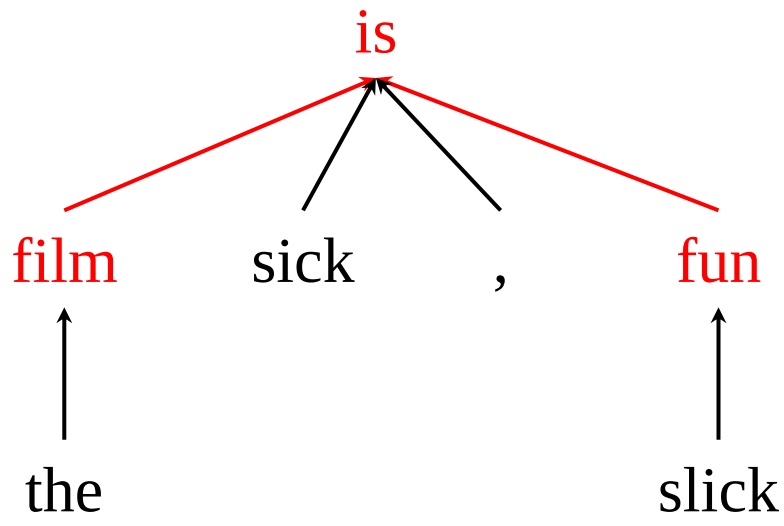
Dependency Forest for Sentiment Analysis

Zhaopeng Tu, Wenbin Jiang, Qun Liu, Shouxun Lin

Key Laboratory of Intelligent Information Processing
Institute of Computing Technology

INSTITUTE OF COMPUTING TECHNOLOGY 中科院计算所

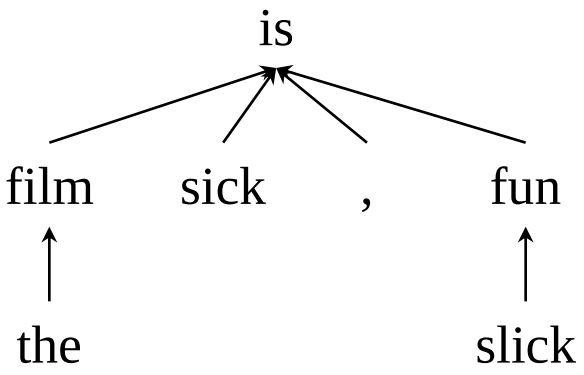
Dependency Grammars in SA

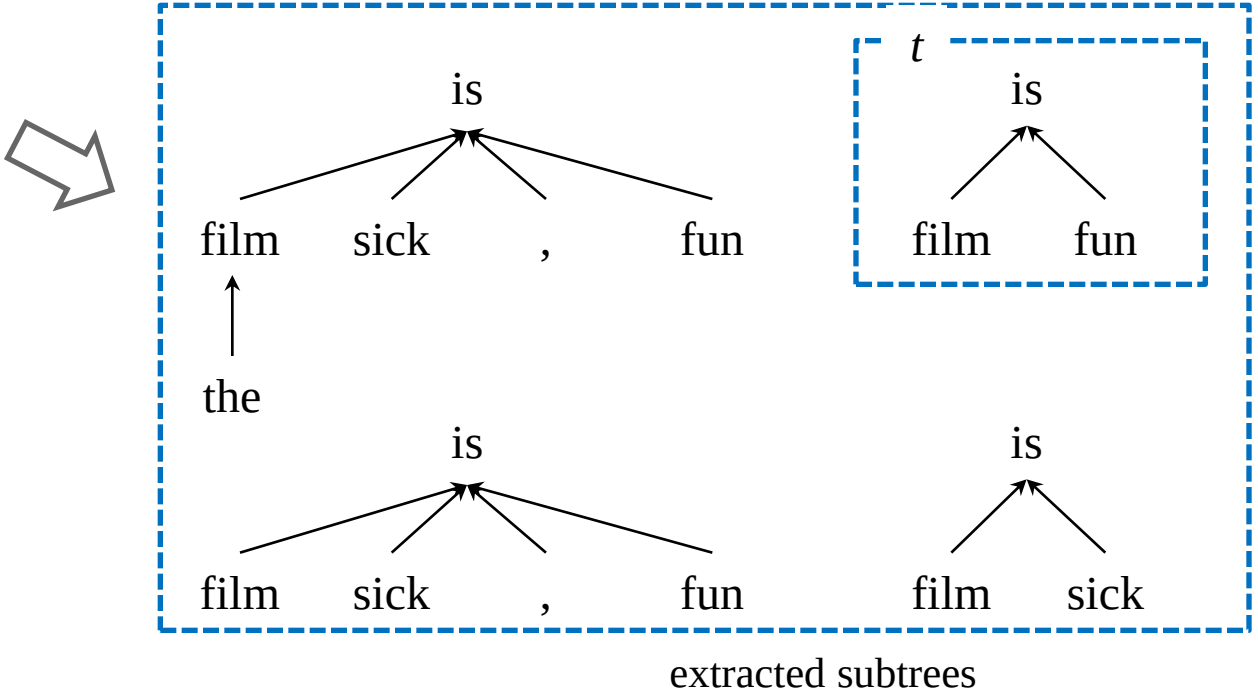


(Matsumoto et al., 2005; Joshi et al., 2009; Liu et al., 2009)

