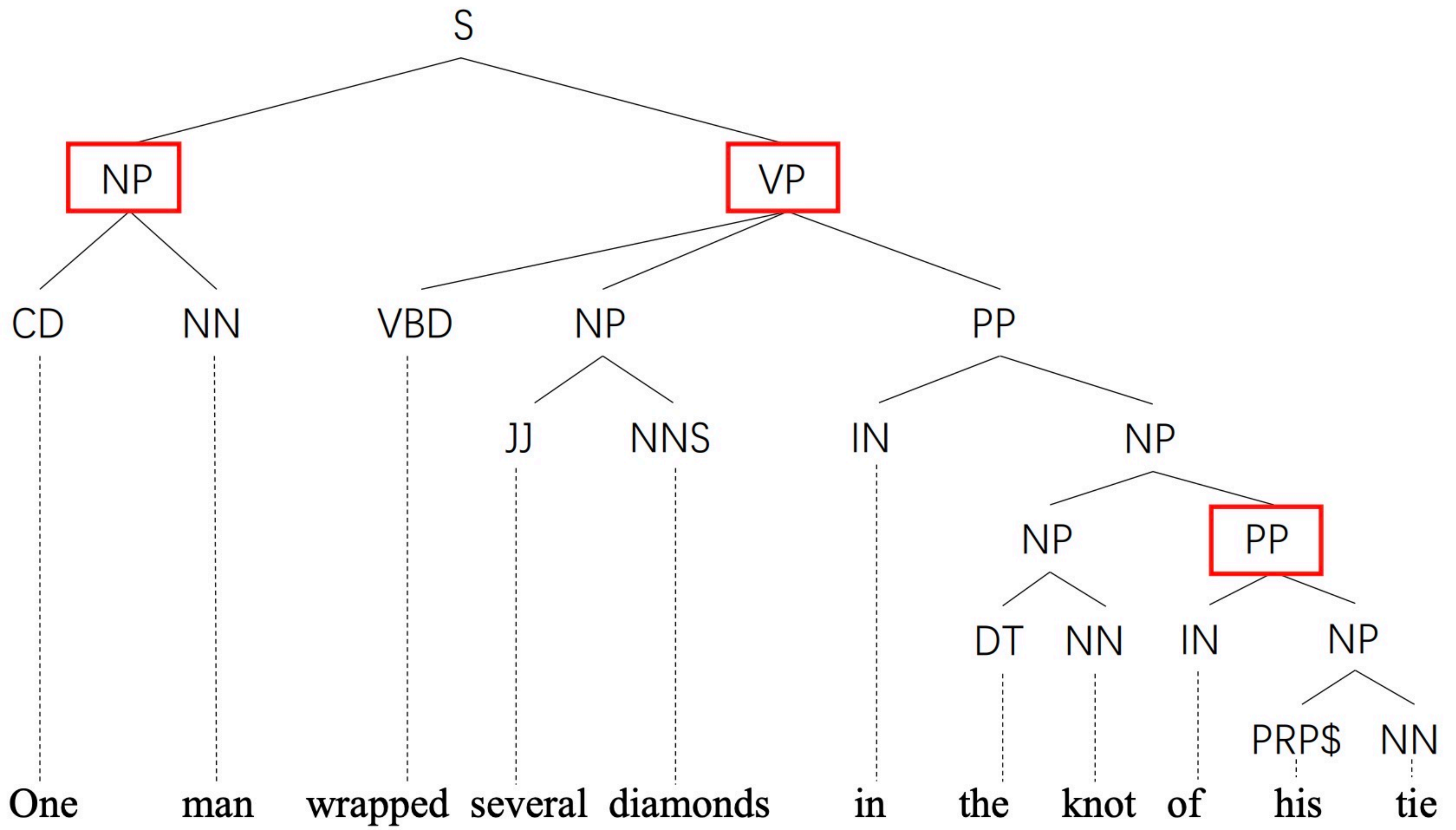
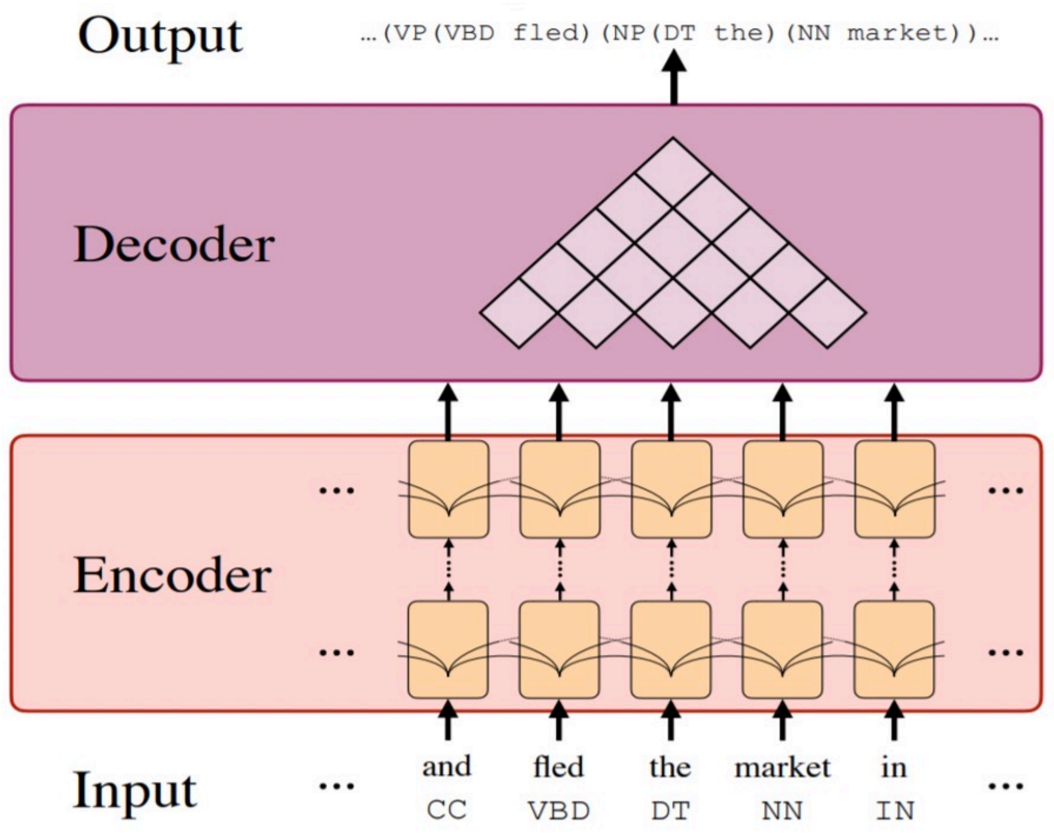
# Investigating Non-local Features for Neural Constituency Parsing

