Cognitively Viable Computational Models of Linguistic Knowledge

Deep Learning and the Nature of Linguistic Representation

Shalom Lappin

University of Gothenburg, Queen Mary University of London, and

King's College London

WeSSLLI 2020, Brandeis University

July 17, 2020

Outline

How Useful are Linguistic Theories for NLP Applications

Machine Learning Models vs Formal Grammar

Explaining Language Acquisition

Conclusions and Future Work

Linguistic Theories and NLP Tasks

- Syntactic and semantic theories offer formal representations of linguistic structure and interpretation, respectively.
- They aim to express central properties of form and meaning that humans make use of in interpreting the sentences of their languages.
- If these theories are formally explicit, it is possible to incorporate their principles into computational models of sentence processing.
- It is reasonable to expect that, to the extent that such theories succeed in capturing core properties of natural language, linguistically informed computational models will show better performance across relevant NLP tasks than systems which do not make use of these theories.

Linguistic Theories and NLP Tasks

- Syntactic and semantic theories offer formal representations of linguistic structure and interpretation, respectively.
- They aim to express central properties of form and meaning that humans make use of in interpreting the sentences of their languages.
- If these theories are formally explicit, it is possible to incorporate their principles into computational models of sentence processing.
- It is reasonable to expect that, to the extent that such theories succeed in capturing core properties of natural language, linguistically informed computational models will show better performance across relevant NLP tasks than systems which do not make use of these theories.

Linguistic Theories and NLP Tasks

- Syntactic and semantic theories offer formal representations of linguistic structure and interpretation, respectively.
- They aim to express central properties of form and meaning that humans make use of in interpreting the sentences of their languages.
- If these theories are formally explicit, it is possible to incorporate their principles into computational models of sentence processing.
- It is reasonable to expect that, to the extent that such theories succeed in capturing core properties of natural language, linguistically informed computational models will show better performance across relevant NLP tasks than systems which do not make use of these theories.

Linguistic Theories and NLP Tasks

- Syntactic and semantic theories offer formal representations of linguistic structure and interpretation, respectively.
- They aim to express central properties of form and meaning that humans make use of in interpreting the sentences of their languages.
- If these theories are formally explicit, it is possible to incorporate their principles into computational models of sentence processing.
- It is reasonable to expect that, to the extent that such theories succeed in capturing core properties of natural language, linguistically informed computational models will show better performance across relevant NLP tasks than systems which do not make use of these theories.

Tree-LSTMs

- Tai et al. (2015) construct two Tree-LSTM models, one of which encodes dependency trees, and the other constituency trees.
- Each model produces its hidden state from an input vector and a set of hidden states corresponding to children of the tree node being processed at that point.
- They apply each model to two tasks: sentiment classification and semantic relatedness.
- Tai et al. claim that one of these models outperforms non-tree DNN baselines on (versions) of both tasks.

Tree-LSTMs

- Tai et al. (2015) construct two Tree-LSTM models, one of which encodes dependency trees, and the other constituency trees.
- Each model produces its hidden state from an input vector and a set of hidden states corresponding to children of the tree node being processed at that point.
- They apply each model to two tasks: sentiment classification and semantic relatedness.
- Tai et al. claim that one of these models outperforms non-tree DNN baselines on (versions) of both tasks.

Tree-LSTMs

- Tai et al. (2015) construct two Tree-LSTM models, one of which encodes dependency trees, and the other constituency trees.
- Each model produces its hidden state from an input vector and a set of hidden states corresponding to children of the tree node being processed at that point.
- They apply each model to two tasks: sentiment classification and semantic relatedness.
- Tai et al. claim that one of these models outperforms non-tree DNN baselines on (versions) of both tasks.

Tree-LSTMs

- Tai et al. (2015) construct two Tree-LSTM models, one of which encodes dependency trees, and the other constituency trees.
- Each model produces its hidden state from an input vector and a set of hidden states corresponding to children of the tree node being processed at that point.
- They apply each model to two tasks: sentiment classification and semantic relatedness.
- Tai et al. claim that one of these models outperforms non-tree DNN baselines on (versions) of both tasks.

Sentiment Classification Results

- Tai et al. train their models on the Stanford Sentiment Treebank (Socher et al., 2013), and test them on a subset of this corpus for both five category and binary classification.
- Their best model, the Constituency Tree-LSTM (with tuned Glove vectors) outperforms the base line systems on the five category task, with a score of 51.0 accuracy, and it achieves 88.0 on the binary task.
- The best non-tree LSTM, a Bidirectional LSTM, achieves 49.1 on the five category task, and 87.5 on the binary.
- Moreover, a (non-tree) multi-channel CNN (Kim, 2014) outperforms both Tree-LSTMs on the binary task, with an accuracy score of 88.1.

Sentiment Classification Results

- Tai et al. train their models on the Stanford Sentiment Treebank (Socher et al., 2013), and test them on a subset of this corpus for both five category and binary classification.
- Their best model, the Constituency Tree-LSTM (with tuned Glove vectors) outperforms the base line systems on the five category task, with a score of 51.0 accuracy, and it achieves 88.0 on the binary task.
- The best non-tree LSTM, a Bidirectional LSTM, achieves 49.1 on the five category task, and 87.5 on the binary.
- Moreover, a (non-tree) multi-channel CNN (Kim, 2014) outperforms both Tree-LSTMs on the binary task, with an accuracy score of 88.1.

Sentiment Classification Results

- Tai et al. train their models on the Stanford Sentiment Treebank (Socher et al., 2013), and test them on a subset of this corpus for both five category and binary classification.
- Their best model, the Constituency Tree-LSTM (with tuned Glove vectors) outperforms the base line systems on the five category task, with a score of 51.0 accuracy, and it achieves 88.0 on the binary task.
- The best non-tree LSTM, a Bidirectional LSTM, achieves 49.1 on the five category task, and 87.5 on the binary.
- Moreover, a (non-tree) multi-channel CNN (Kim, 2014) outperforms both Tree-LSTMs on the binary task, with an accuracy score of 88.1.

Sentiment Classification Results

- Tai et al. train their models on the Stanford Sentiment Treebank (Socher et al., 2013), and test them on a subset of this corpus for both five category and binary classification.
- Their best model, the Constituency Tree-LSTM (with tuned Glove vectors) outperforms the base line systems on the five category task, with a score of 51.0 accuracy, and it achieves 88.0 on the binary task.
- The best non-tree LSTM, a Bidirectional LSTM, achieves 49.1 on the five category task, and 87.5 on the binary.
- Moreover, a (non-tree) multi-channel CNN (Kim, 2014) outperforms both Tree-LSTMs on the binary task, with an accuracy score of 88.1.

Semantic Relatedness Results

- For semantic relatedness Tai et al. train and test their models on the SICK data set (Marelli et al., 2014), which is annotated by human evaluators.
- Their Dependency Tree-LSTM scores highest on this task, with a Pearson correlation of 0.8676, and a Spearman correlation of 0.8083.
- All of the non-Tree LSTMs are above 0.85 on the Pearson metric, and three of the four are above 0.79 on the Spearman (the fourth is at 0.7896).
- The difference in performance between the Tree- and Non-Tree LSTMs on both tasks is marginal.
- It is not clear to what extent encoding tree structure in an LSTM improves its handling of either sentiment classification or semantic relatedness.

Semantic Relatedness Results

- For semantic relatedness Tai et al. train and test their models on the SICK data set (Marelli et al., 2014), which is annotated by human evaluators.
- Their Dependency Tree-LSTM scores highest on this task, with a Pearson correlation of 0.8676, and a Spearman correlation of 0.8083.
- All of the non-Tree LSTMs are above 0.85 on the Pearson metric, and three of the four are above 0.79 on the Spearman (the fourth is at 0.7896).
- The difference in performance between the Tree- and Non-Tree LSTMs on both tasks is marginal.
- It is not clear to what extent encoding tree structure in an LSTM improves its handling of either sentiment classification or semantic relatedness.

Semantic Relatedness Results

- For semantic relatedness Tai et al. train and test their models on the SICK data set (Marelli et al., 2014), which is annotated by human evaluators.
- Their Dependency Tree-LSTM scores highest on this task, with a Pearson correlation of 0.8676, and a Spearman correlation of 0.8083.
- All of the non-Tree LSTMs are above 0.85 on the Pearson metric, and three of the four are above 0.79 on the Spearman (the fourth is at 0.7896).
- The difference in performance between the Tree- and Non-Tree LSTMs on both tasks is marginal.
- It is not clear to what extent encoding tree structure in an LSTM improves its handling of either sentiment classification or semantic relatedness.

Semantic Relatedness Results

- For semantic relatedness Tai et al. train and test their models on the SICK data set (Marelli et al., 2014), which is annotated by human evaluators.
- Their Dependency Tree-LSTM scores highest on this task, with a Pearson correlation of 0.8676, and a Spearman correlation of 0.8083.
- All of the non-Tree LSTMs are above 0.85 on the Pearson metric, and three of the four are above 0.79 on the Spearman (the fourth is at 0.7896).
- The difference in performance between the Tree- and Non-Tree LSTMs on both tasks is marginal.
- It is not clear to what extent encoding tree structure in an LSTM improves its handling of either sentiment classification or semantic relatedness.