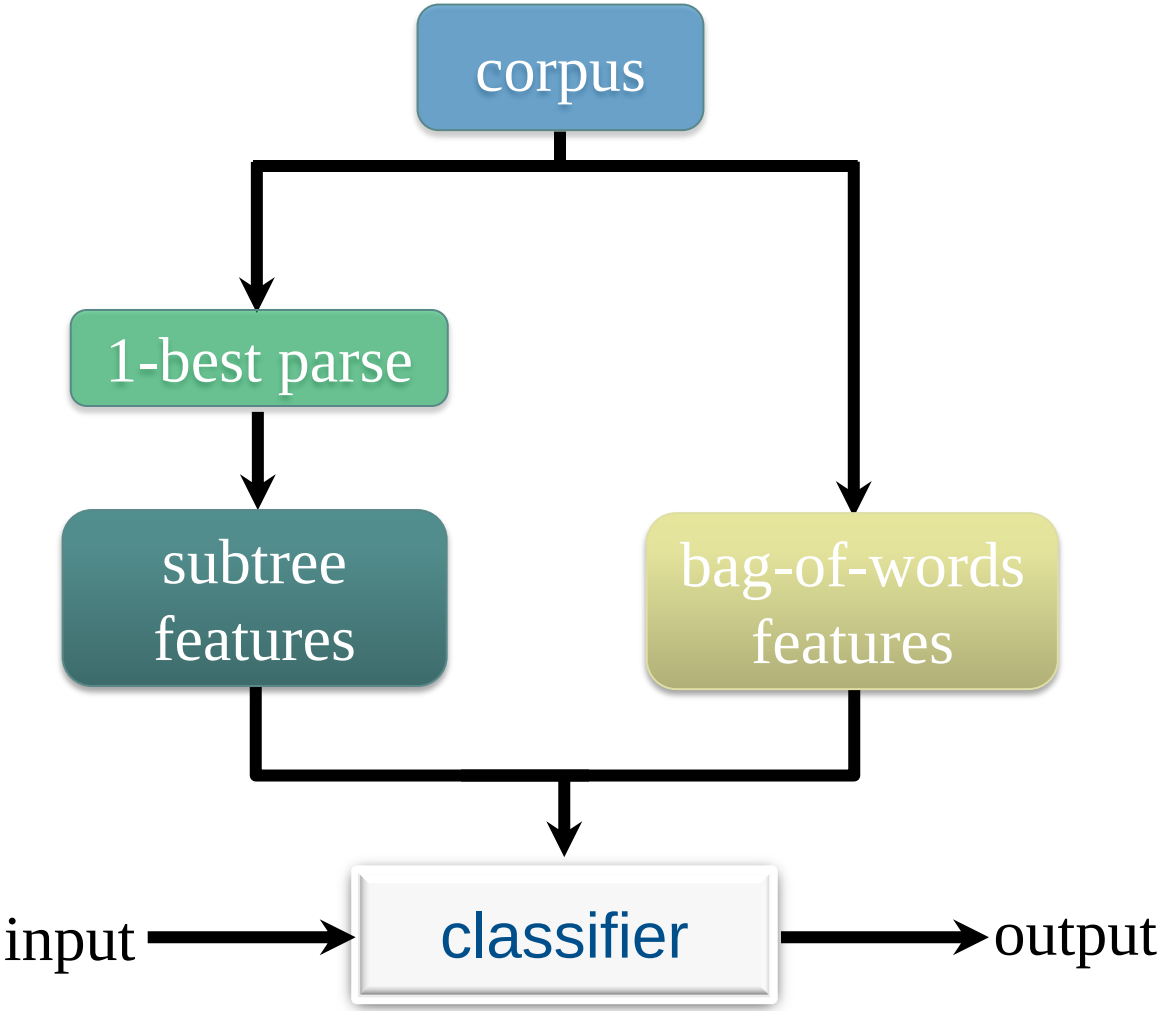
Tree-based Features Extraction

(Matsumoto et al., 2005)

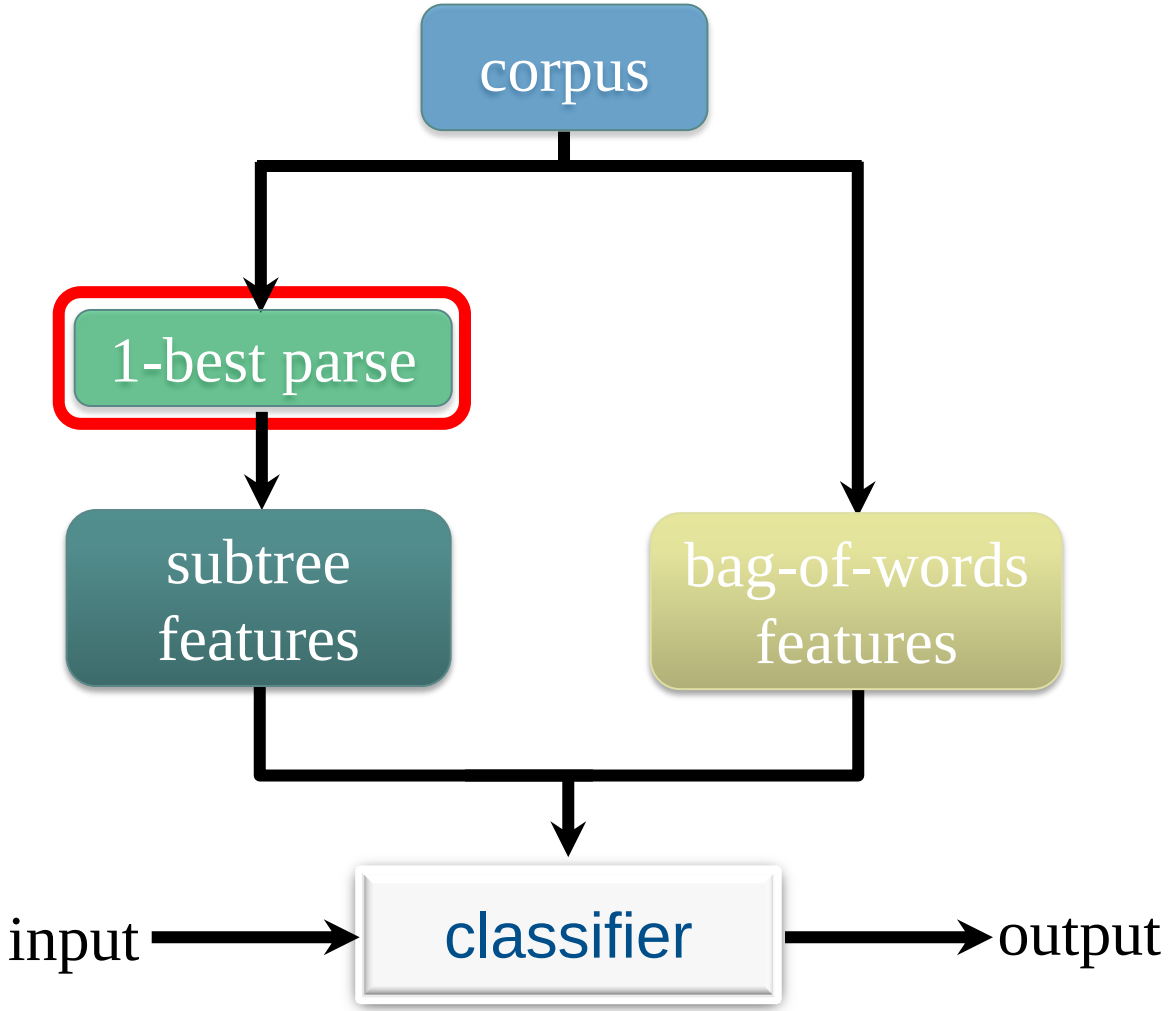




Pipeline


corpus
1-best parse
subtree features
bag-of-words features
input
classifier
output