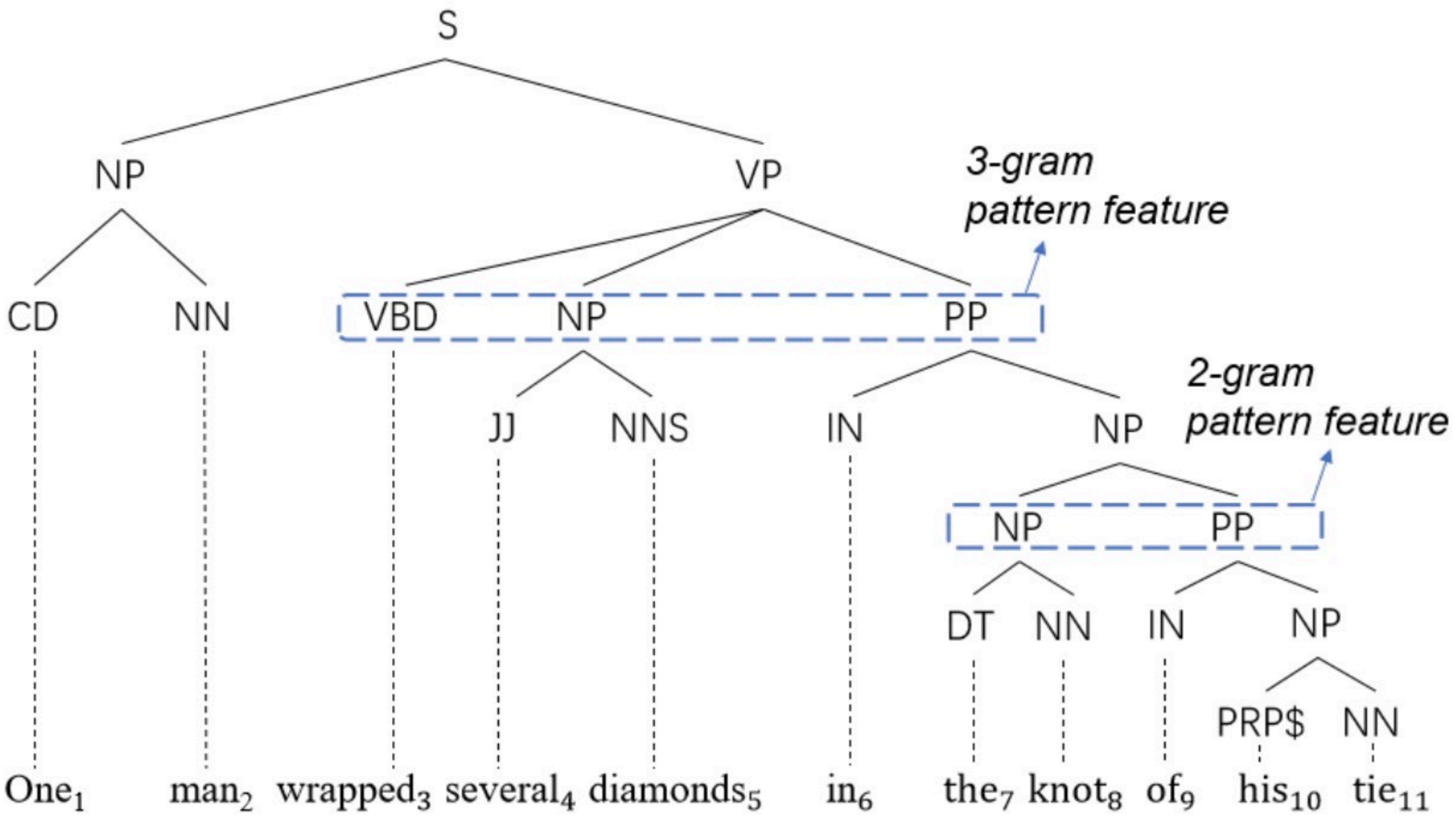
**Leyang Cui**[♡♠*], **Sen Yang**[♠*], **Yue Zhang**[♠◇†]
[♡]Zhejiang University
[♠]School of Engineering, Westlake University
[◇]Institute of Advanced Technology, Westlake Institute for Advanced Study
{cuileyang, zhangyue}@westlake.edu.cn senyang.stu@gmail.com




Constituency Parsing




- Tree Structure $s(T) = \sum_{(i,j,l)\in T} s(i,j,l^c)$
- Token Representation $\mathbf{H}_1^n = \{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_n\}$
- Span Representation $\mathbf{v}_{i,j} = \mathbf{h}_j - \mathbf{h}_i$
- Scoring Function $s(i,j,\cdot) = \mathbf{W}_2^c \mathrm{ReLU}(\mathbf{W}_1^c \mathbf{v}_{i,j} + \mathbf{b}_1^c) + \mathbf{b}_2^c$
- Training $\mathcal{L}_{\text{cons}} = \max\left(0, \max_{T \neq T^*} [s(T) + \Delta(T, T^*)] - s(T^*)\right)$
- Inference $\hat{T} = \operatorname*{argmax}_T s(T)$
- **Research Question**
  - Model makes local prediction on each span representation. How to model the output dependency in the encoder?