Semantic Relatedness Results

- For semantic relatedness Tai et al. train and test their models on the SICK data set (Marelli et al., 2014), which is annotated by human evaluators.
- Their Dependency Tree-LSTM scores highest on this task, with a Pearson correlation of 0.8676, and a Spearman correlation of 0.8083.
- All of the non-Tree LSTMs are above 0.85 on the Pearson metric, and three of the four are above 0.79 on the Spearman (the fourth is at 0.7896).
- The difference in performance between the Tree- and Non-Tree LSTMs on both tasks is marginal.
- It is not clear to what extent encoding tree structure in an LSTM improves its handling of either sentiment classification or semantic relatedness.

More Recent Enriched Models

- As we saw in class 2, BL find that training an LSTM on an impoverished lexicon in which POS tags highlight the subject-verb agreement relation, degrades the performance of the model relative to one trained on a richer lexicon.
- We also observed that, as Williams et al. (2018) show, Choi et al.'s (2018) latent tree RNN outperforms other systems on two NL inference tasks, but the scores of the non-tree LSTM are not far from those of this model.
- Latent tree RNNs generate shallow, parses, which are not consistent, and do not correlate with theoretically motivated constituency structures.
- In class 3 we saw that enriching an LSTM with syntactic or semantic tags, or full dependency trees, degrades its performance on the sentence acceptability task (EBL).

More Recent Enriched Models

- As we saw in class 2, BL find that training an LSTM on an impoverished lexicon in which POS tags highlight the subject-verb agreement relation, degrades the performance of the model relative to one trained on a richer lexicon.
- We also observed that, as Williams et al. (2018) show, Choi et al.'s (2018) latent tree RNN outperforms other systems on two NL inference tasks, but the scores of the non-tree LSTM are not far from those of this model.
- Latent tree RNNs generate shallow, parses, which are not consistent, and do not correlate with theoretically motivated constituency structures.
- In class 3 we saw that enriching an LSTM with syntactic or semantic tags, or full dependency trees, degrades its performance on the sentence acceptability task (EBL).

More Recent Enriched Models

- As we saw in class 2, BL find that training an LSTM on an impoverished lexicon in which POS tags highlight the subject-verb agreement relation, degrades the performance of the model relative to one trained on a richer lexicon.
- We also observed that, as Williams et al. (2018) show, Choi et al.'s (2018) latent tree RNN outperforms other systems on two NL inference tasks, but the scores of the non-tree LSTM are not far from those of this model.
- Latent tree RNNs generate shallow, parses, which are not consistent, and do not correlate with theoretically motivated constituency structures.
- In class 3 we saw that enriching an LSTM with syntactic or semantic tags, or full dependency trees, degrades its performance on the sentence acceptability task (EBL).

More Recent Enriched Models

- As we saw in class 2, BL find that training an LSTM on an impoverished lexicon in which POS tags highlight the subject-verb agreement relation, degrades the performance of the model relative to one trained on a richer lexicon.
- We also observed that, as Williams et al. (2018) show, Choi et al.'s (2018) latent tree RNN outperforms other systems on two NL inference tasks, but the scores of the non-tree LSTM are not far from those of this model.
- Latent tree RNNs generate shallow, parses, which are not consistent, and do not correlate with theoretically motivated constituency structures.
- In class 3 we saw that enriching an LSTM with syntactic or semantic tags, or full dependency trees, degrades its performance on the sentence acceptability task (EBL).

Tree-RNNs Applied to Syntactic Tasks

- McCoy et al. (2020) compare RNNs, LSTMs, and GRUs which incorporate parse tree structure in their respective architectures, with sequential non-tree versions of these DNNs, on two syntactic tasks.
- For question formation they test which of these models identifies moving the main auxiliary verb, rather than moving the first auxiliary in a sequence, in a generalisation test set.

  1a. Don't my yaks that do read giggle?
  b. *Do my yaks that read don't giggle?

- For agreement they test the models on the subject vs the most recent NP as controller of the main verb.

  2a. My zebra by the yaks swims.
  b. *My zebra by the yaks swim.

Tree-RNNs Applied to Syntactic Tasks

- McCoy et al. (2020) compare RNNs, LSTMs, and GRUs which incorporate parse tree structure in their respective architectures, with sequential non-tree versions of these DNNs, on two syntactic tasks.
- For question formation they test which of these models identifies moving the main auxiliary verb, rather than moving the first auxiliary in a sequence, in a generalisation test set.

  1a. Don't my yaks that do read giggle?
  b. *Do my yaks that read don't giggle?

- For agreement they test the models on the subject vs the most recent NP as controller of the main verb.

  2a. My zebra by the yaks swims.
  b. *My zebra by the yaks swim.

Tree-RNNs Applied to Syntactic Tasks

- McCoy et al. (2020) compare RNNs, LSTMs, and GRUs which incorporate parse tree structure in their respective architectures, with sequential non-tree versions of these DNNs, on two syntactic tasks.
- For question formation they test which of these models identifies moving the main auxiliary verb, rather than moving the first auxiliary in a sequence, in a generalisation test set.

  1a. Don't my yaks that do read giggle?
  b. *Do my yaks that read don't giggle?

- For agreement they test the models on the subject vs the most recent NP as controller of the main verb.

  2a. My zebra by the yaks swims.
  b. *My zebra by the yaks swim.

Impoverished vs Rich Training Data

- McCoy et al. report that when they train their DNNs on data that contain only ambiguous examples compatible with either rule, the tree-DNNs generalise correctly to unambiguous cases, but the sequential DNNs do not.
- However, when unambiguous instances of an operation are included in the training data, the sequential and the tree-DNNs perform comparably, with both achieving over 90% accuracy
- These results are hardly surprising, given that the tree models incorporate the parse structure in their architecture, and the non-tree models can only learn the correct forms if they are exposed to them in training.
- It is unclear that these experiments have any consequences for language acquisition, given that there is substantial evidence that human learners are exposed to significant amounts of disambiguating data and reinforcement correction (see Clark and Lappin, 2011 for discussion and references).

Impoverished vs Rich Training Data

- McCoy et al. report that when they train their DNNs on data that contain only ambiguous examples compatible with either rule, the tree-DNNs generalise correctly to unambiguous cases, but the sequential DNNs do not.
- However, when unambiguous instances of an operation are included in the training data, the sequential and the tree-DNNs perform comparably, with both achieving over 90% accuracy
- These results are hardly surprising, given that the tree models incorporate the parse structure in their architecture, and the non-tree models can only learn the correct forms if they are exposed to them in training.
- It is unclear that these experiments have any consequences for language acquisition, given that there is substantial evidence that human learners are exposed to significant amounts of disambiguating data and reinforcement correction (see Clark and Lappin, 2011 for discussion and references).

Impoverished vs Rich Training Data

- McCoy et al. report that when they train their DNNs on data that contain only ambiguous examples compatible with either rule, the tree-DNNs generalise correctly to unambiguous cases, but the sequential DNNs do not.
- However, when unambiguous instances of an operation are included in the training data, the sequential and the tree-DNNs perform comparably, with both achieving over 90% accuracy
- These results are hardly surprising, given that the tree models incorporate the parse structure in their architecture, and the non-tree models can only learn the correct forms if they are exposed to them in training.
- It is unclear that these experiments have any consequences for language acquisition, given that there is substantial evidence that human learners are exposed to significant amounts of disambiguating data and reinforcement correction (see Clark and Lappin, 2011 for discussion and references).

Impoverished vs Rich Training Data

- McCoy et al. report that when they train their DNNs on data that contain only ambiguous examples compatible with either rule, the tree-DNNs generalise correctly to unambiguous cases, but the sequential DNNs do not.
- However, when unambiguous instances of an operation are included in the training data, the sequential and the tree-DNNs perform comparably, with both achieving over 90% accuracy
- These results are hardly surprising, given that the tree models incorporate the parse structure in their architecture, and the non-tree models can only learn the correct forms if they are exposed to them in training.
- It is unclear that these experiments have any consequences for language acquisition, given that there is substantial evidence that human learners are exposed to significant amounts of disambiguating data and reinforcement correction (see Clark and Lappin, 2011 for discussion and references).

Transformers and Implicit Trees

- In class 2 we discussed HM's supervised squared L2 distance and depth probe for dependency tree structure in DNN sentence vectors.
- HM report that BERT and ELMO, but not their baseline systems, predict unlabelled dependency parse trees, with a high degree of accuracy.
- This result suggests that these models encode parse structure information in the distributed representations of their lexical embeddings.
- However, it is not clear that these structures are unambiguously present in the models' output sentence vectors.
- It is, at least in principle, possible that a different supervised probe would reveal entirely distinct parse structures in these vectors.