Pipeline

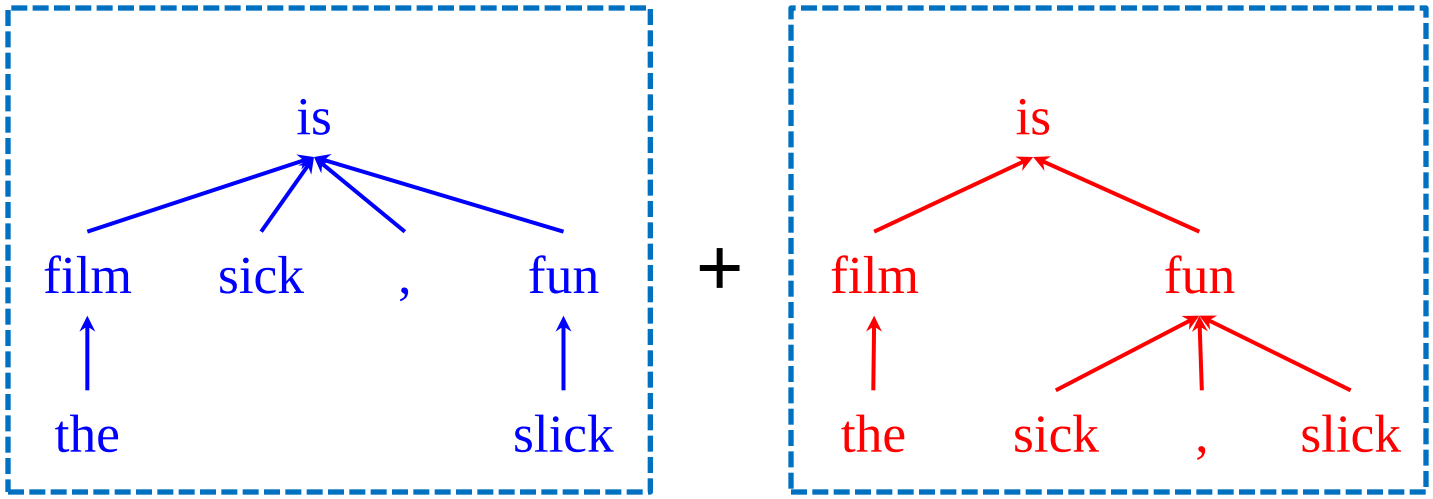

corpus
1-best parse
subtree features
bag-of-words features
input
classifier
output

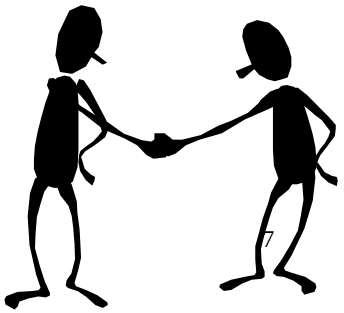
Challenges

- Tree-based approach faces the major challenges:
  - vulnerable to parsing error

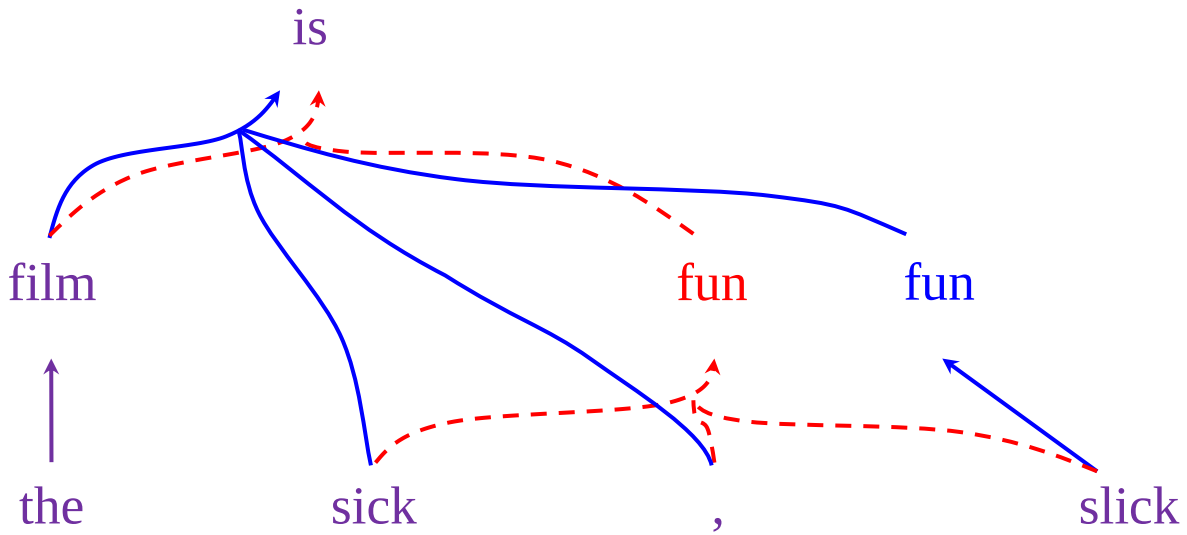
Solution

- Dependency forest provides an elegant solution to this problem (Tu et al, 2010)

