with inductive bias to model hierarchical structures
- **ONLSTM-SYD**: ONLSTM with additional auxiliary task to predict syntactic parses

CPMI-dependency parsing

using syntactically-aware models

Q: Is accuracy higher using models designed to have linguistically-oriented structural bias?

A: No.

baseline

- vanilla LSTM

Also for syntactically aware models, overall attachment score is **no higher than the connect-adjacent baseline**.

random	.22
connect-adjacent	**.49**
vanilla LSTM	.44
ONLSTM	.44
ONLSTM-SYD	.45

Table 1 in paper.
Unlabeled undirected attachment score (UUAS) from syntactically-aware LSTM models on PTB dev split (sec 22).

- **ONLSTM**: LSTM-based language model with inductive bias to model hierarchical structures
- **ONLSTM-SYD**: ONLSTM with additional auxiliary task to predict syntactic parses

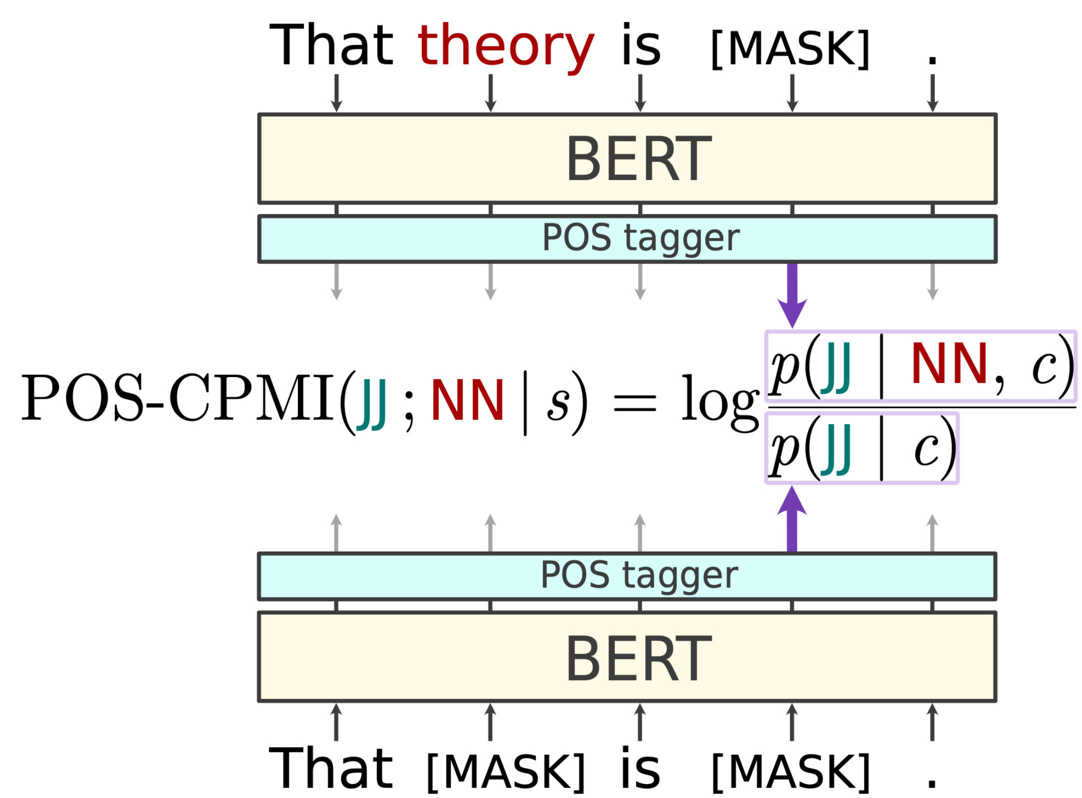