Transformers and Implicit Trees

- In class 2 we discussed HM's supervised squared L2 distance and depth probe for dependency tree structure in DNN sentence vectors.
- HM report that BERT and ELMO, but not their baseline systems, predict unlabelled dependency parse trees, with a high degree of accuracy.
- This result suggests that these models encode parse structure information in the distributed representations of their lexical embeddings.
- However, it is not clear that these structures are unambiguously present in the models' output sentence vectors.
- It is, at least in principle, possible that a different supervised probe would reveal entirely distinct parse structures in these vectors.

Transformers and Implicit Trees

- In class 2 we discussed HM's supervised squared L2 distance and depth probe for dependency tree structure in DNN sentence vectors.
- HM report that BERT and ELMO, but not their baseline systems, predict unlabelled dependency parse trees, with a high degree of accuracy.
- This result suggests that these models encode parse structure information in the distributed representations of their lexical embeddings.
- However, it is not clear that these structures are unambiguously present in the models' output sentence vectors.
- It is, at least in principle, possible that a different supervised probe would reveal entirely distinct parse structures in these vectors.

Transformers and Implicit Trees

- In class 2 we discussed HM's supervised squared L2 distance and depth probe for dependency tree structure in DNN sentence vectors.
- HM report that BERT and ELMO, but not their baseline systems, predict unlabelled dependency parse trees, with a high degree of accuracy.
- This result suggests that these models encode parse structure information in the distributed representations of their lexical embeddings.
- However, it is not clear that these structures are unambiguously present in the models' output sentence vectors.
- It is, at least in principle, possible that a different supervised probe would reveal entirely distinct parse structures in these vectors.

Transformers and Implicit Trees

- In class 2 we discussed HM's supervised squared L2 distance and depth probe for dependency tree structure in DNN sentence vectors.
- HM report that BERT and ELMO, but not their baseline systems, predict unlabelled dependency parse trees, with a high degree of accuracy.
- This result suggests that these models encode parse structure information in the distributed representations of their lexical embeddings.
- However, it is not clear that these structures are unambiguously present in the models' output sentence vectors.
- It is, at least in principle, possible that a different supervised probe would reveal entirely distinct parse structures in these vectors.

A Syntactic Knowledge Distilled BERT

- Kuncoro et al. (2020) adapt the technique of knowledge distilling that Kuncoro et al. (2019) use for LSTMs, to impart an RNNG induced syntactic bias to BERT.
- They employ right to left, and left to right RNNG LMs to estimate the trees, and word probabilities, in the right and left contexts of BERT's training data.
- Kuncoro et al. then use the combined RNNG LM to supervise the predictions of BERT.
- They apply both the syntactic knowledge distilled version of BERT (KD BERT) and non-distilled BERT to six NLP tasks requiring syntactic information, and to the GLUE benchmark (Wang et al., 2018), a suite of eight natural language understanding tasks.

A Syntactic Knowledge Distilled BERT

- Kuncoro et al. (2020) adapt the technique of knowledge distilling that Kuncoro et al. (2019) use for LSTMs, to impart an RNNG induced syntactic bias to BERT.
- They employ right to left, and left to right RNNG LMs to estimate the trees, and word probabilities, in the right and left contexts of BERT's training data.
- Kuncoro et al. then use the combined RNNG LM to supervise the predictions of BERT.
- They apply both the syntactic knowledge distilled version of BERT (KD BERT) and non-distilled BERT to six NLP tasks requiring syntactic information, and to the GLUE benchmark (Wang et al., 2018), a suite of eight natural language understanding tasks.

A Syntactic Knowledge Distilled BERT

- Kuncoro et al. (2020) adapt the technique of knowledge distilling that Kuncoro et al. (2019) use for LSTMs, to impart an RNNG induced syntactic bias to BERT.
- They employ right to left, and left to right RNNG LMs to estimate the trees, and word probabilities, in the right and left contexts of BERT's training data.
- Kuncoro et al. then use the combined RNNG LM to supervise the predictions of BERT.
- They apply both the syntactic knowledge distilled version of BERT (KD BERT) and non-distilled BERT to six NLP tasks requiring syntactic information, and to the GLUE benchmark (Wang et al., 2018), a suite of eight natural language understanding tasks.

A Syntactic Knowledge Distilled BERT

- Kuncoro et al. (2020) adapt the technique of knowledge distilling that Kuncoro et al. (2019) use for LSTMs, to impart an RNNG induced syntactic bias to BERT.
- They employ right to left, and left to right RNNG LMs to estimate the trees, and word probabilities, in the right and left contexts of BERT's training data.
- Kuncoro et al. then use the combined RNNG LM to supervise the predictions of BERT.
- They apply both the syntactic knowledge distilled version of BERT (KD BERT) and non-distilled BERT to six NLP tasks requiring syntactic information, and to the GLUE benchmark (Wang et al., 2018), a suite of eight natural language understanding tasks.

KD vs Non-KD BERT

- In five of the six tasks that Kuncoro et al. test, KD BERT outperforms non-KD BERT by less than 1%.
- In the sixth task, CCG super tagging, it scores 1.32% higher.
- Non-KD BERT narrowly outperforms KD BERT on an average of eight tasks for GLUE, with 80.3% to 80%.
- These results are consistent with the pattern that we observed in our discussion in Class 2 of Kuncoro et al. (2019)'s Syntax-Aware LSTM.
- Despite Kuncoro et al.'s claim that induced syntactic bias improves the performance of DNN LMs, their results across a wide range of NLP tasks suggest that this bias does not significantly increase accuracy.

KD vs Non-KD BERT

- In five of the six tasks that Kuncoro et al. test, KD BERT outperforms non-KD BERT by less than 1%.
- In the sixth task, CCG super tagging, it scores 1.32% higher.
- Non-KD BERT narrowly outperforms KD BERT on an average of eight tasks for GLUE, with 80.3% to 80%.
- These results are consistent with the pattern that we observed in our discussion in Class 2 of Kuncoro et al. (2019)'s Syntax-Aware LSTM.
- Despite Kuncoro et al.'s claim that induced syntactic bias improves the performance of DNN LMs, their results across a wide range of NLP tasks suggest that this bias does not significantly increase accuracy.

KD vs Non-KD BERT

- In five of the six tasks that Kuncoro et al. test, KD BERT outperforms non-KD BERT by less than 1%.
- In the sixth task, CCG super tagging, it scores 1.32% higher.
- Non-KD BERT narrowly outperforms KD BERT on an average of eight tasks for GLUE, with 80.3% to 80%.
- These results are consistent with the pattern that we observed in our discussion in Class 2 of Kuncoro et al. (2019)'s Syntax-Aware LSTM.
- Despite Kuncoro et al.'s claim that induced syntactic bias improves the performance of DNN LMs, their results across a wide range of NLP tasks suggest that this bias does not significantly increase accuracy.

KD vs Non-KD BERT

- In five of the six tasks that Kuncoro et al. test, KD BERT outperforms non-KD BERT by less than 1%.
- In the sixth task, CCG super tagging, it scores 1.32% higher.
- Non-KD BERT narrowly outperforms KD BERT on an average of eight tasks for GLUE, with 80.3% to 80%.
- These results are consistent with the pattern that we observed in our discussion in Class 2 of Kuncoro et al. (2019)'s Syntax-Aware LSTM.
- Despite Kuncoro et al.'s claim that induced syntactic bias improves the performance of DNN LMs, their results across a wide range of NLP tasks suggest that this bias does not significantly increase accuracy.

KD vs Non-KD BERT

- In five of the six tasks that Kuncoro et al. test, KD BERT outperforms non-KD BERT by less than 1%.
- In the sixth task, CCG super tagging, it scores 1.32% higher.
- Non-KD BERT narrowly outperforms KD BERT on an average of eight tasks for GLUE, with 80.3% to 80%.
- These results are consistent with the pattern that we observed in our discussion in Class 2 of Kuncoro et al. (2019)'s Syntax-Aware LSTM.
- Despite Kuncoro et al.'s claim that induced syntactic bias improves the performance of DNN LMs, their results across a wide range of NLP tasks suggest that this bias does not significantly increase accuracy.

Gradience in Linguistic Judgments

- In classes 3 and 4 we saw that human sentence acceptability judgments are consistently gradient, both at the individual and the aggregate level (LCL, BLL, LALPS).
- We use acceptability rather than grammaticality in our crowdsource experiments, because the former is directly observable.
- By contrast, grammaticality is a theoretical property, and so it is not directly accessible .
- We seek to avoid biasing the judgments of human annotators with theoretical or prescriptive commitments that they may hold.
- Theoretical linguists have generally appealed to speakers' acceptability intuitions to motivate their claims.

Gradience in Linguistic Judgments

- In classes 3 and 4 we saw that human sentence acceptability judgments are consistently gradient, both at the individual and the aggregate level (LCL, BLL, LALPS).
- We use acceptability rather than grammaticality in our crowdsource experiments, because the former is directly observable.
- By contrast, grammaticality is a theoretical property, and so it is not directly accessible .
- We seek to avoid biasing the judgments of human annotators with theoretical or prescriptive commitments that they may hold.
- Theoretical linguists have generally appealed to speakers' acceptability intuitions to motivate their claims.