Constituency Parsing

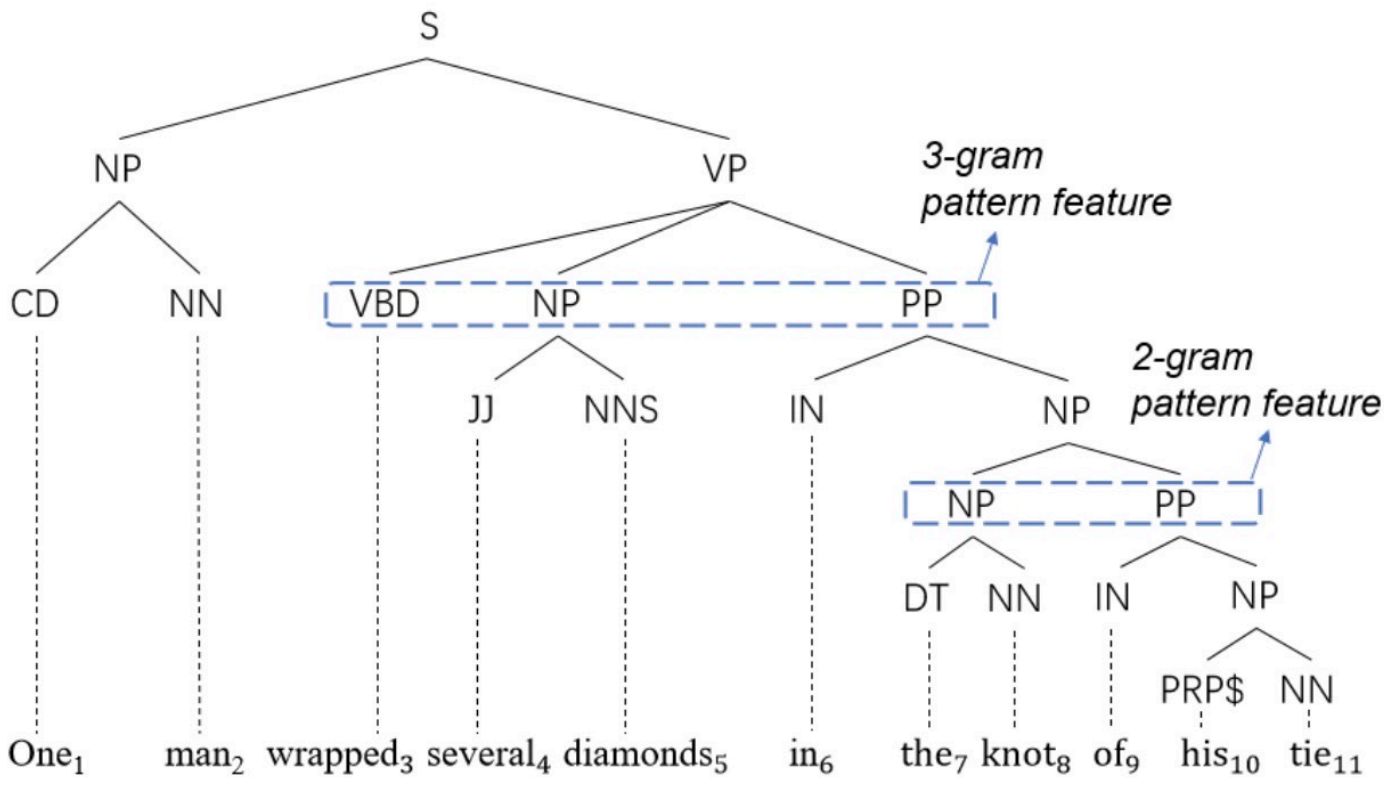
- **Auxiliary training objective 1: Pattern Prediction**

  - Pattern prediction $$\hat{p}_{i,j} = \text{Softmax}\left(\mathbf{W}_2^{\text{p}}\text{R}\small{\text{E}}\normalsize\text{LU}(\mathbf{W}_1^{\text{p}}\mathbf{v}_{i,j} + \mathbf{b}_1^{\text{p}}) + \mathbf{b}_2^{\text{p}}\right)$$

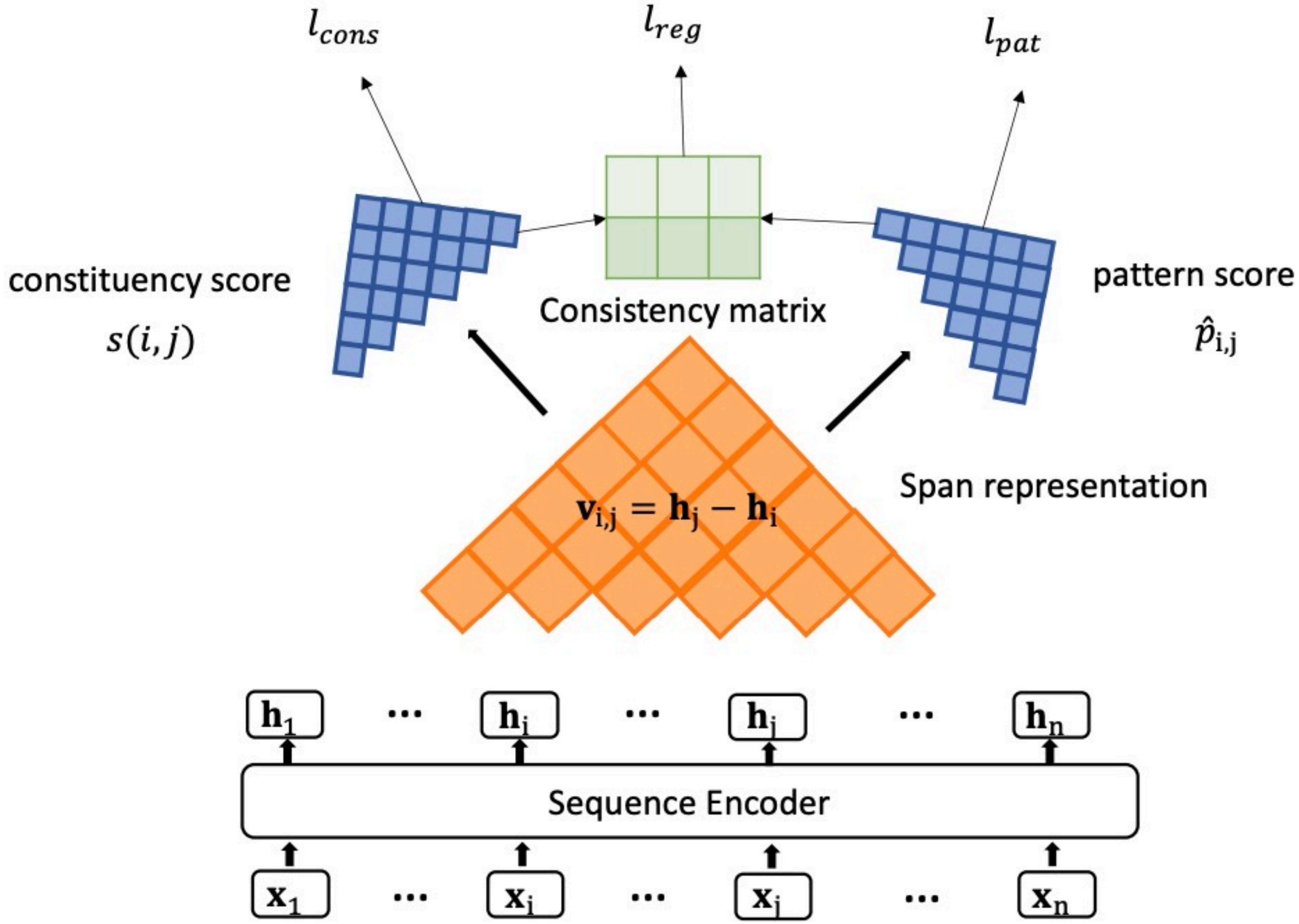
  - Pattern Loss $$\mathcal{L}_{pat} = -\sum_{i=1}^{n}\sum_{j=1}^{n} p_{i,j} \log \hat{p}_{i,j}$$

- **Auxiliary training objective 2: legality between pattern and constituent**
  - Constituent span $(i_t, j_t, l_t^c)$ is a subtree of pattern span $(i_{t'}, j_{t'}, l_{t'}^p)$
    -> $l_t^c$ is legal to co-occurrence with $l_{t'}^p$.

    Both NNS and NP are legal to occur as sub-trees of the 3-gram pattern {VBD NP PP}
    S or ADJP cannot be contained with in {VBD NP PP} based on the grammar rule.