POS-CPMI-dependency parsing

a delexicalized version of CPMI

POS-CPMI-dependency parsing

a delexicalized version of CPMI

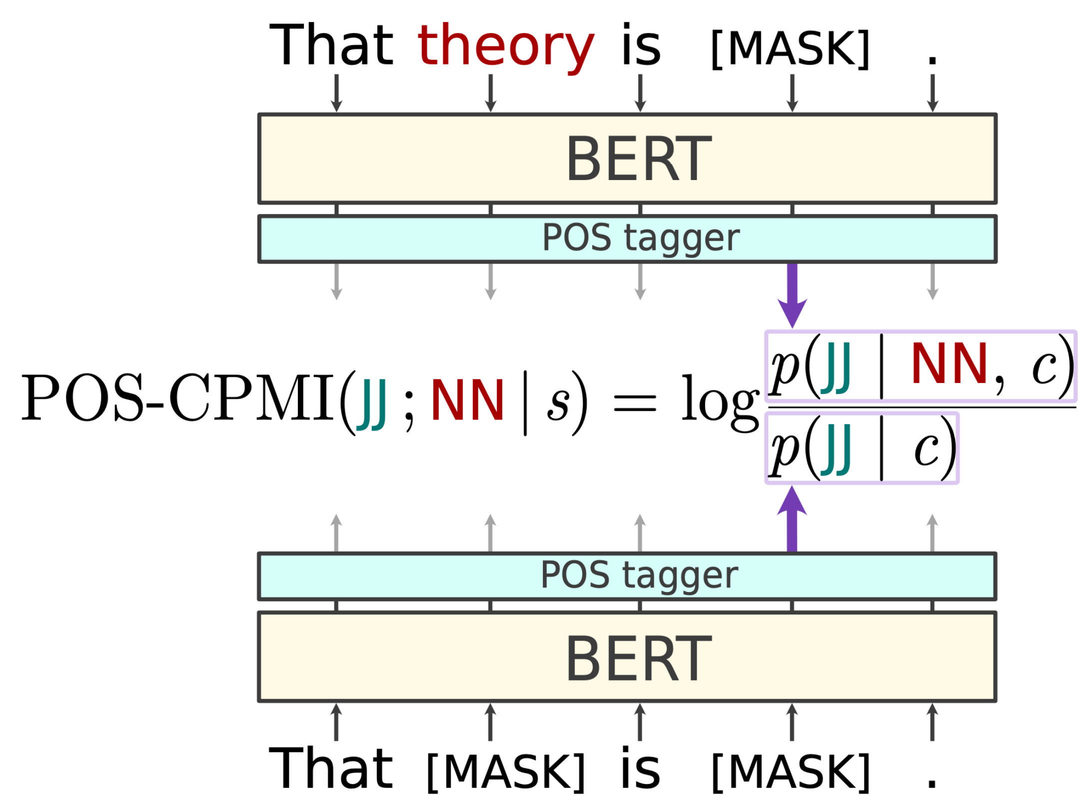
- Perhaps “*actual lexical items are too semantically charged to represent workable units of syntactic structure*” (Klein and Manning, 2004).
- Many statistical studies have used POS tags instead of words to deal with data sparsity issues (including Futrell et al 2019)



POS-CPMI-dependency parsing

a delexicalized version of CPMI

- Perhaps “*actual lexical items are too semantically charged to represent workable units of syntactic structure*” (Klein and Manning, 2004).
- Many statistical studies have used POS tags instead of words to deal with data sparsity issues (including Futrell et al 2019)
- We construct **POS-CPMI** a delexicalized version of CPMI based on probability estimates over POS tags rather than words.