Gradience in Linguistic Judgments

- In classes 3 and 4 we saw that human sentence acceptability judgments are consistently gradient, both at the individual and the aggregate level (LCL, BLL, LALPS).
- We use acceptability rather than grammaticality in our crowdsource experiments, because the former is directly observable.
- By contrast, grammaticality is a theoretical property, and so it is not directly accessible .
- We seek to avoid biasing the judgments of human annotators with theoretical or prescriptive commitments that they may hold.
- Theoretical linguists have generally appealed to speakers' acceptability intuitions to motivate their claims.

Gradience in Linguistic Judgments

- In classes 3 and 4 we saw that human sentence acceptability judgments are consistently gradient, both at the individual and the aggregate level (LCL, BLL, LALPS).
- We use acceptability rather than grammaticality in our crowdsource experiments, because the former is directly observable.
- By contrast, grammaticality is a theoretical property, and so it is not directly accessible .
- We seek to avoid biasing the judgments of human annotators with theoretical or prescriptive commitments that they may hold.
- Theoretical linguists have generally appealed to speakers' acceptability intuitions to motivate their claims.

Gradience in Linguistic Judgments

- In classes 3 and 4 we saw that human sentence acceptability judgments are consistently gradient, both at the individual and the aggregate level (LCL, BLL, LALPS).
- We use acceptability rather than grammaticality in our crowdsource experiments, because the former is directly observable.
- By contrast, grammaticality is a theoretical property, and so it is not directly accessible .
- We seek to avoid biasing the judgments of human annotators with theoretical or prescriptive commitments that they may hold.
- Theoretical linguists have generally appealed to speakers' acceptability intuitions to motivate their claims.

Predicting Sentence Acceptability

- We also saw that deep neural language models achieve encouraging results in predicting mean human acceptability ratings, both in and out of document context.
- Bidirectional transformers approach estimated human performance for this task on test sets derived by round trip MT on Wikipedia text (LALPS).
- They surpass human performance for this task when performance is estimated by a one-vs-rest metric unflitered for outlier judgments
- The bidirectional transformer models also perform robustly on out of domain prediction of mean human judgments for linguists' example test sets.

Predicting Sentence Acceptability

- We also saw that deep neural language models achieve encouraging results in predicting mean human acceptability ratings, both in and out of document context.
- Bidirectional transformers approach estimated human performance for this task on test sets derived by round trip MT on Wikipedia text (LALPS).
- They surpass human performance for this task when performance is estimated by a one-vs-rest metric unflitered for outlier judgments
- The bidirectional transformer models also perform robustly on out of domain prediction of mean human judgments for linguists' example test sets.

Predicting Sentence Acceptability

- We also saw that deep neural language models achieve encouraging results in predicting mean human acceptability ratings, both in and out of document context.
- Bidirectional transformers approach estimated human performance for this task on test sets derived by round trip MT on Wikipedia text (LALPS).
- They surpass human performance for this task when performance is estimated by a one-vs-rest metric unflitered for outlier judgments
- The bidirectional transformer models also perform robustly on out of domain prediction of mean human judgments for linguists' example test sets.

Predicting Sentence Acceptability

- We also saw that deep neural language models achieve encouraging results in predicting mean human acceptability ratings, both in and out of document context.
- Bidirectional transformers approach estimated human performance for this task on test sets derived by round trip MT on Wikipedia text (LALPS).
- They surpass human performance for this task when performance is estimated by a one-vs-rest metric unflitered for outlier judgments
- The bidirectional transformer models also perform robustly on out of domain prediction of mean human judgments for linguists' example test sets.

Formal Grammars and Gradience

- Classical formal grammars specify recursive definitions of the set of well-formed sentences in a language.
- They are binary decision procedures for membership in this class, and so they cannot, in themselves, accommodate gradience.
- Classical binary theories of grammar must consign gradience to external processing and performance factors.
- This approach is, in principle, plausible, but it must formulate a precise, integrated theory of grammar and processing that predicts the observed phenomenon in detail, to have any explanatory content.
- To date, no such account has been forthcoming.

Formal Grammars and Gradience

- Classical formal grammars specify recursive definitions of the set of well-formed sentences in a language.
- They are binary decision procedures for membership in this class, and so they cannot, in themselves, accommodate gradience.
- Classical binary theories of grammar must consign gradience to external processing and performance factors.
- This approach is, in principle, plausible, but it must formulate a precise, integrated theory of grammar and processing that predicts the observed phenomenon in detail, to have any explanatory content.
- To date, no such account has been forthcoming.

Formal Grammars and Gradience

- Classical formal grammars specify recursive definitions of the set of well-formed sentences in a language.
- They are binary decision procedures for membership in this class, and so they cannot, in themselves, accommodate gradience.
- Classical binary theories of grammar must consign gradience to external processing and performance factors.
- This approach is, in principle, plausible, but it must formulate a precise, integrated theory of grammar and processing that predicts the observed phenomenon in detail, to have any explanatory content.
- To date, no such account has been forthcoming.

Formal Grammars and Gradience

- Classical formal grammars specify recursive definitions of the set of well-formed sentences in a language.
- They are binary decision procedures for membership in this class, and so they cannot, in themselves, accommodate gradience.
- Classical binary theories of grammar must consign gradience to external processing and performance factors.
- This approach is, in principle, plausible, but it must formulate a precise, integrated theory of grammar and processing that predicts the observed phenomenon in detail, to have any explanatory content.
- To date, no such account has been forthcoming.

Formal Grammars and Gradience

- Classical formal grammars specify recursive definitions of the set of well-formed sentences in a language.
- They are binary decision procedures for membership in this class, and so they cannot, in themselves, accommodate gradience.
- Classical binary theories of grammar must consign gradience to external processing and performance factors.
- This approach is, in principle, plausible, but it must formulate a precise, integrated theory of grammar and processing that predicts the observed phenomenon in detail, to have any explanatory content.
- To date, no such account has been forthcoming.

A Criticism of the Neural LM Approach to Sentence Acceptability

- Sprouse et al. (2018) (SYIFB) argue that LCL's models capture gradience in human acceptability ratings at the cost of accuracy in binary classification of sentences as acceptable or unacceptable.
- They train LCL's RNN on the BNC, and they test it, with SLOR, on three corpora.
- These corpora include 1. Sprouse et al. (2013)'s set of 150 sentence pairs (good and bad) from *Linguistic Inquiry* articles (LI), 2. Adger (2003)'s example pairs, and 3. 120 permutations of the words in *Colorless green ideas sleep furiously* (CGI).

A Criticism of the Neural LM Approach to Sentence Acceptability

- Sprouse et al. (2018) (SYIFB) argue that LCL's models capture gradience in human acceptability ratings at the cost of accuracy in binary classification of sentences as acceptable or unacceptable.
- They train LCL's RNN on the BNC, and they test it, with SLOR, on three corpora.
- These corpora include 1. Sprouse et al. (2013)'s set of 150 sentence pairs (good and bad) from *Linguistic Inquiry* articles (LI), 2. Adger (2003)'s example pairs, and 3. 120 permutations of the words in *Colorless green ideas sleep furiously* (CGI).

A Criticism of the Neural LM Approach to Sentence Acceptability

- Sprouse et al. (2018) (SYIFB) argue that LCL's models capture gradience in human acceptability ratings at the cost of accuracy in binary classification of sentences as acceptable or unacceptable.
- They train LCL's RNN on the BNC, and they test it, with SLOR, on three corpora.
- These corpora include 1. Sprouse et al. (2013)'s set of 150 sentence pairs (good and bad) from *Linguistic Inquiry* articles (LI), 2. Adger (2003)'s example pairs, and 3. 120 permutations of the words in *Colorless green ideas sleep furiously* (CGI).

A Defence of Binary Formal Grammars

- SYIFB report that LCL's RNN + SLOR achieves Pearson correlations of 0.36 for the mean human ratings of the LI test set, 0.55 for Adger's set, and 0.44 for CGI.
- They then use the RNN as a binary classifier for the LI and Adger sets, comparing its performance with that of what they describe as a "binary grammar".
- SYIFB's binary grammar is a measure of the Pearson correlation between the linguists' judgments, reported in the LI articles and Adger's textbook, with the mean crowd source acceptability ratings of these sentences.
- While the Pearson r scores of the RNN are 0.4 for LI and 0.51 for Adger, SYIFB's binary grammar metric achieves 0.71 for the former and 0.87 for the latter.
- SYIFB claim that LCL's RNN performs badly in binary acceptability classification in comparison to their binary grammar system.

A Defence of Binary Formal Grammars

- SYIFB report that LCL's RNN + SLOR achieves Pearson correlations of 0.36 for the mean human ratings of the LI test set, 0.55 for Adger's set, and 0.44 for CGI.
- They then use the RNN as a binary classifier for the LI and Adger sets, comparing its performance with that of what they describe as a "binary grammar".
- SYIFB's binary grammar is a measure of the Pearson correlation between the linguists' judgments, reported in the LI articles and Adger's textbook, with the mean crowd source acceptability ratings of these sentences.
- While the Pearson r scores of the RNN are 0.4 for LI and 0.51 for Adger, SYIFB's binary grammar metric achieves 0.71 for the former and 0.87 for the latter.
- SYIFB claim that LCL's RNN performs badly in binary acceptability classification in comparison to their binary grammar system.