- **Auxiliary training objective 2: legality between pattern and constituent**
  - Constituent span $(i_t\ , j_t\ ,\ l_t^c)$ is a subtree of pattern span $(i_{t'}, j_{t'},\ l_{t'}^p)$ -> $l_t^c$ is legal to co-occurrence with $l_{t'}^p$.
  - Instance-level Consistency $\hat{\mathbf{Y}} = \mathrm{Sigmoid}\big((\mathbf{W}_2^{\mathrm{c}\,\mathsf{T}}\mathbf{U}_1\mathbf{V})(\mathbf{V}^{\mathsf{T}}\mathbf{U}_2\mathbf{W}_2^{\mathrm{p}})\big)$
  - Corpus-level Consistency $\hat{\mathbf{Y}} = \mathrm{Sigmoid}\big(\mathbf{W}_2^{\mathrm{c}}\mathbf{U}\mathbf{W}_2^{\mathrm{p}\,\mathsf{T}}\big)$
  - Consistency Loss $\mathcal{L}_{reg} = -\sum_{a=1}^{|L^{\mathrm{c}}|}\sum_{b=1}^{|L^{\mathrm{p}}|} y_{a,b}\log \hat{y}_{a,b}$




The three training objectives in NFC

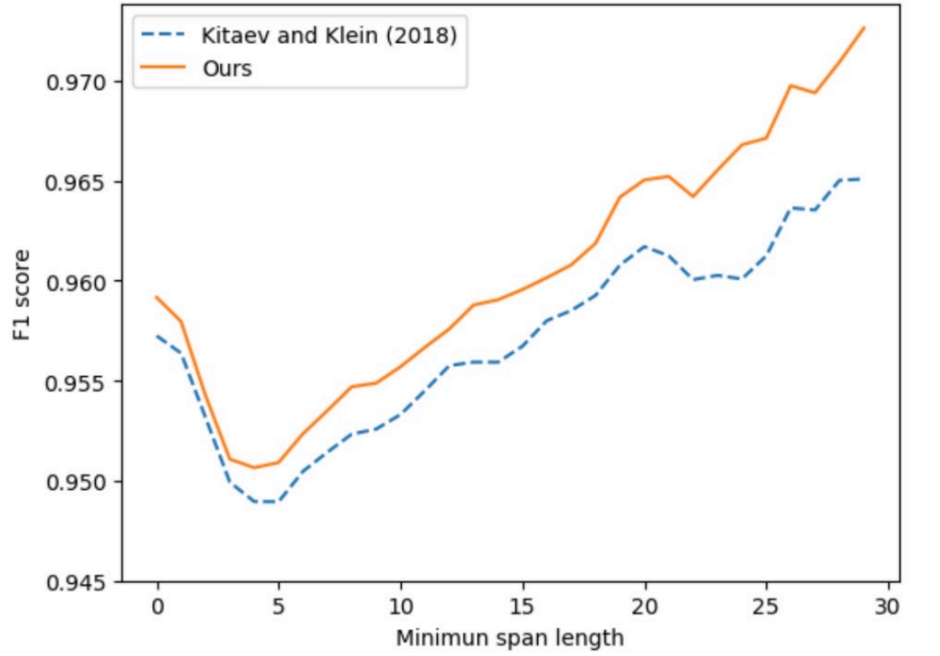
| Model | LR | LP | F1 |
|---|---|---|---|
| Liu and Zhang (2017) ◇ | - | - | 95.71 |
| Kitaev and Klein (2018) | 95.46 | 95.73 | 95.59 |
| Zhou and Zhao (2019) | 95.51 | 95.93 | 95.72 |
| Zhou and Zhao (2019) * | 95.70 | 95.98 | 95.84 |
| Zhang et al. (2020b) | 95.53 | 95.85 | 95.69 |
| Nguyen et al. (2020) | - | - | 95.48 |
| Tian et al. (2020) | 95.58 | 96.11 | 95.85 |
| This work | | | |
| Kitaev and Klein (2018) † | 95.56 | 95.89 | 95.72 |
| NFC w/o $\mathcal{L}_{reg}$ | 95.49 | 96.07 | 95.78 |
| NFC | **95.70** | **96.14** | **95.92** |

Model performance on PTB

| Model | LR | LP | F1 |
|---|---|---|---|
| Liu and Zhang (2017) ◇ | - | - | 91.81 |
| Kitaev and Klein (2018) | 91.55 | 91.96 | 91.75 |
| Zhang et al. (2020b) | 92.04 | 92.51 | 92.27 |
| Zhou and Zhao (2019) | 91.14 | 93.09 | 92.10 |
| Tian et al. (2020) | 92.14 | 92.25 | 92.20 |
| This work | | | |
| Kitaev and Klein (2018) † | 91.80 | 92.23 | 91.98 |
| NFC w/o $\mathcal{L}_{reg}$ | 91.87 | 92.40 | 92.13 |
| NFC | **92.17** | **92.45** | **92.31** |
| w/ External Dependency Supervision | | | |
| Zhou and Zhao (2019) * | 92.03 | 92.33 | 92.18 |
| Mrini et al. (2020)* | 91.85 | 93.45 | 92.64 |

Model performance on CTB

## In Domain Syntactic Parsing




F1 scores versus minimum constituent span length on PTB test set

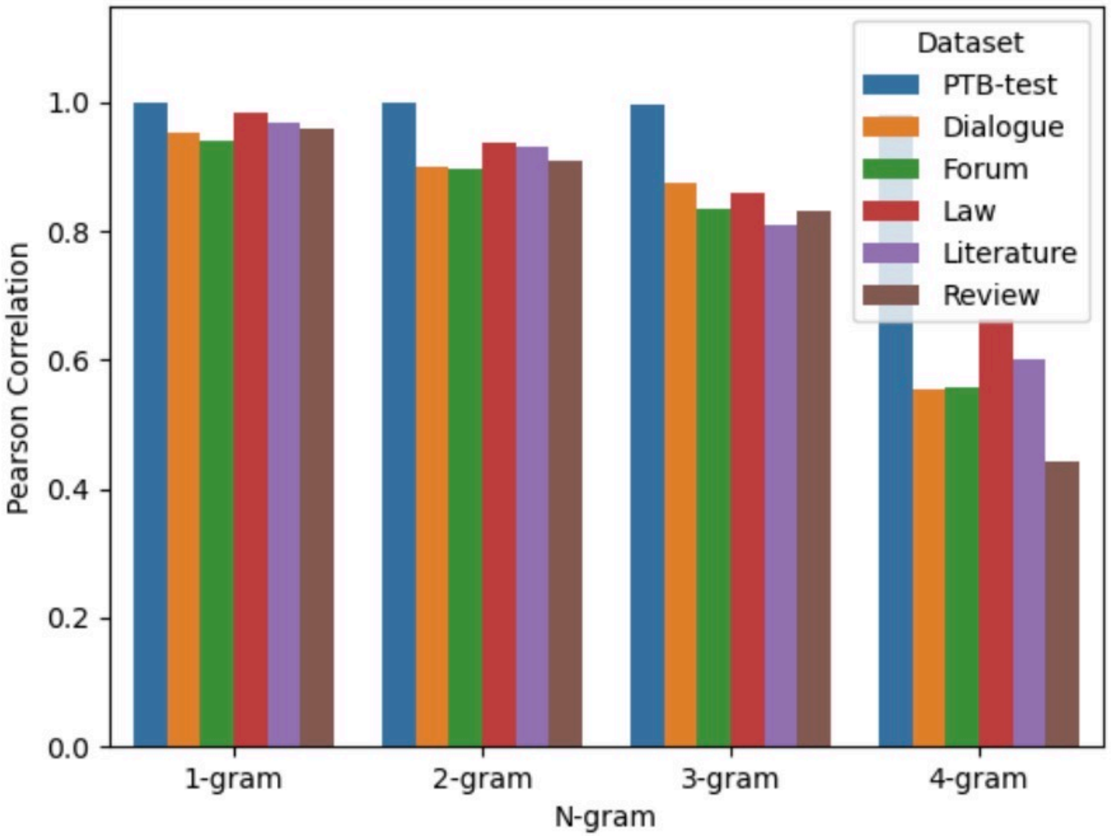
- NFC significantly outperform baseline when the minimum span length increases.