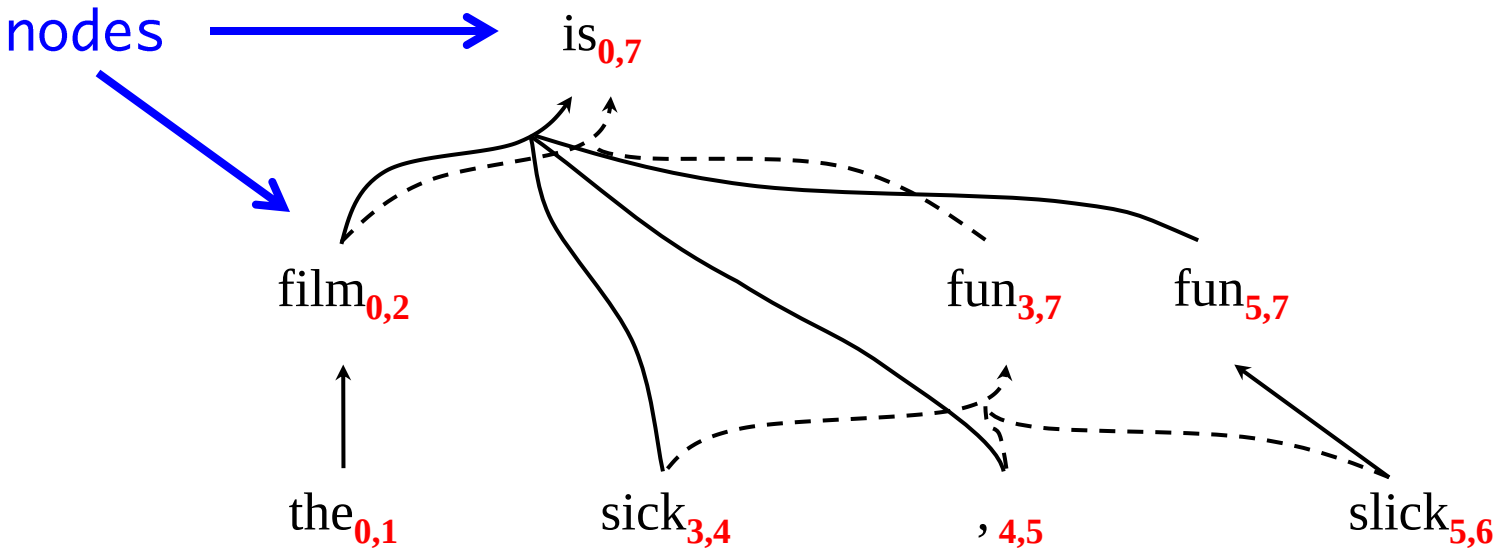




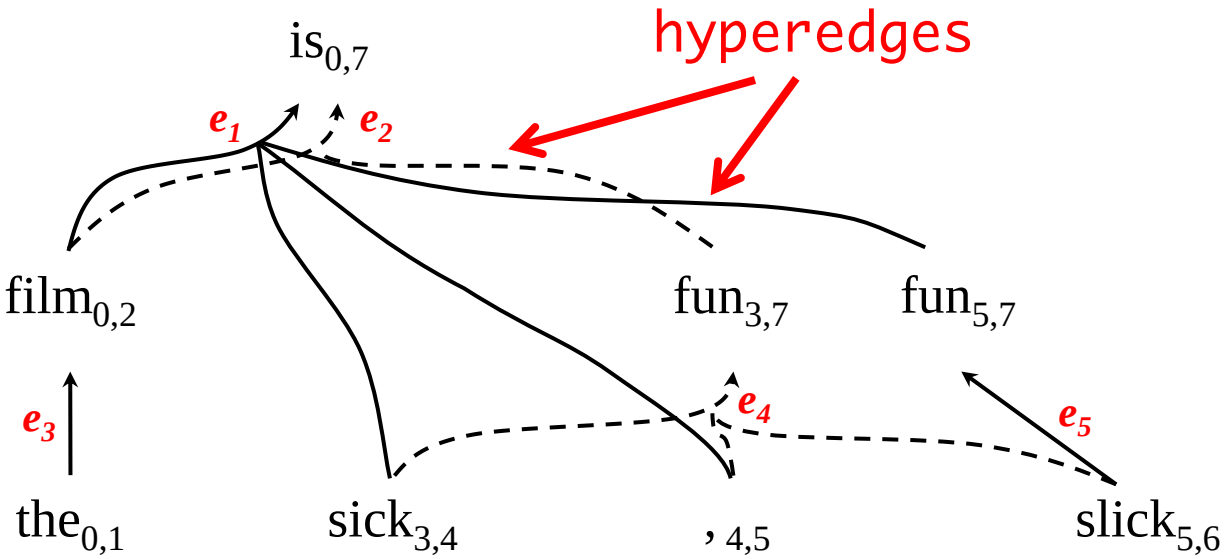
Dependency Forest



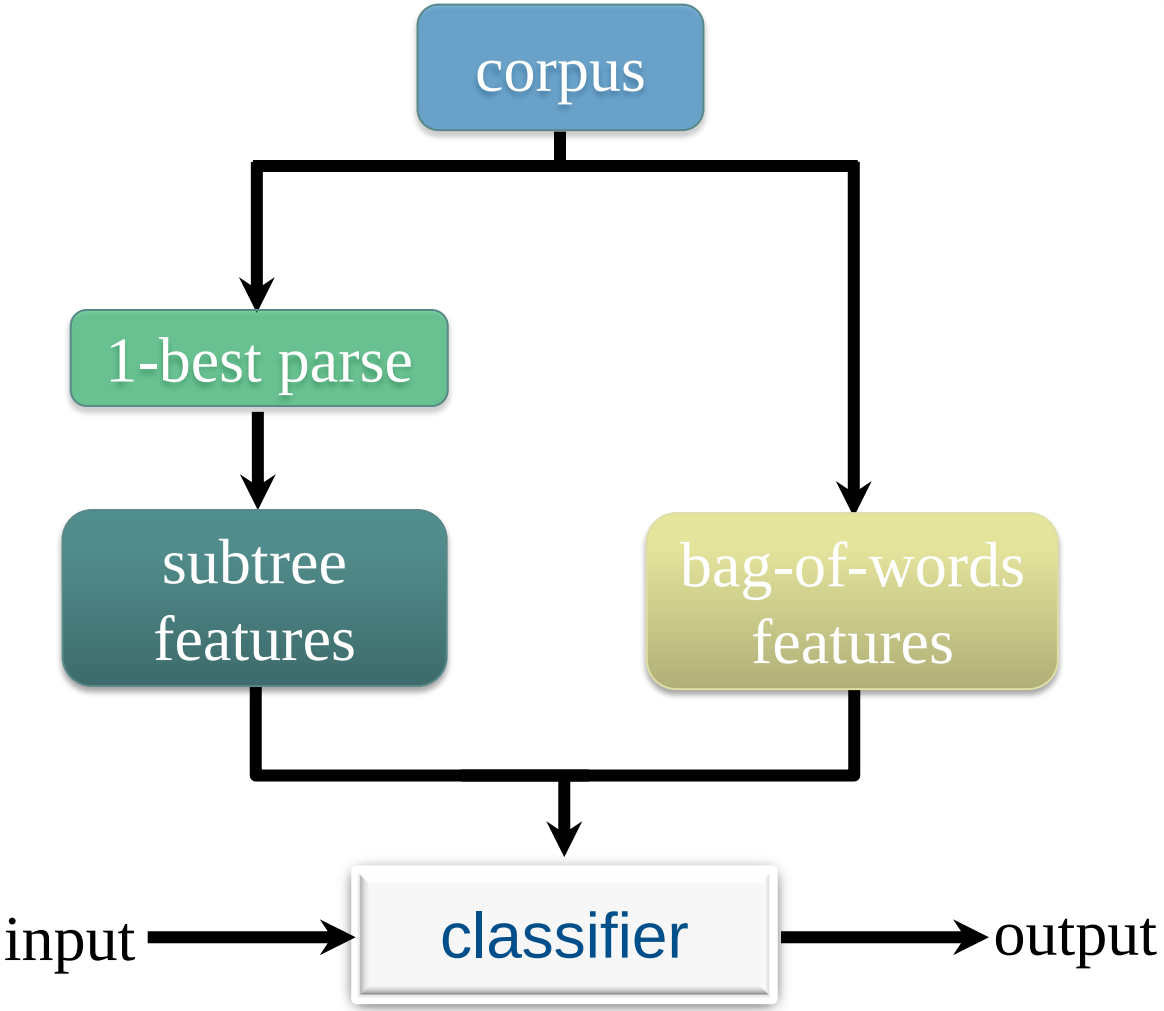
Dependency Forest



Dependency Forest

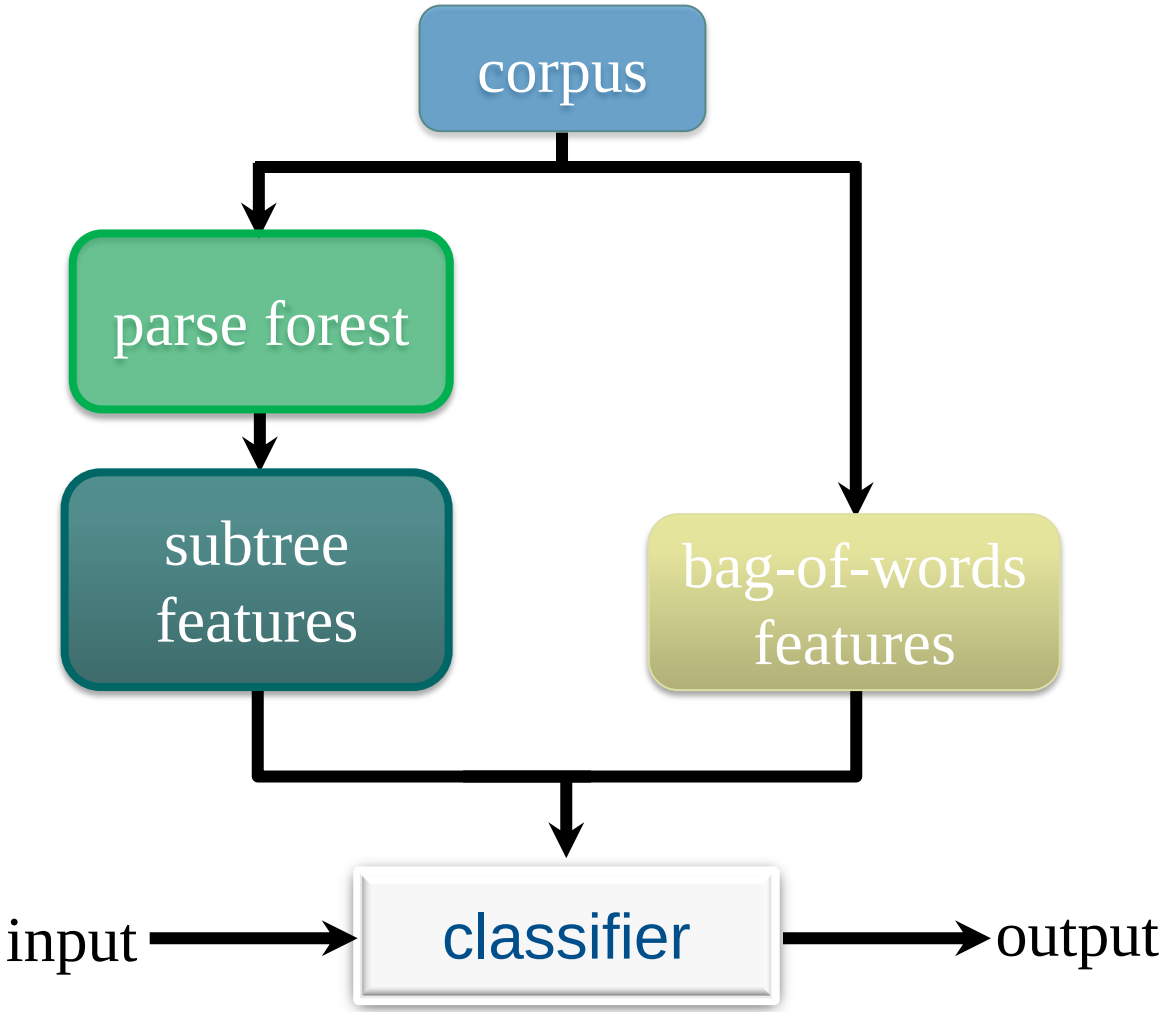


Pipeline



tree-based

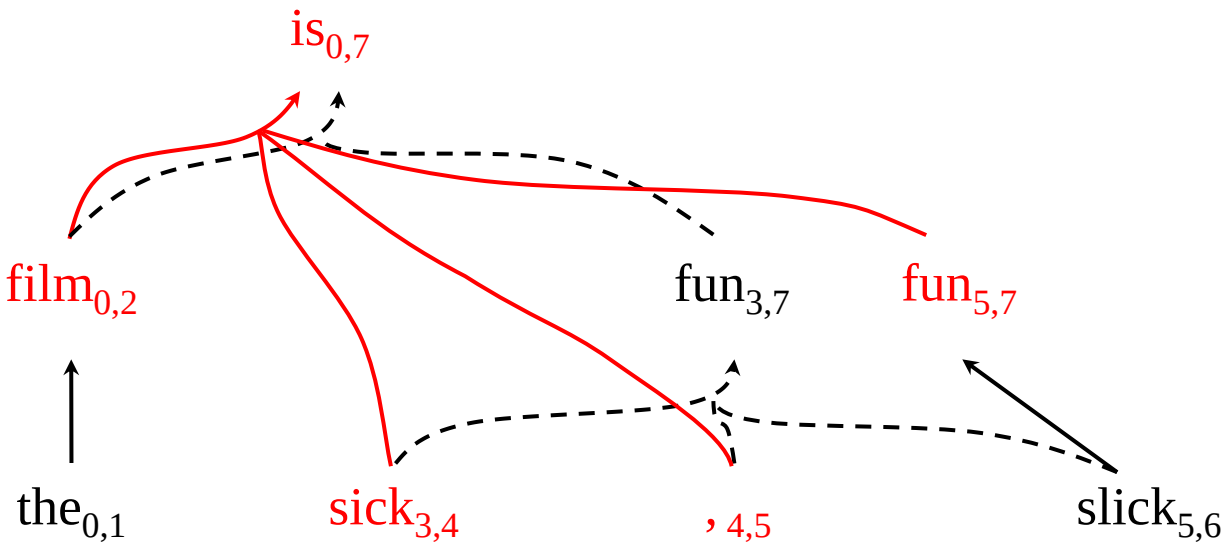
Pipeline



forest-based

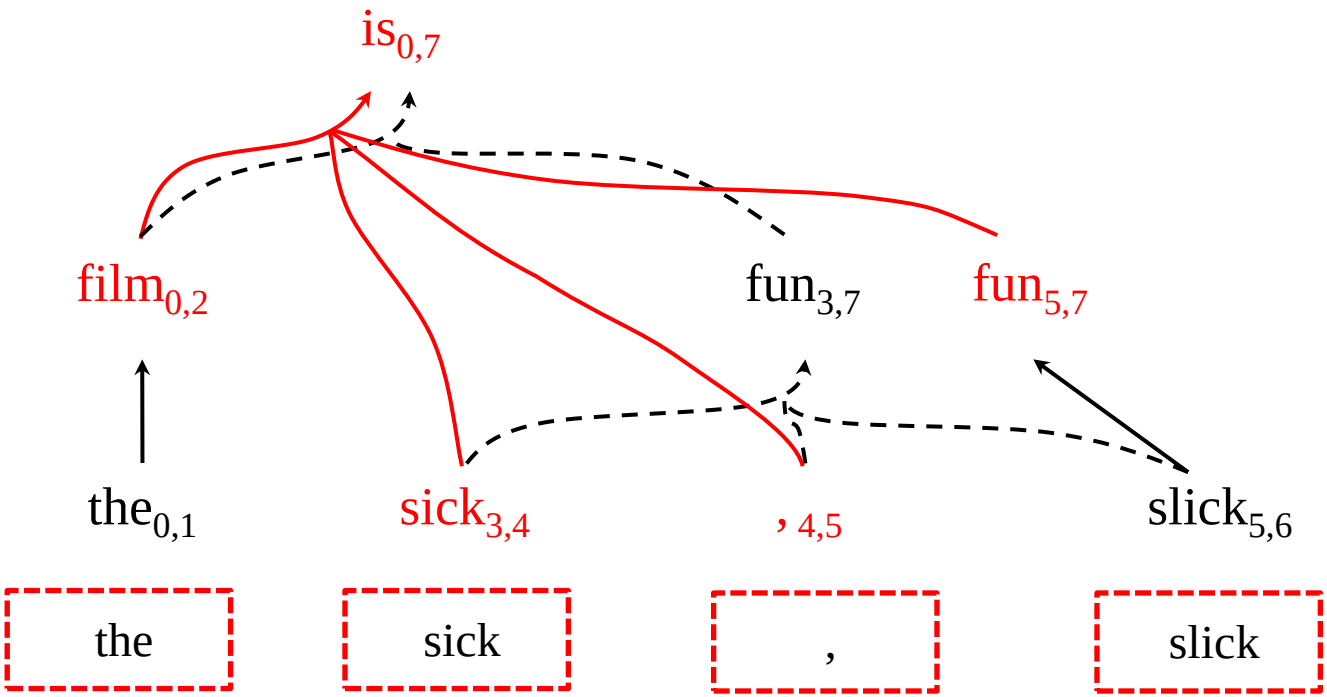
Forest-based Feature Extraction

bottom-up style



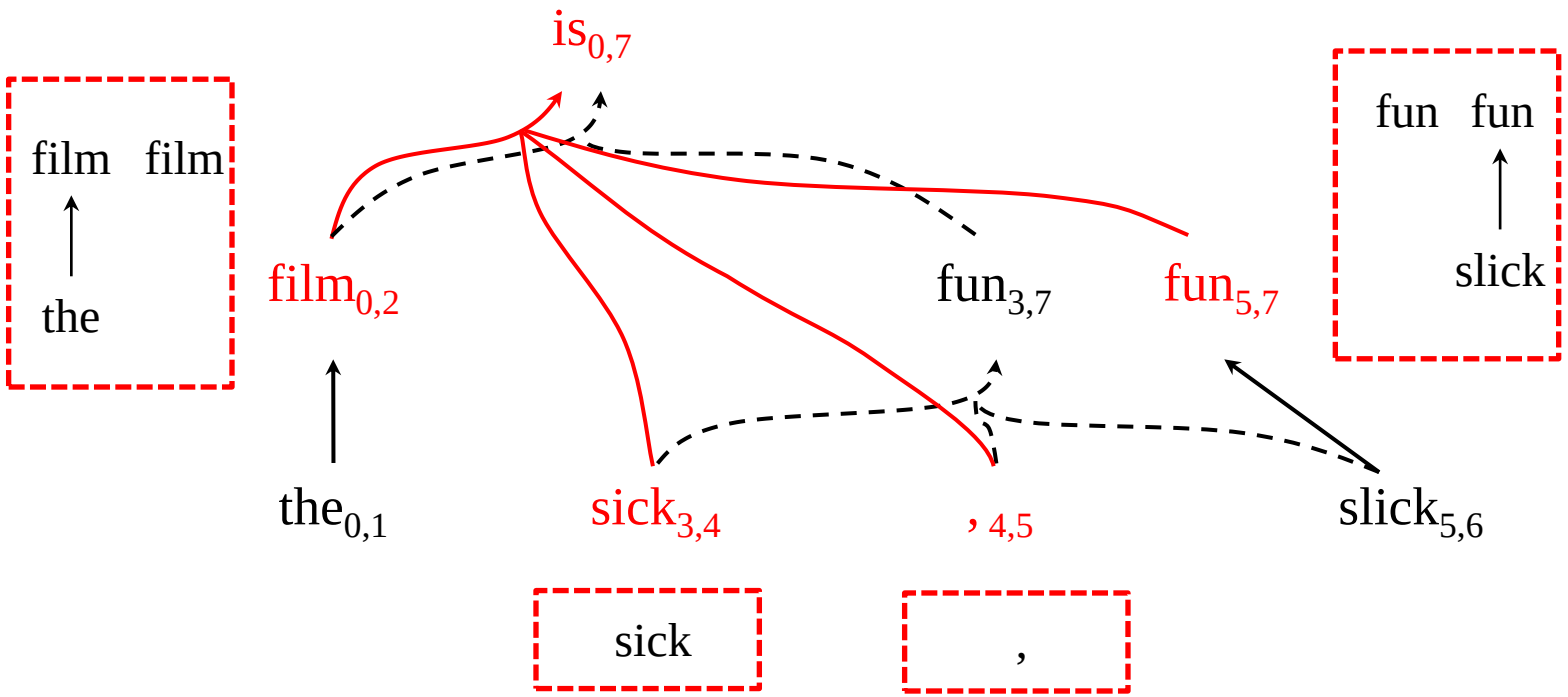
Forest-based Feature Extraction

bottom-up style



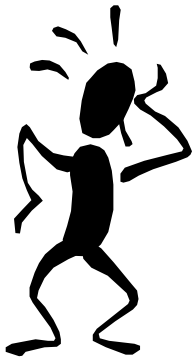
Forest-based Feature Extraction

bottom-up style

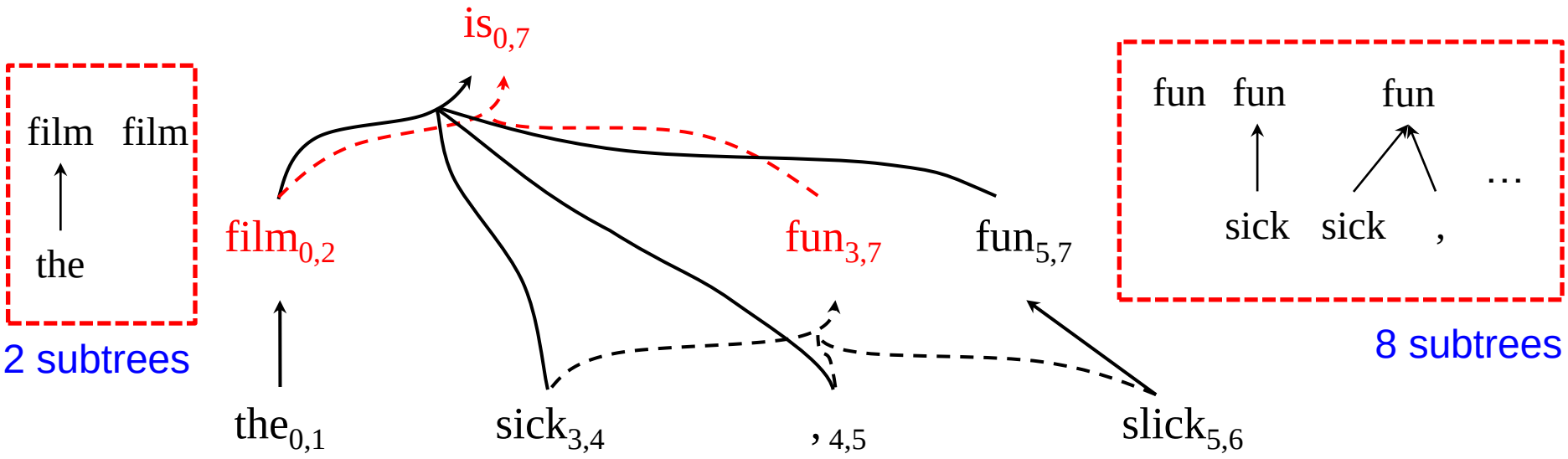


Difficulty in Finding Subtrees

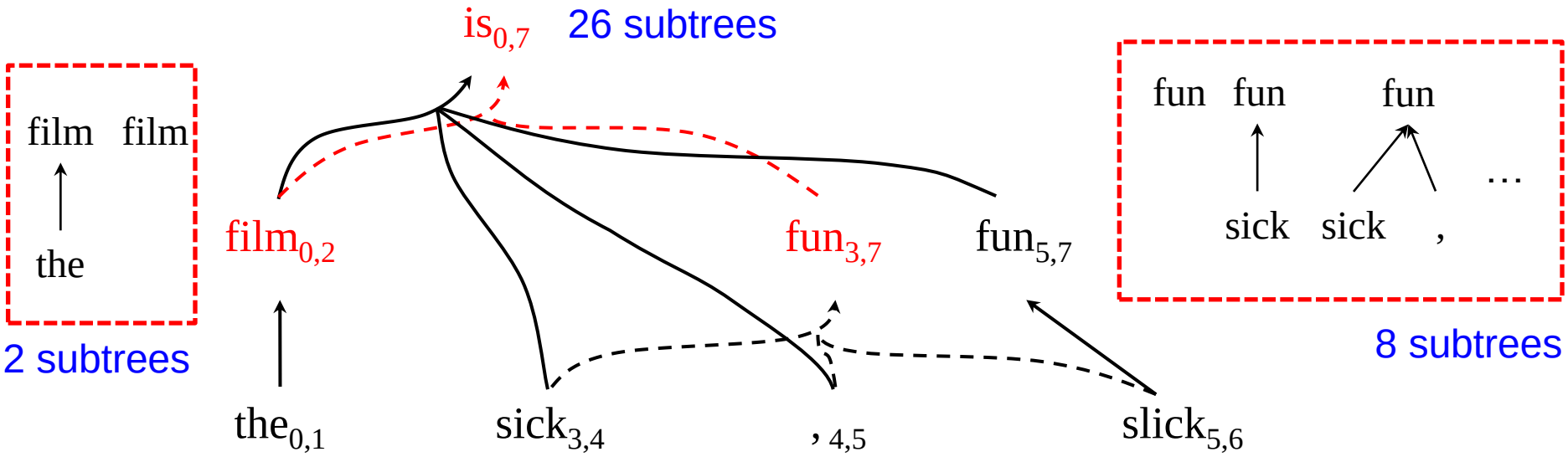
too many choices



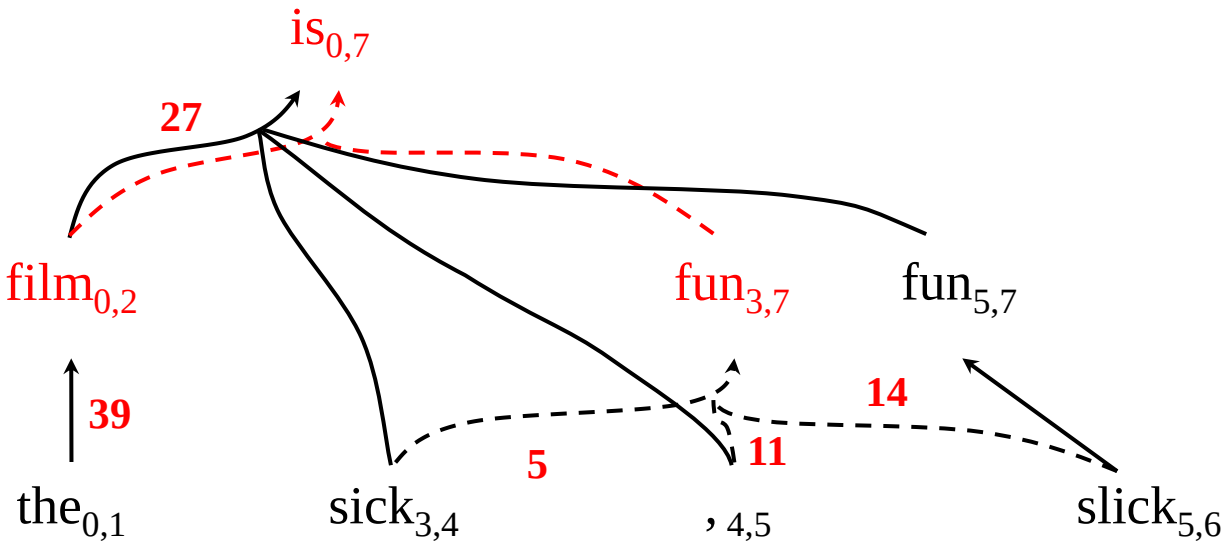