A Defence of Binary Formal Grammars

- SYIFB report that LCL's RNN + SLOR achieves Pearson correlations of 0.36 for the mean human ratings of the LI test set, 0.55 for Adger's set, and 0.44 for CGI.
- They then use the RNN as a binary classifier for the LI and Adger sets, comparing its performance with that of what they describe as a "binary grammar".
- SYIFB's binary grammar is a measure of the Pearson correlation between the linguists' judgments, reported in the LI articles and Adger's textbook, with the mean crowd source acceptability ratings of these sentences.
- While the Pearson r scores of the RNN are 0.4 for LI and 0.51 for Adger, SYIFB's binary grammar metric achieves 0.71 for the former and 0.87 for the latter.
- SYIFB claim that LCL's RNN performs badly in binary acceptability classification in comparison to their binary grammar system.

A Defence of Binary Formal Grammars

- SYIFB report that LCL's RNN + SLOR achieves Pearson correlations of 0.36 for the mean human ratings of the LI test set, 0.55 for Adger's set, and 0.44 for CGI.
- They then use the RNN as a binary classifier for the LI and Adger sets, comparing its performance with that of what they describe as a "binary grammar".
- SYIFB's binary grammar is a measure of the Pearson correlation between the linguists' judgments, reported in the LI articles and Adger's textbook, with the mean crowd source acceptability ratings of these sentences.
- While the Pearson r scores of the RNN are 0.4 for LI and 0.51 for Adger, SYIFB's binary grammar metric achieves 0.71 for the former and 0.87 for the latter.
- SYIFB claim that LCL's RNN performs badly in binary acceptability classification in comparison to their binary grammar system.

A Defence of Binary Formal Grammars

- SYIFB report that LCL's RNN + SLOR achieves Pearson correlations of 0.36 for the mean human ratings of the LI test set, 0.55 for Adger's set, and 0.44 for CGI.
- They then use the RNN as a binary classifier for the LI and Adger sets, comparing its performance with that of what they describe as a "binary grammar".
- SYIFB's binary grammar is a measure of the Pearson correlation between the linguists' judgments, reported in the LI articles and Adger's textbook, with the mean crowd source acceptability ratings of these sentences.
- While the Pearson r scores of the RNN are 0.4 for LI and 0.51 for Adger, SYIFB's binary grammar metric achieves 0.71 for the former and 0.87 for the latter.
- SYIFB claim that LCL's RNN performs badly in binary acceptability classification in comparison to their binary grammar system.

Why the Defence does Not Succeed

- Lappin and Lau (2018) (LL) point out that SYIFB's defence of binary grammar as a classifier for sentence acceptability is without force.
- The "binary grammar metric" which they use as a standard of comparison is neither a grammar nor a model.
- Instead it is a version of the one-vs-rest correlation for estimating an upper bound on any model's expected performance.
- SYIFB suggest that it is an idealised, if unspecified, categorical grammar from which the linguists' judgments are derived.
- To assume such a grammar without formulating it is entirely circular, as its existence is the question at issue.

Why the Defence does Not Succeed

- Lappin and Lau (2018) (LL) point out that SYIFB's defence of binary grammar as a classifier for sentence acceptability is without force.
- The "binary grammar metric" which they use as a standard of comparison is neither a grammar nor a model.
- Instead it is a version of the one-vs-rest correlation for estimating an upper bound on any model's expected performance.
- SYIFB suggest that it is an idealised, if unspecified, categorical grammar from which the linguists' judgments are derived.
- To assume such a grammar without formulating it is entirely circular, as its existence is the question at issue.

Why the Defence does Not Succeed

- Lappin and Lau (2018) (LL) point out that SYIFB's defence of binary grammar as a classifier for sentence acceptability is without force.
- The "binary grammar metric" which they use as a standard of comparison is neither a grammar nor a model.
- Instead it is a version of the one-vs-rest correlation for estimating an upper bound on any model's expected performance.
- SYIFB suggest that it is an idealised, if unspecified, categorical grammar from which the linguists' judgments are derived.
- To assume such a grammar without formulating it is entirely circular, as its existence is the question at issue.

Why the Defence does Not Succeed

- Lappin and Lau (2018) (LL) point out that SYIFB's defence of binary grammar as a classifier for sentence acceptability is without force.
- The "binary grammar metric" which they use as a standard of comparison is neither a grammar nor a model.
- Instead it is a version of the one-vs-rest correlation for estimating an upper bound on any model's expected performance.
- SYIFB suggest that it is an idealised, if unspecified, categorical grammar from which the linguists' judgments are derived.
- To assume such a grammar without formulating it is entirely circular, as its existence is the question at issue.

Why the Defence does Not Succeed

- Lappin and Lau (2018) (LL) point out that SYIFB's defence of binary grammar as a classifier for sentence acceptability is without force.
- The "binary grammar metric" which they use as a standard of comparison is neither a grammar nor a model.
- Instead it is a version of the one-vs-rest correlation for estimating an upper bound on any model's expected performance.
- SYIFB suggest that it is an idealised, if unspecified, categorical grammar from which the linguists' judgments are derived.
- To assume such a grammar without formulating it is entirely circular, as its existence is the question at issue.

A Corpus of Linguists' Sentences

- Warstadt et al. (2019) (WSB) assembled a corpus of Linguistic Acceptability (CoLA), a set of 10,657 linguists' sentences labelled for grammaticality/ungrammaticality.
- They extend it to include out of domain sentences randomly selected from syntax textbooks and research articles.
- WSB use five linguistics PhD students to rate a subset of 200 sentences of CoLA for binary acceptability value, and they find that the majority annotator scores diverge from the linguists' annotations for 13% of the subcorpus.
- This is a comparatively high rate of divergence, which raises the question of the reliability of linguists' grammaticality judgments as evidence for syntactic theories (Gibson and Fedorenko, 2013; Sprouse and Almeida, 2013; Gibson et al., 2013).

A Corpus of Linguists' Sentences

- Warstadt et al. (2019) (WSB) assembled a corpus of Linguistic Acceptability (CoLA), a set of 10,657 linguists' sentences labelled for grammaticality/ungrammaticality.
- They extend it to include out of domain sentences randomly selected from syntax textbooks and research articles.
- WSB use five linguistics PhD students to rate a subset of 200 sentences of CoLA for binary acceptability value, and they find that the majority annotator scores diverge from the linguists' annotations for 13% of the subcorpus.
- This is a comparatively high rate of divergence, which raises the question of the reliability of linguists' grammaticality judgments as evidence for syntactic theories (Gibson and Fedorenko, 2013; Sprouse and Almeida, 2013; Gibson et al., 2013).

A Corpus of Linguists' Sentences

- Warstadt et al. (2019) (WSB) assembled a corpus of Linguistic Acceptability (CoLA), a set of 10,657 linguists' sentences labelled for grammaticality/ungrammaticality.
- They extend it to include out of domain sentences randomly selected from syntax textbooks and research articles.
- WSB use five linguistics PhD students to rate a subset of 200 sentences of CoLA for binary acceptability value, and they find that the majority annotator scores diverge from the linguists' annotations for 13% of the subcorpus.
- This is a comparatively high rate of divergence, which raises the question of the reliability of linguists' grammaticality judgments as evidence for syntactic theories (Gibson and Fedorenko, 2013; Sprouse and Almeida, 2013; Gibson et al., 2013).

A Corpus of Linguists' Sentences

- Warstadt et al. (2019) (WSB) assembled a corpus of Linguistic Acceptability (CoLA), a set of 10,657 linguists' sentences labelled for grammaticality/ungrammaticality.
- They extend it to include out of domain sentences randomly selected from syntax textbooks and research articles.
- WSB use five linguistics PhD students to rate a subset of 200 sentences of CoLA for binary acceptability value, and they find that the majority annotator scores diverge from the linguists' annotations for 13% of the subcorpus.
- This is a comparatively high rate of divergence, which raises the question of the reliability of linguists' grammaticality judgments as evidence for syntactic theories (Gibson and Fedorenko, 2013; Sprouse and Almeida, 2013; Gibson et al., 2013).

Predicting Acceptability for Linguists' Sentences

- WSB do semi-supervised learning for a variety of LSTM and transformer neural LMs by first training them on the sentences of the BNC and ill formed variants of these sentences derived by permutation.
- They use rich pre-trained word embeddings in this part of the training process.
- They then transfer the sentence vectors obtained, to train a binary classifier on their linguists examples for part of CoLA.
- They test their models, and LCL's RNN (with SLOR and the lexical unigram scoring functions), on the remainder of the CoLA corpus.

Predicting Acceptability for Linguists' Sentences

- WSB do semi-supervised learning for a variety of LSTM and transformer neural LMs by first training them on the sentences of the BNC and ill formed variants of these sentences derived by permutation.
- They use rich pre-trained word embeddings in this part of the training process.
- They then transfer the sentence vectors obtained, to train a binary classifier on their linguists examples for part of CoLA.
- They test their models, and LCL's RNN (with SLOR and the lexical unigram scoring functions), on the remainder of the CoLA corpus.