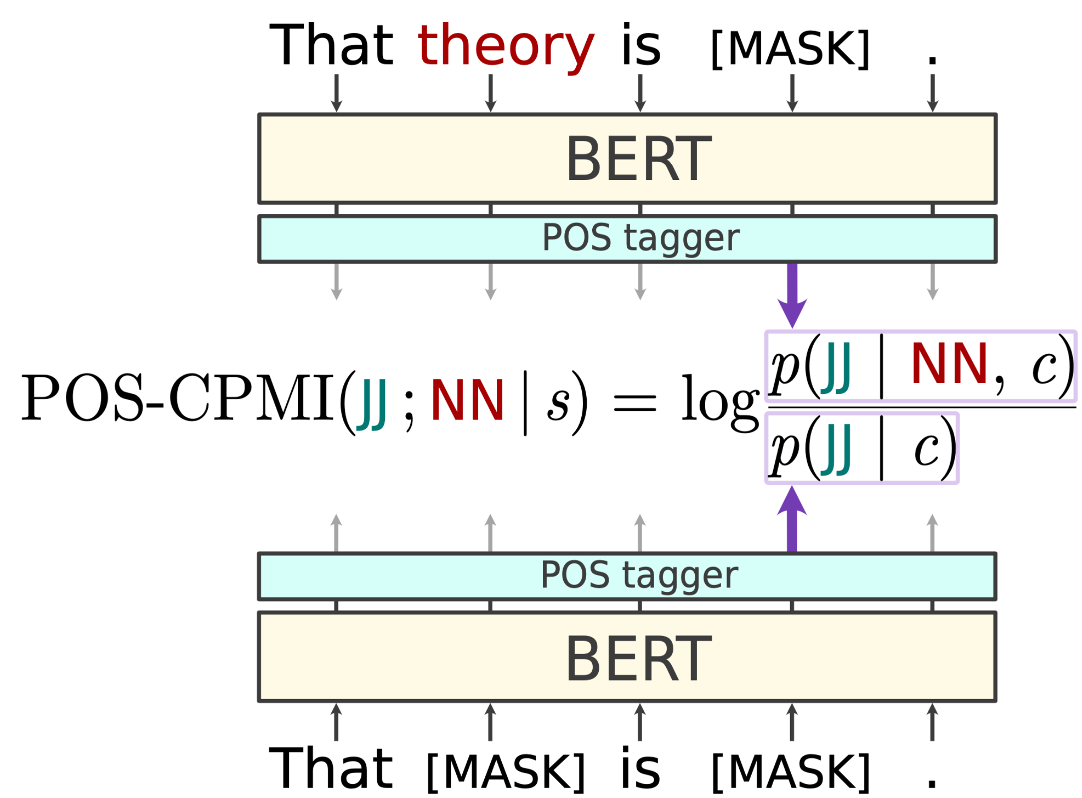
POS-CPMI-dependency parsing

a delexicalized version of CPMI

- Perhaps "*actual lexical items are too semantically charged to represent workable units of syntactic structure*" (Klein and Manning, 2004).
- Many statistical studies have used POS tags instead of words to deal with data sparsity issues (including Futrell et al 2019)
- We construct **POS-CPMI** a delexicalized version of CPMI based on probability estimates over POS tags rather than words.
- Results: POS-CPMI accuracy **no higher than CPMI (nor connect-adjacent)**

IB-POS	**BERT** base	.41
	BERT large	.41
	XLNet base	.40
	XLNet large	.36

Table 3: Total UUAS for POS-CPMI



$$\text{POS-CPMI}(\text{JJ}\,;\text{NN}\mid s) = \log \frac{p(\text{JJ}\mid \text{NN},\, c)}{p(\text{JJ}\mid c)}$$