| Model | In-domain | Cross-domain | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | PTB | Bio | Dialogue | Forum | Law | Literature | Review | Avg |
| Liu and Zhang (2017) | 95.65 | 86.33 | 85.56 | 85.42 | 91.50 | 84.84 | 83.53 | 86.20 |
| Kitaev and Klein (2018) | 95.72 | 86.61 | 86.30 | 86.29 | 92.08 | 86.10 | 83.88 | 86.88 |
| NFC | **95.92** | 86.43 | **89.85** | **88.52** | **95.43** | **90.75** | **88.10** | **89.85** |

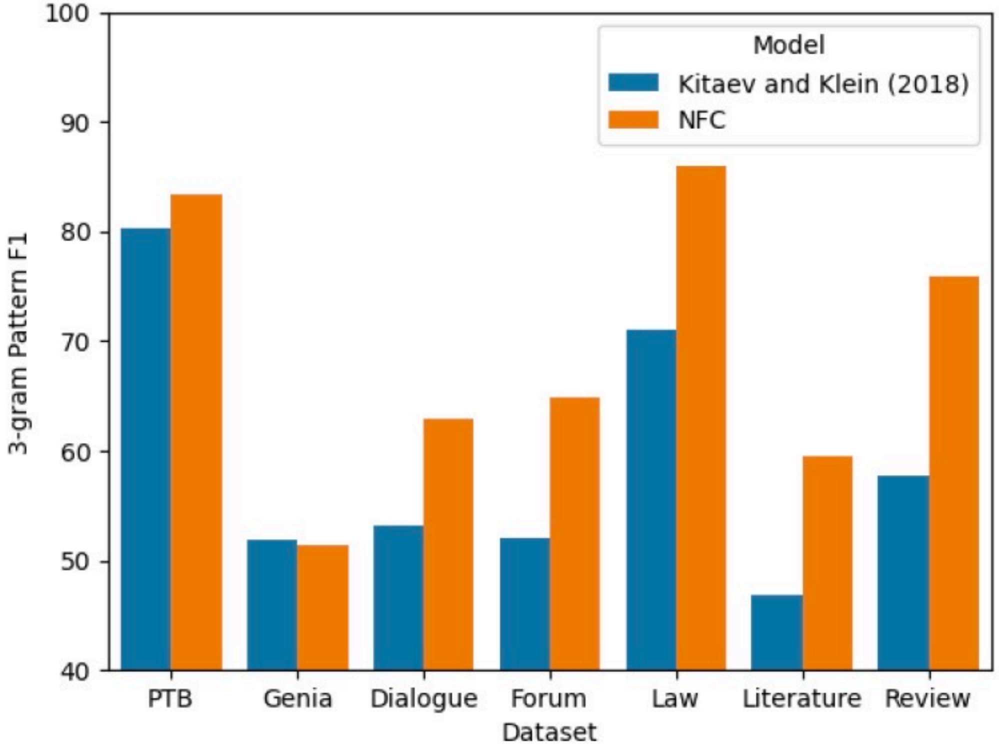
Zero-shot performance on cross-domain test set

| Model | Rich resource | | | | Low Resource | | | | Avg |
|---|---|---|---|---|---|---|---|---|---|
| | French | German | Korean | Avg | Hungarian | Basque | Polish | Avg | |
| Kitaev and Klein (2018) | 87.42 | 90.20 | 88.80 | 88.81 | 94.90 | 91.63 | **96.36** | 94.30 | 91.55 |
| Nguyen et al. (2020) | 86.69 | 90.28 | 88.71 | 88.56 | 94.24 | **92.02** | 96.14 | 94.13 | 91.34 |
| Kitaev and Klein (2018) † | 87.38 | 90.25 | 88.91 | 88.85 | 94.56 | 91.66 | 96.14 | 94.12 | 91.48 |
| NFC | **87.51** | **90.43** | **89.07** | **89.00** | **94.95** | 91.73 | 96.33 | **94.34** | **91.67** |

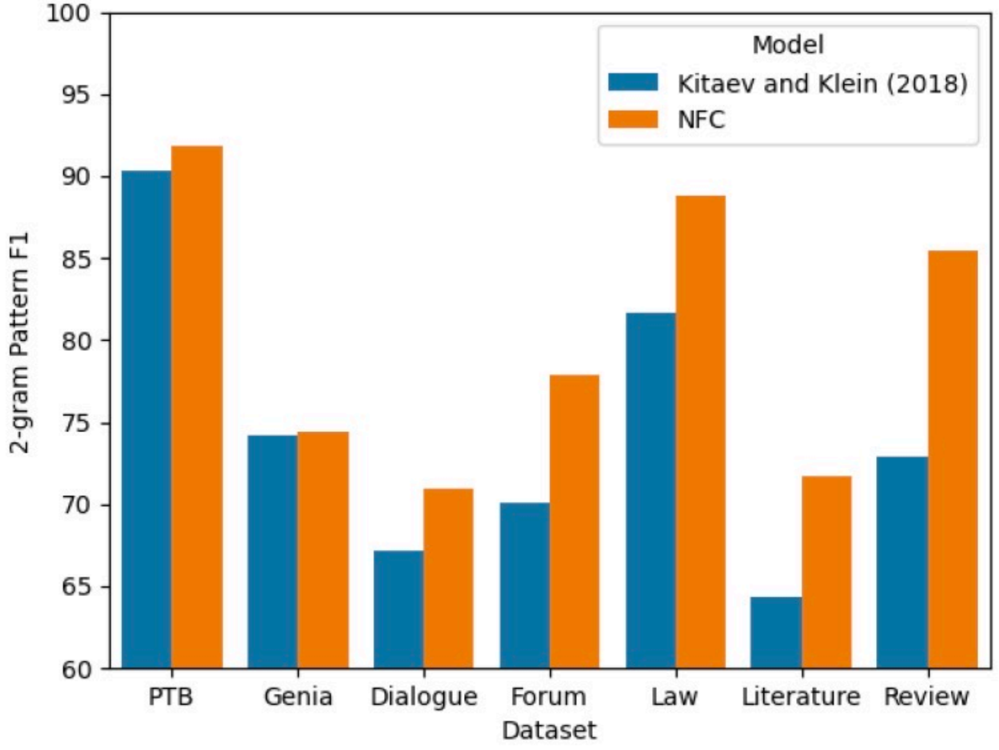
Multi-lingual performance




- Pearson correlation of n-gram pattern distribution between PTB training set and different test set.