Predicting Acceptability for Linguists' Sentences

- WSB do semi-supervised learning for a variety of LSTM and transformer neural LMs by first training them on the sentences of the BNC and ill formed variants of these sentences derived by permutation.
- They use rich pre-trained word embeddings in this part of the training process.
- They then transfer the sentence vectors obtained, to train a binary classifier on their linguists examples for part of CoLA.
- They test their models, and LCL's RNN (with SLOR and the lexical unigram scoring functions), on the remainder of the CoLA corpus.

Predicting Acceptability for Linguists' Sentences

- WSB do semi-supervised learning for a variety of LSTM and transformer neural LMs by first training them on the sentences of the BNC and ill formed variants of these sentences derived by permutation.
- They use rich pre-trained word embeddings in this part of the training process.
- They then transfer the sentence vectors obtained, to train a binary classifier on their linguists examples for part of CoLA.
- They test their models, and LCL's RNN (with SLOR and the lexical unigram scoring functions), on the remainder of the CoLA corpus.

Model Performance on CoLA

- WSB report that their models generally outperform LCL's RNN.
- Given the supervised training of these models on CoLA as binary classifiers, and the power of some of WSB's transformers, this is not a surprising result.
- Interestingly, LCL's RNN is competitive on the out of domain part of the CoLA test set.
- Also, it outperforms WSB's models on predicting ratings for three of the five syntactic constructions that they consider.

Model Performance on CoLA

- WSB report that their models generally outperform LCL's RNN.
- Given the supervised training of these models on CoLA as binary classifiers, and the power of some of WSB's transformers, this is not a surprising result.
- Interestingly, LCL's RNN is competitive on the out of domain part of the CoLA test set.
- Also, it outperforms WSB's models on predicting ratings for three of the five syntactic constructions that they consider.

Model Performance on CoLA

- WSB report that their models generally outperform LCL's RNN.
- Given the supervised training of these models on CoLA as binary classifiers, and the power of some of WSB's transformers, this is not a surprising result.
- Interestingly, LCL's RNN is competitive on the out of domain part of the CoLA test set.
- Also, it outperforms WSB's models on predicting ratings for three of the five syntactic constructions that they consider.

Model Performance on CoLA

- WSB report that their models generally outperform LCL's RNN.
- Given the supervised training of these models on CoLA as binary classifiers, and the power of some of WSB's transformers, this is not a surprising result.
- Interestingly, LCL's RNN is competitive on the out of domain part of the CoLA test set.
- Also, it outperforms WSB's models on predicting ratings for three of the five syntactic constructions that they consider.

Evaluating Neural LMs on Targeted Syntactic Tests

- Hu et al. (2020) test the capacity of a number of neural language models to generalise correctly for a set of syntactic phenomena.
- The models include, *inter alia*, a vanilla LSTM, an RNNG, GPT-2, and GPT-2-XL (both GPT-2 LMs are pre-trained and untuned).
- The syntactic phenomena which they test are agreement, licensing (negative polarity and reflexive pronouns), garden path effects, the expectation of large syntactic categories, centre embedding, and long distance dependencies (filler-gap structures and cleft verb dependency).
- Hu et al. report that the two GPT-2 transformer models outperform the other models on most of the tasks, and score the highest average accuracy across the tests suites (approximately 0.8).

Evaluating Neural LMs on Targeted Syntactic Tests

- Hu et al. (2020) test the capacity of a number of neural language models to generalise correctly for a set of syntactic phenomena.
- The models include, *inter alia*, a vanilla LSTM, an RNNG, GPT-2, and GPT-2-XL (both GPT-2 LMs are pre-trained and untuned).
- The syntactic phenomena which they test are agreement, licensing (negative polarity and reflexive pronouns), garden path effects, the expectation of large syntactic categories, centre embedding, and long distance dependencies (filler-gap structures and cleft verb dependency).
- Hu et al. report that the two GPT-2 transformer models outperform the other models on most of the tasks, and score the highest average accuracy across the tests suites (approximately 0.8).

Evaluating Neural LMs on Targeted Syntactic Tests

- Hu et al. (2020) test the capacity of a number of neural language models to generalise correctly for a set of syntactic phenomena.
- The models include, *inter alia*, a vanilla LSTM, an RNNG, GPT-2, and GPT-2-XL (both GPT-2 LMs are pre-trained and untuned).
- The syntactic phenomena which they test are agreement, licensing (negative polarity and reflexive pronouns), garden path effects, the expectation of large syntactic categories, centre embedding, and long distance dependencies (filler-gap structures and cleft verb dependency).
- Hu et al. report that the two GPT-2 transformer models outperform the other models on most of the tasks, and score the highest average accuracy across the tests suites (approximately 0.8).

Evaluating Neural LMs on Targeted Syntactic Tests

- Hu et al. (2020) test the capacity of a number of neural language models to generalise correctly for a set of syntactic phenomena.
- The models include, *inter alia*, a vanilla LSTM, an RNNG, GPT-2, and GPT-2-XL (both GPT-2 LMs are pre-trained and untuned).
- The syntactic phenomena which they test are agreement, licensing (negative polarity and reflexive pronouns), garden path effects, the expectation of large syntactic categories, centre embedding, and long distance dependencies (filler-gap structures and cleft verb dependency).
- Hu et al. report that the two GPT-2 transformer models outperform the other models on most of the tasks, and score the highest average accuracy across the tests suites (approximately 0.8).

ML and Universal Grammar

- Both SYIFB and WSB suggest that if it is necessary to enrich the training data of ML systems with symbolic features such as POS tags or syntactic trees, then these features will correspond to domain specific learning biases.
- They take these biases to be the conditions required for human language acquisition.
- SYIFB and WSB identify them with the learning theoretic content of an innate Universal Grammar (UG).
- In fact this claim is not at all warranted.

ML and Universal Grammar

- Both SYIFB and WSB suggest that if it is necessary to enrich the training data of ML systems with symbolic features such as POS tags or syntactic trees, then these features will correspond to domain specific learning biases.
- They take these biases to be the conditions required for human language acquisition.
- SYIFB and WSB identify them with the learning theoretic content of an innate Universal Grammar (UG).
- In fact this claim is not at all warranted.

ML and Universal Grammar

- Both SYIFB and WSB suggest that if it is necessary to enrich the training data of ML systems with symbolic features such as POS tags or syntactic trees, then these features will correspond to domain specific learning biases.
- They take these biases to be the conditions required for human language acquisition.
- SYIFB and WSB identify them with the learning theoretic content of an innate Universal Grammar (UG).
- In fact this claim is not at all warranted.

ML and Universal Grammar

- Both SYIFB and WSB suggest that if it is necessary to enrich the training data of ML systems with symbolic features such as POS tags or syntactic trees, then these features will correspond to domain specific learning biases.
- They take these biases to be the conditions required for human language acquisition.
- SYIFB and WSB identify them with the learning theoretic content of an innate Universal Grammar (UG).
- In fact this claim is not at all warranted.

ML Learning of Symbolic Features

- As we have seen, it is not at all clear that enriching training data with syntactic/semantic markers, or with trees significantly improves the performance of DNN models on most NLP tasks, and, in the case of the sentence acceptability task, such enrichment degrades performance.
- Even if such features were necessary for human level accuracy on NLP tasks, they can be learned from data, without strong domain specific learning biases.
- ML systems can efficiently learn POS tags (for example, Clark, 2003).
- Similarly, Clark (2015) shows that distributional learning can induce tree structures on strings (strong learning), for a subclass of Context-Free Grammars, and current work seeks to extend these results to Parallel Context-Free Grammars (Clark and Yoshinaka, 2014).

ML Learning of Symbolic Features

- As we have seen, it is not at all clear that enriching training data with syntactic/semantic markers, or with trees significantly improves the performance of DNN models on most NLP tasks, and, in the case of the sentence acceptability task, such enrichment degrades performance.
- Even if such features were necessary for human level accuracy on NLP tasks, they can be learned from data, without strong domain specific learning biases.
- ML systems can efficiently learn POS tags (for example, Clark, 2003).
- Similarly, Clark (2015) shows that distributional learning can induce tree structures on strings (strong learning), for a subclass of Context-Free Grammars, and current work seeks to extend these results to Parallel Context-Free Grammars (Clark and Yoshinaka, 2014).

ML Learning of Symbolic Features

- As we have seen, it is not at all clear that enriching training data with syntactic/semantic markers, or with trees significantly improves the performance of DNN models on most NLP tasks, and, in the case of the sentence acceptability task, such enrichment degrades performance.
- Even if such features were necessary for human level accuracy on NLP tasks, they can be learned from data, without strong domain specific learning biases.
- ML systems can efficiently learn POS tags (for example, Clark, 2003).
- Similarly, Clark (2015) shows that distributional learning can induce tree structures on strings (strong learning), for a subclass of Context-Free Grammars, and current work seeks to extend these results to Parallel Context-Free Grammars (Clark and Yoshinaka, 2014).