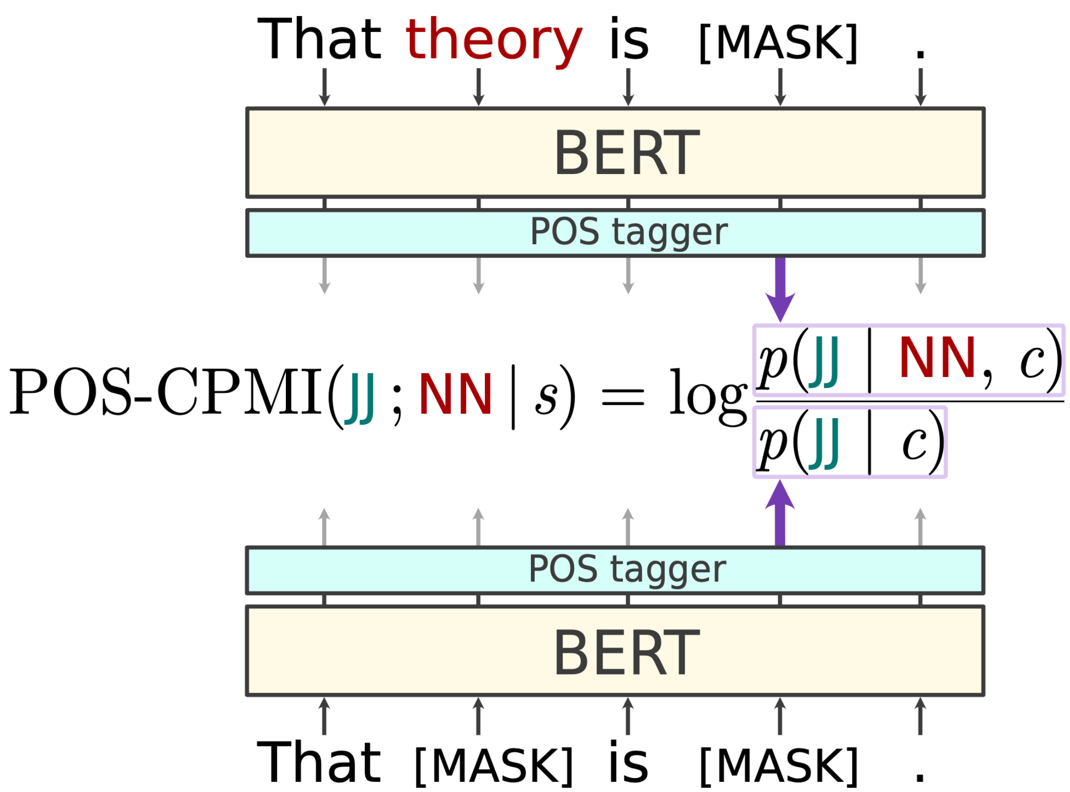
POS-CPMI-dependency parsing

a delexicalized version of CPMI

- Perhaps "*actual lexical items are too semantically charged to represent workable units of syntactic structure*" (Klein and Manning, 2004).
- Many statistical studies have used POS tags instead of words to deal with data sparsity issues (including Futrell et al 2019)
- We construct **POS-CPMI** a delexicalized version of CPMI based on probability estimates over POS tags rather than words.
- Results: POS-CPMI accuracy **no higher than CPMI (nor connect-adjacent)**

IB-POS	**BERT** base	.41
	BERT large	.41
	XLNet base	.40
	XLNet large	.36

Table 3: Total UUAS for POS-CPMI



CPMI-dependency parsing

more detailed analyses of large pretrained LM results

Looking more closely:

CPMI-dependency parsing

more detailed analyses of large pretrained LM results

Looking more closely:

- CPMI-dependencies overpredict connections between adjacent words (length = 1)
- especially BERT