(a) F1 scores measured by 3-gram pattern.

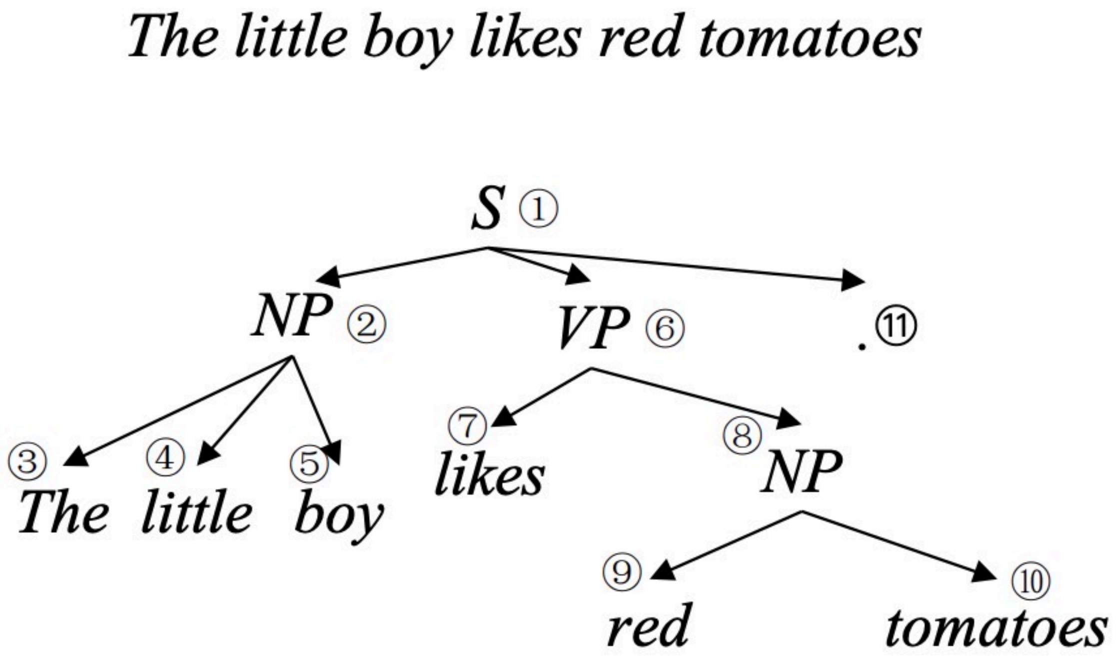



(b) F1 scores measured by 2-gram pattern.

- NFC significantly outperform baseline measured by pattern-level f1.

# Learned Incremental Representations for Parsing

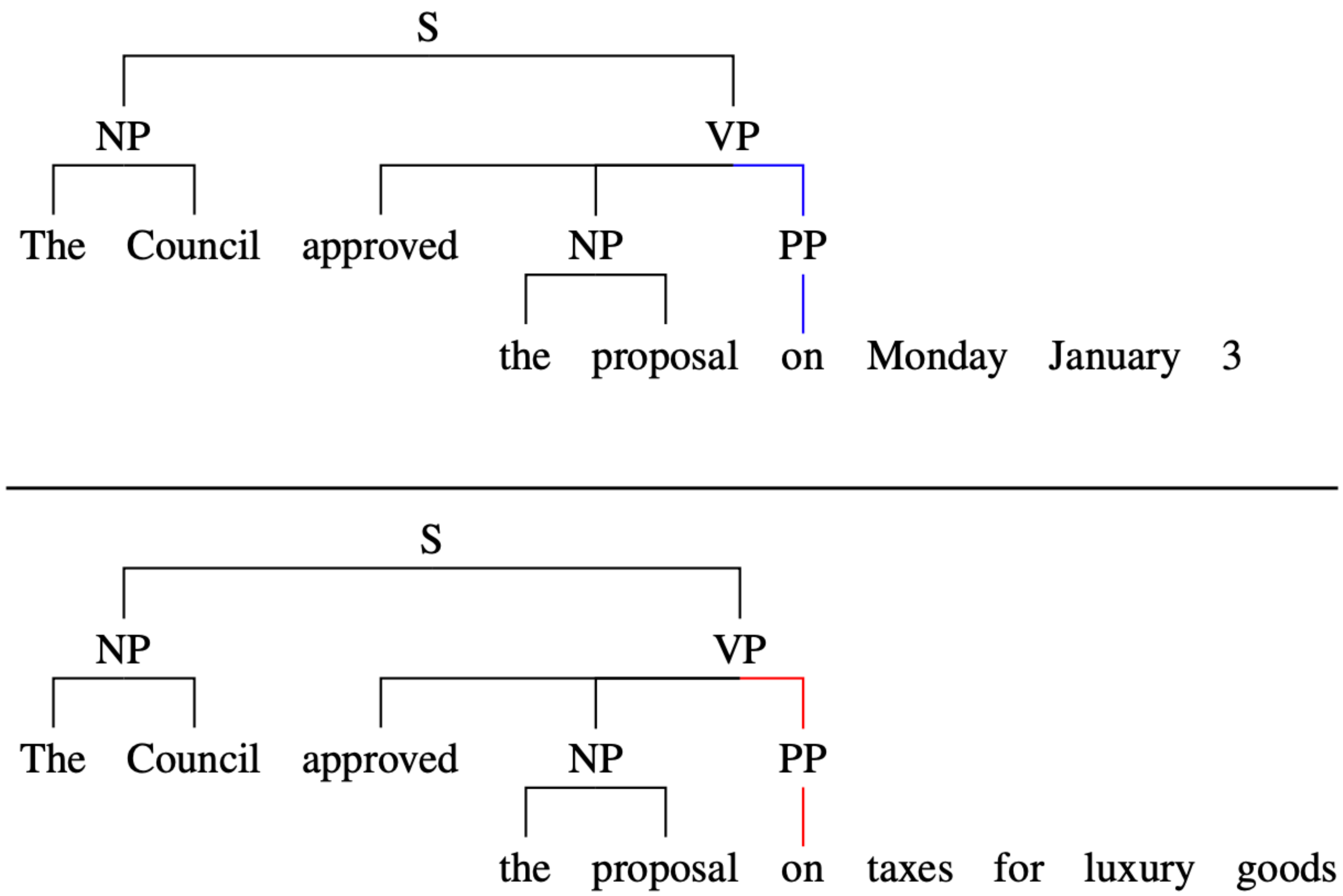
**Nikita Kitaev** **Thomas Lu** **Dan Klein**
Computer Science Division
University of California, Berkeley
{kitaev,tlu2000,klein}@berkeley.edu