Forest-based Feature Extraction



Forest-based Feature Extraction

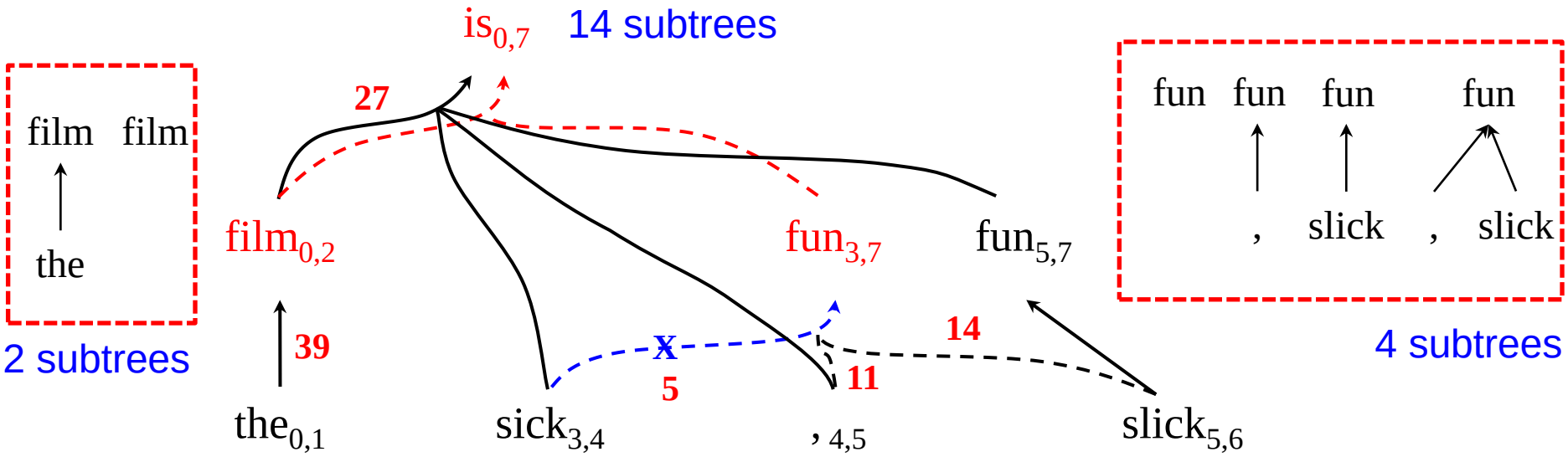


Forest-based Feature Extraction



gathering the occurrences of edges

Forest-based Feature Extraction



gathering the occurrences of edges

Experiments

Setup

- Movie Review Corpus (Pang et al., 2004)
  - 1000 positive documents and 1000 negative documents
- Unigrams:
  - appear at least 4 times
- Tree-based Subtrees:
  - appear in at least 10 distinct sentences
- Forest-based Subtrees:
  - appear in at least 10 distinct sentences
  - the fractional count should be not lower than 0.01 (Tu et al., 2010)

Results

Features	Number	Accuracy	Time
Unigram	17,704	86.2	–
Subtree$_{1-best}$	12,282	74.2	0.33