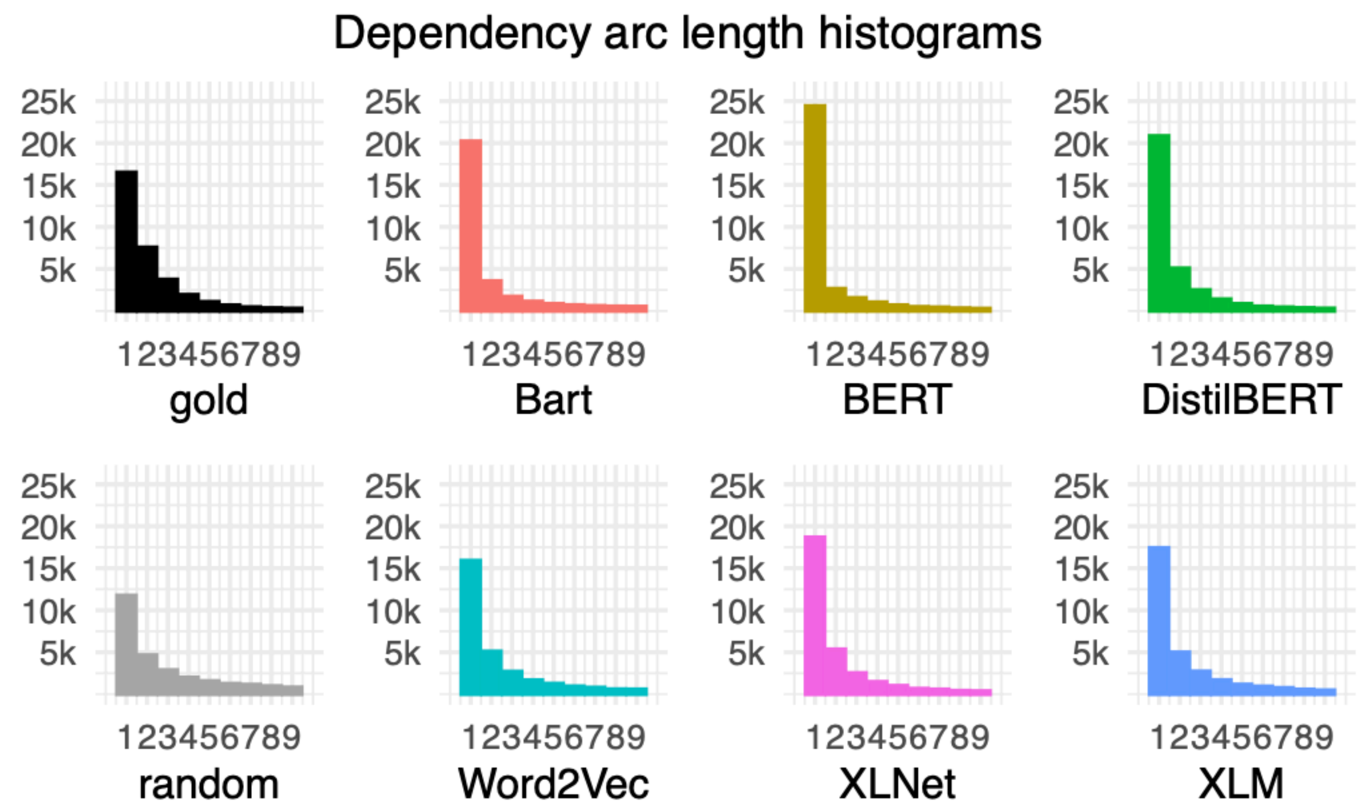


Figure 7: Histograms of arc length. Note, 49% of the gold arcs are length 1, whereas all of the CPMI dependencies had a higher proportion. BERT (base), in particular has 72%. For Word2Vec (which does not have access to word order), 47% are length 1. For the connect-adjacent baseline (not shown) the histogram is trivial: all arcs are length 1.