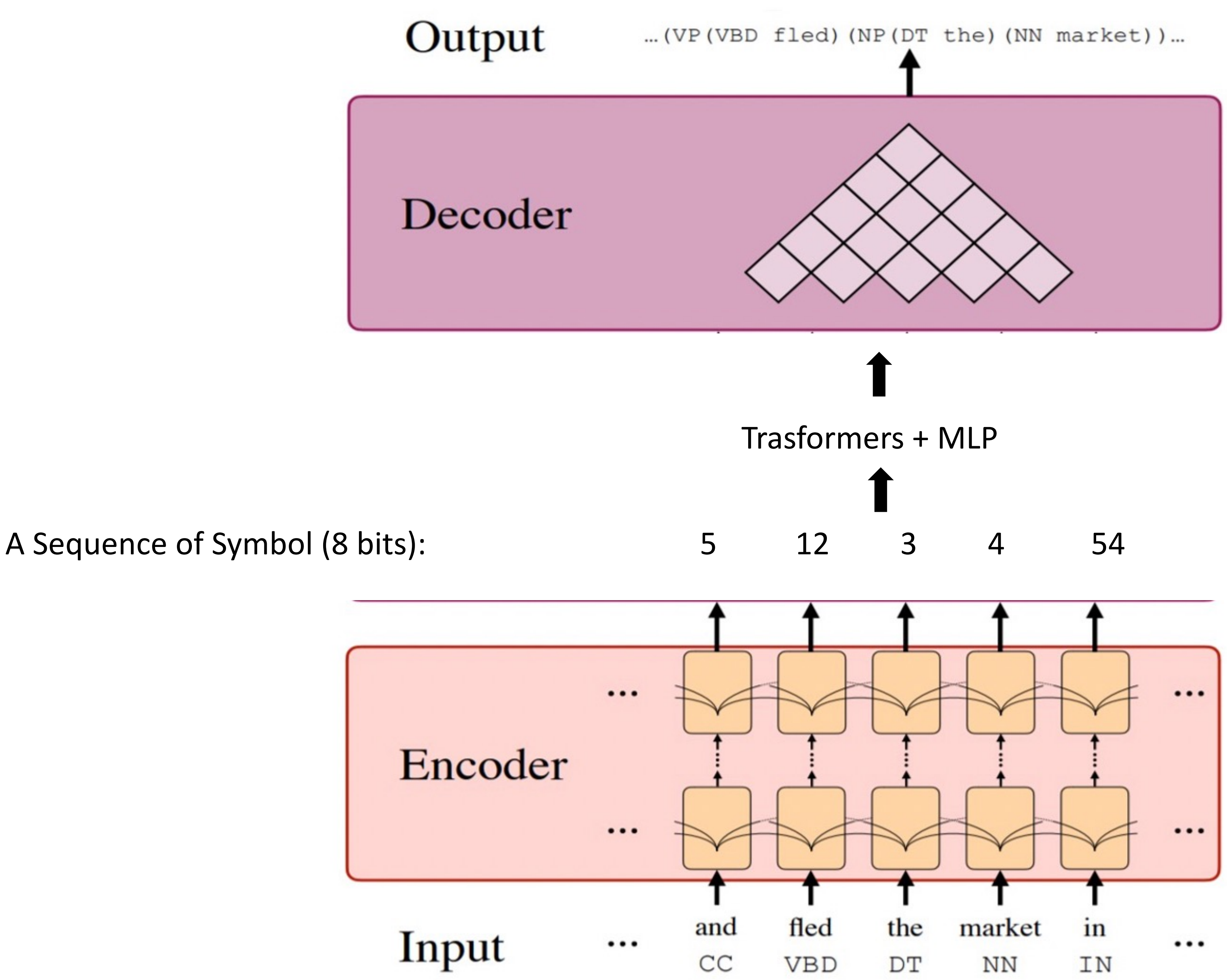

Syntactic tree

| stack | buffer | action | node |
|---:|---|:---:|:---:|
| [] | [The little ...] | NT-S | ① |
| [(S] | [The little ...] | NT-NP | ② |
| [(S (NP] | [The little ...] | SHIFT | ③ |
| [... (NP The] | [little boy ...] | SHIFT | ④ |
| [... The little] | [boy likes ...] | SHIFT | ⑤ |
| [... little boy] | [likes red ...] | REDUCE | / |
| ... | ... | ... | ... |