ML Learning of Symbolic Features

- As we have seen, it is not at all clear that enriching training data with syntactic/semantic markers, or with trees significantly improves the performance of DNN models on most NLP tasks, and, in the case of the sentence acceptability task, such enrichment degrades performance.
- Even if such features were necessary for human level accuracy on NLP tasks, they can be learned from data, without strong domain specific learning biases.
- ML systems can efficiently learn POS tags (for example, Clark, 2003).
- Similarly, Clark (2015) shows that distributional learning can induce tree structures on strings (strong learning), for a subclass of Context-Free Grammars, and current work seeks to extend these results to Parallel Context-Free Grammars (Clark and Yoshinaka, 2014).

Arguments on the Limitations of ML Methods

- SYIFB argue from the putative inadequacy of LCL's RNN to the non-viability of ML methods in general, as an approach to modelling language acquisition and linguistic representation.
- Aside from the fact that their argument concerning the RNN does not go through, it also over reaches in making claims about the full class of ML methods.
- LALPS show that bidirectional transformers approach human performance on the sentence acceptability task.
- The history of linguistics and cognitive science is replete with unsound arguments from the limitations of a particular class of models to the non-viability of the entire approach to learning and representation that these models exemplify.

Arguments on the Limitations of ML Methods

- SYIFB argue from the putative inadequacy of LCL's RNN to the non-viability of ML methods in general, as an approach to modelling language acquisition and linguistic representation.
- Aside from the fact that their argument concerning the RNN does not go through, it also over reaches in making claims about the full class of ML methods.
- LALPS show that bidirectional transformers approach human performance on the sentence acceptability task.
- The history of linguistics and cognitive science is replete with unsound arguments from the limitations of a particular class of models to the non-viability of the entire approach to learning and representation that these models exemplify.

Arguments on the Limitations of ML Methods

- SYIFB argue from the putative inadequacy of LCL's RNN to the non-viability of ML methods in general, as an approach to modelling language acquisition and linguistic representation.
- Aside from the fact that their argument concerning the RNN does not go through, it also over reaches in making claims about the full class of ML methods.
- LALPS show that bidirectional transformers approach human performance on the sentence acceptability task.
- The history of linguistics and cognitive science is replete with unsound arguments from the limitations of a particular class of models to the non-viability of the entire approach to learning and representation that these models exemplify.

Arguments on the Limitations of ML Methods

- SYIFB argue from the putative inadequacy of LCL's RNN to the non-viability of ML methods in general, as an approach to modelling language acquisition and linguistic representation.
- Aside from the fact that their argument concerning the RNN does not go through, it also over reaches in making claims about the full class of ML methods.
- LALPS show that bidirectional transformers approach human performance on the sentence acceptability task.
- The history of linguistics and cognitive science is replete with unsound arguments from the limitations of a particular class of models to the non-viability of the entire approach to learning and representation that these models exemplify.

Example 1: Chomsky (1957) on Probabilistic Bigrams

- Chomsky (1957) observes that simple probabilistic bigram models that use raw frequency counts of lexical items assign nil probability to both

  1a. Colourless green ideas sleep furiously.

  b. Furiously sleep ideas green colourlessly.

- He concludes that no probabilistic characterisation of grammaticality can succeed, a view that has been widely accepted among theoretical linguists over many years.
- Pereira (2000) shows that if bigram models are extended to include smoothing for unseen events, and hidden variables for word classes (identified from the data through word distributions), then a bigram model trained on newspaper text assigns a significantly higher probability value to 1a than to 1b.

Example 1: Chomsky (1957) on Probabilistic Bigrams

- Chomsky (1957) observes that simple probabilistic bigram models that use raw frequency counts of lexical items assign nil probability to both
  1a. Colourless green ideas sleep furiously.
  b. Furiously sleep ideas green colourlessly.
- He concludes that no probabilistic characterisation of grammaticality can succeed, a view that has been widely accepted among theoretical linguists over many years.
- Pereira (2000) shows that if bigram models are extended to include smoothing for unseen events, and hidden variables for word classes (identified from the data through word distributions), then a bigram model trained on newspaper text assigns a significantly higher probability value to 1a than to 1b.

Example 1: Chomsky (1957) on Probabilistic Bigrams

- Chomsky (1957) observes that simple probabilistic bigram models that use raw frequency counts of lexical items assign nil probability to both
  1a. Colourless green ideas sleep furiously.
  b. Furiously sleep ideas green colourlessly.
- He concludes that no probabilistic characterisation of grammaticality can succeed, a view that has been widely accepted among theoretical linguists over many years.
- Pereira (2000) shows that if bigram models are extended to include smoothing for unseen events, and hidden variables for word classes (identified from the data through word distributions), then a bigram model trained on newspaper text assigns a significantly higher probability value to 1a than to 1b.

Example 2: Gold's (1967) Identification in the Limit

- Gold (1967) shows that, given his Identification In the Limit (IIL) learning paradigm, and presentations of positive evidence only, a learner can acquire the class of finite languages, and a finite class of (possibly infinite) languages.
- However, on IIL with positive evidence only, a learner cannot learn a suprafinite class, which contains the class of finite languages and at least one infinite language.
- Therefore none of the language classes of the Chomsky Hierarchy are learnable through induction from positive evidence.
- Some advocates of UG take Gold's results to demonstrate that strong innate, domain specific constraints on learning are a necessary condition for human language acquisition (for example, Crain and Thornton,1998).

Example 2: Gold's (1967) Identification in the Limit

- Gold (1967) shows that, given his Identification In the Limit (IIL) learning paradigm, and presentations of positive evidence only, a learner can acquire the class of finite languages, and a finite class of (possibly infinite) languages.
- However, on IIL with positive evidence only, a learner cannot learn a suprafinite class, which contains the class of finite languages and at least one infinite language.
- Therefore none of the language classes of the Chomsky Hierarchy are learnable through induction from positive evidence.
- Some advocates of UG take Gold's results to demonstrate that strong innate, domain specific constraints on learning are a necessary condition for human language acquisition (for example, Crain and Thornton,1998).

Example 2: Gold's (1967) Identification in the Limit

- Gold (1967) shows that, given his Identification In the Limit (IIL) learning paradigm, and presentations of positive evidence only, a learner can acquire the class of finite languages, and a finite class of (possibly infinite) languages.
- However, on IIL with positive evidence only, a learner cannot learn a suprafinite class, which contains the class of finite languages and at least one infinite language.
- Therefore none of the language classes of the Chomsky Hierarchy are learnable through induction from positive evidence.
- Some advocates of UG take Gold's results to demonstrate that strong innate, domain specific constraints on learning are a necessary condition for human language acquisition (for example, Crain and Thornton,1998).

Example 2: Gold's (1967) Identification in the Limit

- Gold (1967) shows that, given his Identification In the Limit (IIL) learning paradigm, and presentations of positive evidence only, a learner can acquire the class of finite languages, and a finite class of (possibly infinite) languages.
- However, on IIL with positive evidence only, a learner cannot learn a suprafinite class, which contains the class of finite languages and at least one infinite language.
- Therefore none of the language classes of the Chomsky Hierarchy are learnable through induction from positive evidence.
- Some advocates of UG take Gold's results to demonstrate that strong innate, domain specific constraints on learning are a necessary condition for human language acquisition (for example, Crain and Thornton,1998).

Alternatives to IIL

- Gold's paradigm relies on a number of highly implausible assumptions concerning the nature of learning, and the evidence available to the language learner.
- When IIL is replaced by models specified in terms of a more realistic view of the learning process, then it is possible to prove that a much richer class of languages (and of grammars) can be efficiently acquired through data driven induction procedures.
- These models do not posit strong domain specific learning biases of the kind encoded in UG.
- Clark and Lappin (2011, 2013) provide detailed discussion of the problems with IIL, and they consider more plausible learning models.

Alternatives to IIL

- Gold's paradigm relies on a number of highly implausible assumptions concerning the nature of learning, and the evidence available to the language learner.
- When IIL is replaced by models specified in terms of a more realistic view of the learning process, then it is possible to prove that a much richer class of languages (and of grammars) can be efficiently acquired through data driven induction procedures.
- These models do not posit strong domain specific learning biases of the kind encoded in UG.
- Clark and Lappin (2011, 2013) provide detailed discussion of the problems with IIL, and they consider more plausible learning models.

Alternatives to IIL

- Gold's paradigm relies on a number of highly implausible assumptions concerning the nature of learning, and the evidence available to the language learner.
- When IIL is replaced by models specified in terms of a more realistic view of the learning process, then it is possible to prove that a much richer class of languages (and of grammars) can be efficiently acquired through data driven induction procedures.
- These models do not posit strong domain specific learning biases of the kind encoded in UG.
- Clark and Lappin (2011, 2013) provide detailed discussion of the problems with IIL, and they consider more plausible learning models.