CPMI-dependency parsing

more detailed analyses of large pretrained LM results

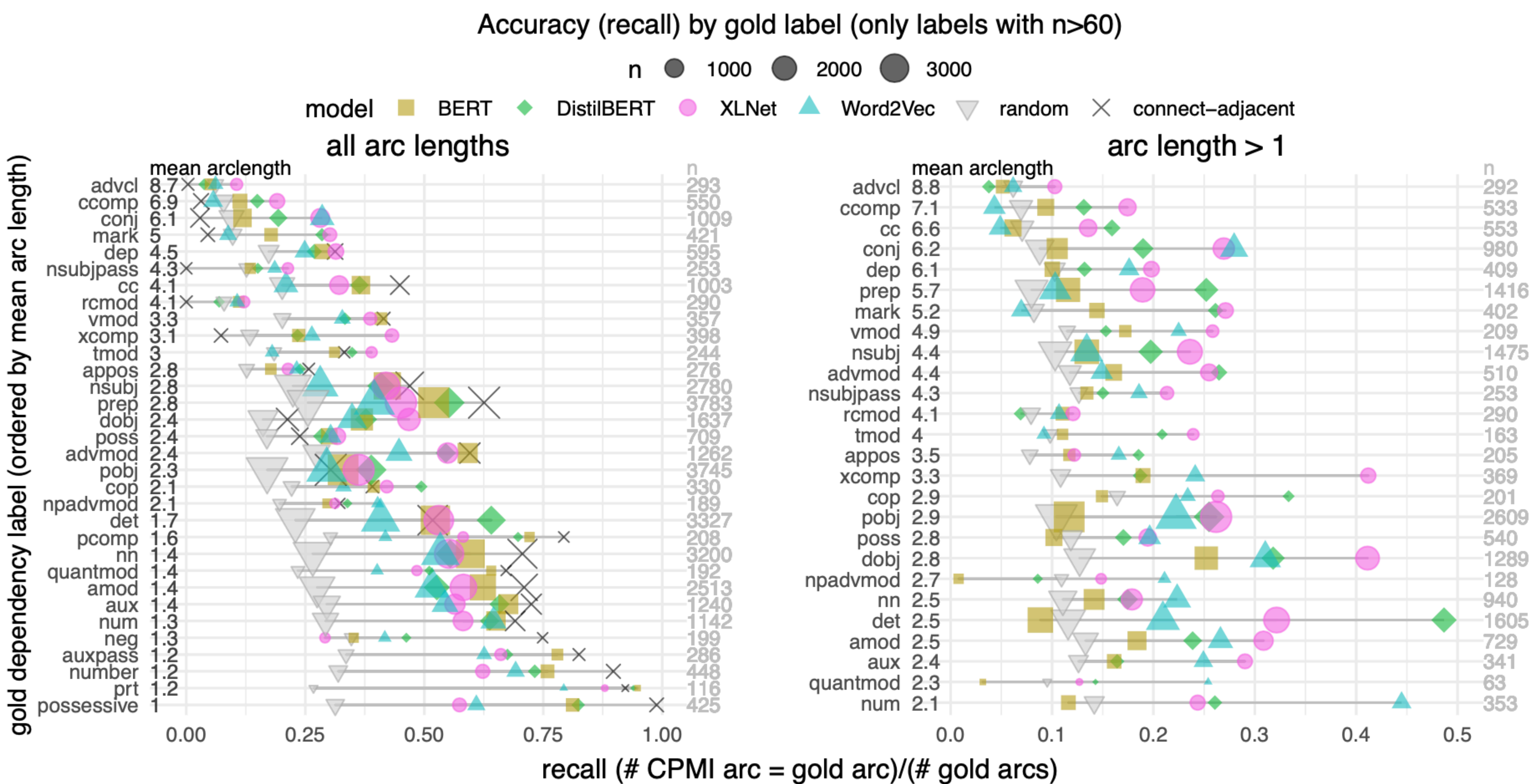
Looking more closely:

CPMI-dependency parsing

more detailed analyses of large pretrained LM results

Looking more closely:

- no relation has particularly high accuracy, beyond just connecting adjacent



Conclusion

Conclusion

What we did: We used large pretrained LMs to examine whether words that are *statistically* dependent are likely to be in a *linguistic* dependency relationship.

Conclusion

What we did: We used large pretrained LMs to examine whether words that are *statistically* dependent are likely to be in a *linguistic* dependency relationship.

Takeaways:

Conclusion

What we did: We used large pretrained LMs to examine whether words that are *statistically* dependent are likely to be in a *linguistic* dependency relationship.