Unigram + Subtree$_{1-best}$	29,986	90.3†	–
Subtree$_{100-best}$	24,006	81.9	35.47
Unigram + Subtree$_{100-best}$	41,710	90.2†	–
Subtree$_{forest}$	18,968	81.2	6.93
Unigram + Subtree$_{forest}$	36,674	**91.6**‡	–
Pang et al. [18]	–	87.1	–
Ng et al. [17]	–	90.5	–
Yessenalina et al. [23]	–	91.8	–

Forest VS Bag-of-words

Features	Number	Accuracy	Time
Unigram	17,704	86.2	–
Subtree$_{1-best}$	12,282	74.2	0.33
Unigram + Subtree$_{1-best}$	29,986	90.3†	–
Subtree$_{100-best}$	24,006	81.9	35.47
Unigram + Subtree$_{100-best}$	41,710	90.2†	–
Subtree$_{forest}$	18,968	81.2	6.93
Unigram + Subtree$_{forest}$	36,674	**91.6**‡	–
Pang et al. [18]	–	87.1	–
Ng et al. [17]	–	90.5	–
Yessenalina et al. [23]	–	91.8	–

Forest VS 1-best tree

Features	Number	Accuracy	Time
Unigram	17,704	86.2	–
Subtree$_{1-best}$	12,282	74.2	0.33
Unigram + Subtree$_{1-best}$	29,986	90.3†	–
Subtree$_{100-best}$	24,006	81.9	35.47
Unigram + Subtree$_{100-best}$	41,710	90.2†	–
Subtree$_{forest}$	18,968	81.2	6.93
Unigram + Subtree$_{forest}$	36,674	**91.6**‡	–
Pang et al. [18]	–	87.1	–
Ng et al. [17]	–	90.5	–
Yessenalina et al. [23]	–	91.8	–

Forest VS *k*-best trees

Features	Number	Accuracy	Time
Unigram	17,704	86.2	–
Subtree$_{1-best}$	12,282	74.2	0.33
Unigram + Subtree$_{1-best}$	29,986	90.3†	–
Subtree$_{100-best}$	24,006	81.9	35.47
Unigram + Subtree$_{100-best}$	41,710	90.2†	–
Subtree$_{forest}$	18,968	81.2	6.93
Unigram + Subtree$_{forest}$	36,674	**91.6**‡	–
Pang et al. [18]	–	87.1	–
Ng et al. [17]	–	90.5	–
Yessenalina et al. [23]	–	91.8	–

Forest VS State-of-the-art

Features	Number	Accuracy	Time
Unigram	17,704	86.2	–
Subtree$_{1-best}$	12,282	74.2	0.33
Unigram + Subtree$_{1-best}$	29,986	90.3†	–
Subtree$_{100-best}$	24,006	81.9	35.47
Unigram + Subtree$_{100-best}$	41,710	90.2†	–
Subtree$_{forest}$	18,968	81.2	6.93
Unigram + Subtree$_{forest}$	36,674	**91.6**‡	–
Pang et al. [18]	–	87.1	–
Ng et al. [17]	–	90.5	–
Yessenalina et al. [23]	–	91.8	–

Conclusion and Future Work

- dependency forest provides an elegant solution to the problem of *parsing error propagation*
- very simple idea, but works very well in practice
  - 5.4 points in accuracy better than bag-of-words
  - 1.3 points in accuracy better than 1-best trees

Forest offers more alternatives.

Thank you!

Thanks to the anonymous reviewers.