Alternatives to IIL

- Gold's paradigm relies on a number of highly implausible assumptions concerning the nature of learning, and the evidence available to the language learner.
- When IIL is replaced by models specified in terms of a more realistic view of the learning process, then it is possible to prove that a much richer class of languages (and of grammars) can be efficiently acquired through data driven induction procedures.
- These models do not posit strong domain specific learning biases of the kind encoded in UG.
- Clark and Lappin (2011, 2013) provide detailed discussion of the problems with IIL, and they consider more plausible learning models.

Example 3: Fodor and Pylyshyn on Connectionism

- Fodor and Pylyshyn (1988), and Fodor (2000) point out the serious limitations in the learning abilities of simple feed forward NNs with a single layer of hidden units.
- They conclude that neural networks in general are incapable of acquiring human level knowledge in most AI applications, particularly in natural language processing.
- Again, the argument involves an unsound inference from the limitations of a particular subclass of ML systems to the non-viability of ML in general as a way of modelling human learning and representation.
- As we have seen, DNNs, particularly bidirectional transformers, approach, and, in some cases, surpass human performance across a variety of NLP tasks.

Example 3: Fodor and Pylyshyn on Connectionism

- Fodor and Pylyshyn (1988), and Fodor (2000) point out the serious limitations in the learning abilities of simple feed forward NNs with a single layer of hidden units.
- They conclude that neural networks in general are incapable of acquiring human level knowledge in most AI applications, particularly in natural language processing.
- Again, the argument involves an unsound inference from the limitations of a particular subclass of ML systems to the non-viability of ML in general as a way of modelling human learning and representation.
- As we have seen, DNNs, particularly bidirectional transformers, approach, and, in some cases, surpass human performance across a variety of NLP tasks.

Example 3: Fodor and Pylyshyn on Connectionism

- Fodor and Pylyshyn (1988), and Fodor (2000) point out the serious limitations in the learning abilities of simple feed forward NNs with a single layer of hidden units.
- They conclude that neural networks in general are incapable of acquiring human level knowledge in most AI applications, particularly in natural language processing.
- Again, the argument involves an unsound inference from the limitations of a particular subclass of ML systems to the non-viability of ML in general as a way of modelling human learning and representation.
- As we have seen, DNNs, particularly bidirectional transformers, approach, and, in some cases, surpass human performance across a variety of NLP tasks.

Example 3: Fodor and Pylyshyn on Connectionism

- Fodor and Pylyshyn (1988), and Fodor (2000) point out the serious limitations in the learning abilities of simple feed forward NNs with a single layer of hidden units.
- They conclude that neural networks in general are incapable of acquiring human level knowledge in most AI applications, particularly in natural language processing.
- Again, the argument involves an unsound inference from the limitations of a particular subclass of ML systems to the non-viability of ML in general as a way of modelling human learning and representation.
- As we have seen, DNNs, particularly bidirectional transformers, approach, and, in some cases, surpass human performance across a variety of NLP tasks.

Comparing ML Models and Formal Grammars

- Work on DNN models for the learning and representation of natural language is still in its infancy.
- It is reasonable to expect that entirely new types of machine learning architectures will replace current DNNs, and that these may well yield significant gains in modelling ability across a range of linguistic applications.
- It is necessary for advocates of a categorial grammar, derived from a strong bias UG view of language acquisition, to produce a genuine computational model that provides a non-trivial classifier for acceptability.
- Only when such a system is available can we compare a UG approach to ML models for performance in language acquisition, and the handling of NLP applications.

Comparing ML Models and Formal Grammars

- Work on DNN models for the learning and representation of natural language is still in its infancy.
- It is reasonable to expect that entirely new types of machine learning architectures will replace current DNNs, and that these may well yield significant gains in modelling ability across a range of linguistic applications.
- It is necessary for advocates of a categorial grammar, derived from a strong bias UG view of language acquisition, to produce a genuine computational model that provides a non-trivial classifier for acceptability.
- Only when such a system is available can we compare a UG approach to ML models for performance in language acquisition, and the handling of NLP applications.

Comparing ML Models and Formal Grammars

- Work on DNN models for the learning and representation of natural language is still in its infancy.
- It is reasonable to expect that entirely new types of machine learning architectures will replace current DNNs, and that these may well yield significant gains in modelling ability across a range of linguistic applications.
- It is necessary for advocates of a categorial grammar, derived from a strong bias UG view of language acquisition, to produce a genuine computational model that provides a non-trivial classifier for acceptability.
- Only when such a system is available can we compare a UG approach to ML models for performance in language acquisition, and the handling of NLP applications.

Comparing ML Models and Formal Grammars

- Work on DNN models for the learning and representation of natural language is still in its infancy.
- It is reasonable to expect that entirely new types of machine learning architectures will replace current DNNs, and that these may well yield significant gains in modelling ability across a range of linguistic applications.
- It is necessary for advocates of a categorial grammar, derived from a strong bias UG view of language acquisition, to produce a genuine computational model that provides a non-trivial classifier for acceptability.
- Only when such a system is available can we compare a UG approach to ML models for performance in language acquisition, and the handling of NLP applications.

Conclusions

- DNNs have achieved significant progress in modelling human linguistic knowledge across a wide range of NLP tasks.
- Adding formal syntactic and semantic representations to training data, or to the internal architecture of a model, does not seem to improve the performance of DNNs for most tasks.
- DNNs learn hierarchical syntactic structure, long range dependencies, and semantic relations, which they represent in distributed form, through vectors derived from rich lexical embeddings.
- It is not possible to compare ML models to formal grammars as systems for learning and representation until the latter have been encoded in wide coverage computational devices.

Conclusions

- DNNs have achieved significant progress in modelling human linguistic knowledge across a wide range of NLP tasks.
- Adding formal syntactic and semantic representations to training data, or to the internal architecture of a model, does not seem to improve the performance of DNNs for most tasks.
- DNNs learn hierarchical syntactic structure, long range dependencies, and semantic relations, which they represent in distributed form, through vectors derived from rich lexical embeddings.
- It is not possible to compare ML models to formal grammars as systems for learning and representation until the latter have been encoded in wide coverage computational devices.

Conclusions

- DNNs have achieved significant progress in modelling human linguistic knowledge across a wide range of NLP tasks.
- Adding formal syntactic and semantic representations to training data, or to the internal architecture of a model, does not seem to improve the performance of DNNs for most tasks.
- DNNs learn hierarchical syntactic structure, long range dependencies, and semantic relations, which they represent in distributed form, through vectors derived from rich lexical embeddings.
- It is not possible to compare ML models to formal grammars as systems for learning and representation until the latter have been encoded in wide coverage computational devices.

Conclusions

- DNNs have achieved significant progress in modelling human linguistic knowledge across a wide range of NLP tasks.
- Adding formal syntactic and semantic representations to training data, or to the internal architecture of a model, does not seem to improve the performance of DNNs for most tasks.
- DNNs learn hierarchical syntactic structure, long range dependencies, and semantic relations, which they represent in distributed form, through vectors derived from rich lexical embeddings.
- It is not possible to compare ML models to formal grammars as systems for learning and representation until the latter have been encoded in wide coverage computational devices.

Future Work

- Applying DNNs to multimodal NLP tasks involving text, images, and sound is an emerging field of work showing considerable promise.
- These tasks include generating detailed image descriptions, question answering about visual scenes, and learning to perform actions in response to verbal or textual commands relating to video simulations.
- To the extent that DNNs can master these tasks, they will achieve a deeper level of linguistic understanding than most of the current systems.
- Work on Multi-task DNNs, which have a generic architecture and are fine tuned for different tasks, transferring information among tasks, is another important area of innovative research that is relevant to linguistic learning and representation.

Future Work

- Applying DNNs to multimodal NLP tasks involving text, images, and sound is an emerging field of work showing considerable promise.
- These tasks include generating detailed image descriptions, question answering about visual scenes, and learning to perform actions in response to verbal or textual commands relating to video simulations.
- To the extent that DNNs can master these tasks, they will achieve a deeper level of linguistic understanding than most of the current systems.
- Work on Multi-task DNNs, which have a generic architecture and are fine tuned for different tasks, transferring information among tasks, is another important area of innovative research that is relevant to linguistic learning and representation.

Future Work

- Applying DNNs to multimodal NLP tasks involving text, images, and sound is an emerging field of work showing considerable promise.
- These tasks include generating detailed image descriptions, question answering about visual scenes, and learning to perform actions in response to verbal or textual commands relating to video simulations.
- To the extent that DNNs can master these tasks, they will achieve a deeper level of linguistic understanding than most of the current systems.
- Work on Multi-task DNNs, which have a generic architecture and are fine tuned for different tasks, transferring information among tasks, is another important area of innovative research that is relevant to linguistic learning and representation.

Future Work

- Applying DNNs to multimodal NLP tasks involving text, images, and sound is an emerging field of work showing considerable promise.
- These tasks include generating detailed image descriptions, question answering about visual scenes, and learning to perform actions in response to verbal or textual commands relating to video simulations.
- To the extent that DNNs can master these tasks, they will achieve a deeper level of linguistic understanding than most of the current systems.
- Work on Multi-task DNNs, which have a generic architecture and are fine tuned for different tasks, transferring information among tasks, is another important area of innovative research that is relevant to linguistic learning and representation.