Takeaways:

1. CPMI-dependency **accuracy only at most about as good as** a simple **connect-adjacent baseline**.

Conclusion

What we did: We used large pretrained LMs to examine whether words that are *statistically* dependent are likely to be in a *linguistic* dependency relationship.

Takeaways:

1. CPMI-dependency **accuracy only at most about as good as** a simple **connect-adjacent baseline**.
   - True **across languages**,

Conclusion

What we did: We used large pretrained LMs to examine whether words that are *statistically* dependent are likely to be in a *linguistic* dependency relationship.

Takeaways:

1. CPMI-dependency **accuracy only at most about as good as** a simple **connect-adjacent baseline**.
    - True **across languages**,
    - True for **syntactically-aware LMs**,

Conclusion

What we did: We used large pretrained LMs to examine whether words that are *statistically* dependent are likely to be in a *linguistic* dependency relationship.

Takeaways:

1. CPMI-dependency **accuracy only at most about as good as** a simple **connect-adjacent baseline**.
    - True **across languages**,
    - True for **syntactically-aware LMs**,
    - True about statistical dependencies **between POS tags** as well as wordforms

Conclusion

What we did: We used large pretrained LMs to examine whether words that are *statistically* dependent are likely to be in a *linguistic* dependency relationship.

Takeaways:

1. CPMI-dependency **accuracy only at most about as good as** a simple **connect-adjacent baseline**.
    - True **across languages**,
    - True for **syntactically-aware LMs**,
    - True about statistical dependencies **between POS tags** as well as wordforms
2. **statistical dependencies differ substantially between LMs**.

Conclusion

What we did: We used large pretrained LMs to examine whether words that are *statistically* dependent are likely to be in a *linguistic* dependency relationship.

Takeaways:

1. CPMI-dependency **accuracy only at most about as good as** a simple **connect-adjacent baseline**.
   - True **across languages**,
   - True for **syntactically-aware LMs**,
   - True about statistical dependencies **between POS tags** as well as wordforms
2. **statistical dependencies differ substantially between LMs**.
   - looking at differences in CPMI-dependencies can be a tool to understand these networks model statistical dependencies

Thank you!

paper: arxiv.org/abs/2104.08685
code: github.com/mcqll/cpmi-dependencies

References

Du, Wenyu, Zhouhan Lin, Yikang Shen, Timothy J. O'Donnell, Yoshua Bengio, and Yue Zhang. 2020. "Exploiting Syntactic Structure for Better Language Modeling: A Syntactic Distance Approach." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6611–28. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.591.

Futrell, Richard, Peng Qian, Edward Gibson, Evelina Fedorenko, and Idan Blank. 2019. "Syntactic Dependencies Correspond to Word Pairs with High Mutual Information." In P*roceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, 3–13. Paris, France: Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-7703.

Klein, Dan, and Christopher Manning. 2004. "Corpus-Based Induction of Syntactic Structure: Models of Dependency and Constituency." In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, 478–85. Barcelona, Spain. https://doi.org/10.3115/1218955.1219016.

Magerman, David M, and Mitchell P Marcus. 1990. "Parsing a Natural Language Using Mutual Information Statistics." In AAAI, 90:984–89. https://www.aaai.org/Library/AAAI/1990/aaai90-147.php.

de Paiva Alves, Eduardo. 1996. "The Selection of the Most Probable Dependency Structure in Japanese Using Mutual Information." In *34th Annual Meeting of the Association for Computational Linguistics*, 372–74. Santa Cruz, California, USA: Association for Computational Linguistics. https://doi.org/10.3115/981863.981919.

Zhang, Tianyi, and Tatsunori Hashimoto. 2021. "On the Inductive Bias of Masked Language Modeling: From Statistical to Syntactic Dependencies." In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. https://doi.org/10.18653/v1/2021.naacl-main.404.