Shift-reduce parser

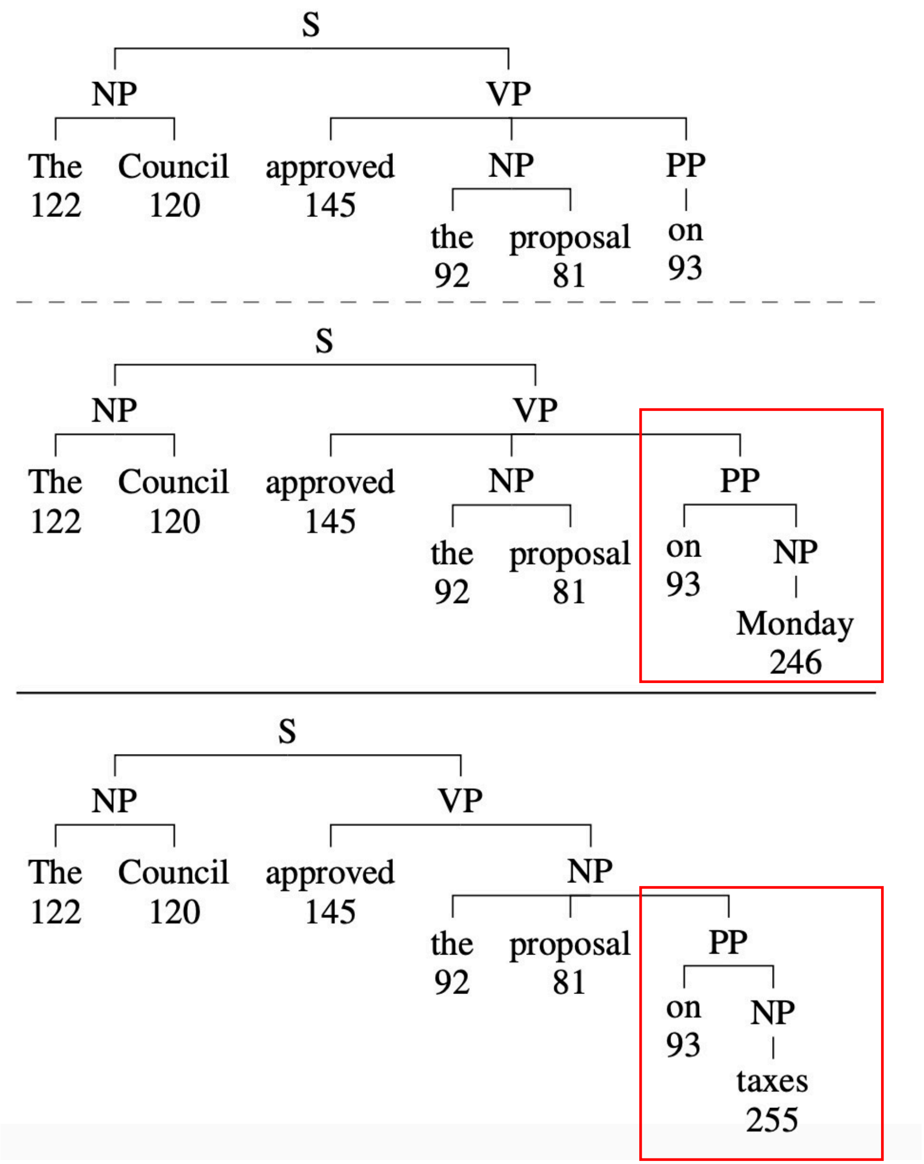


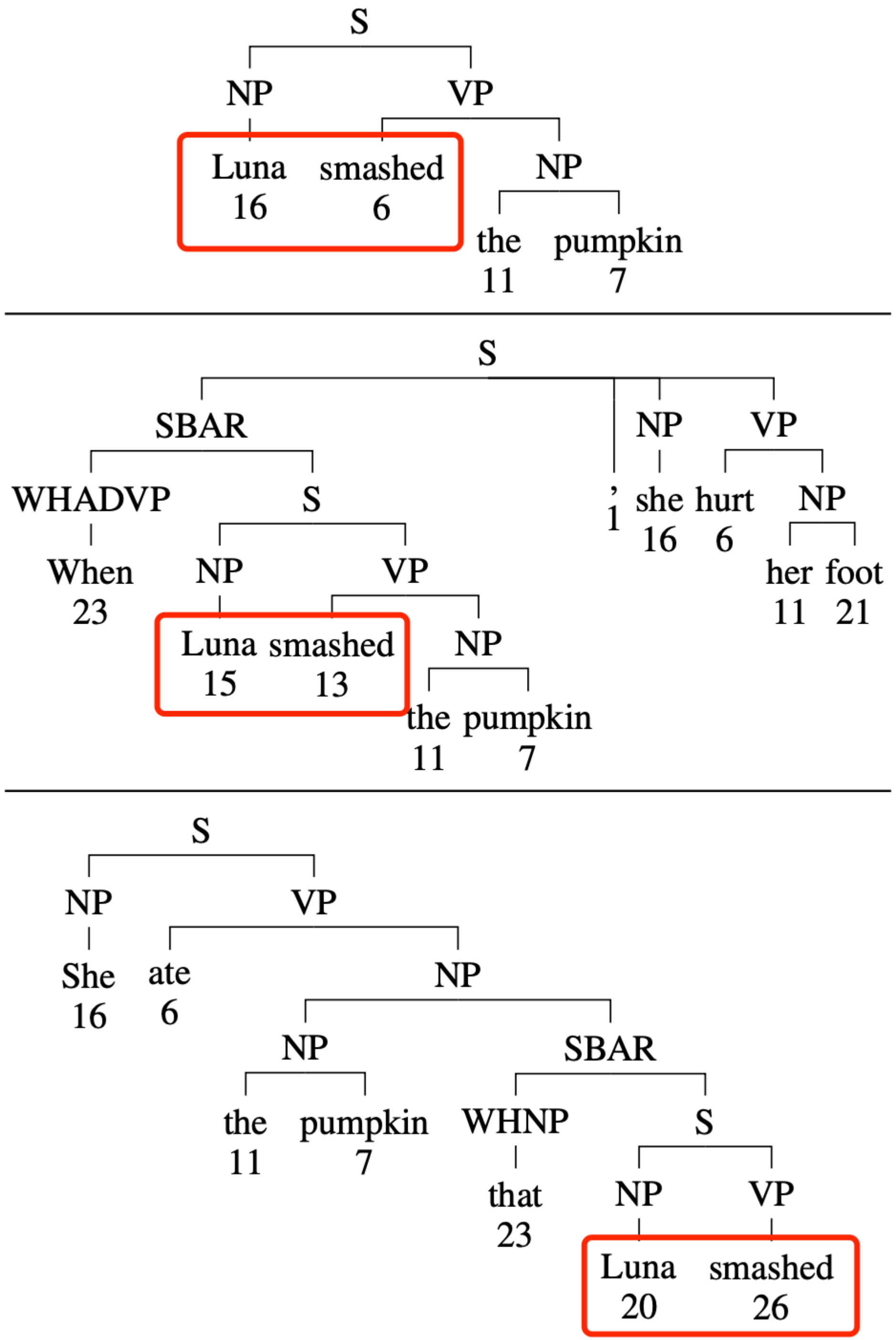

whether the preposition “on” attaches to noun “proposal” or the verb “approved.”


Output
…(VP(VBD fled)(NP(DT the)(NN market))…
Decoder
Trasformers + MLP
A Sequence of Symbol (8 bits):
5 12 3 4 54
Encoder
Input
and CC
fled VBD
the DT
market NN
in IN

| | Encoder Type | | |
| --- | --- | --- | --- |
| | Bi (↔) | | Uni (→) |
| Representation | BERT | GPT-2 | GPT-2 |
| Span Classification (Kitaev et al., 2019) | 95.59 | **95.10**† | 93.95† |
| Attach-Juxtapose (Yang and Deng, 2020) | **95.79** | 94.53† | 87.66† |
| Learned (This work) | 95.55 | – | **94.97** |

Model Performance on PTB.


S
NP
VP
The 122
Council 120
approved 145
NP
PP
the 92
proposal 81
on 93
S
NP
VP
The 122
Council 120
approved 145
NP
PP
the 92
proposal 81
on 93
NP
Monday 246
S
NP
VP
The 122
Council 120
approved 145
NP
the 92
proposal 81
PP
on 93
NP
taxes 255




- In main clauses, subjects and verbs are assigned symbols 16 and 6.
- Subordinate clauses, however, tend to use alternate symbols 15 and 13 for subject nouns and verb, respectively.
- Relative clauses use 20 and 26.

Tags capture structural context beyond the current word!

# Thank You