Université de Montréal

GOFAISUM
A Symbolic Summarizer

Fabrizio Gotti, Guy Lapalme
Université de Montréal

Luka Nerima, Éric Wehrli
Université de Genève

rali

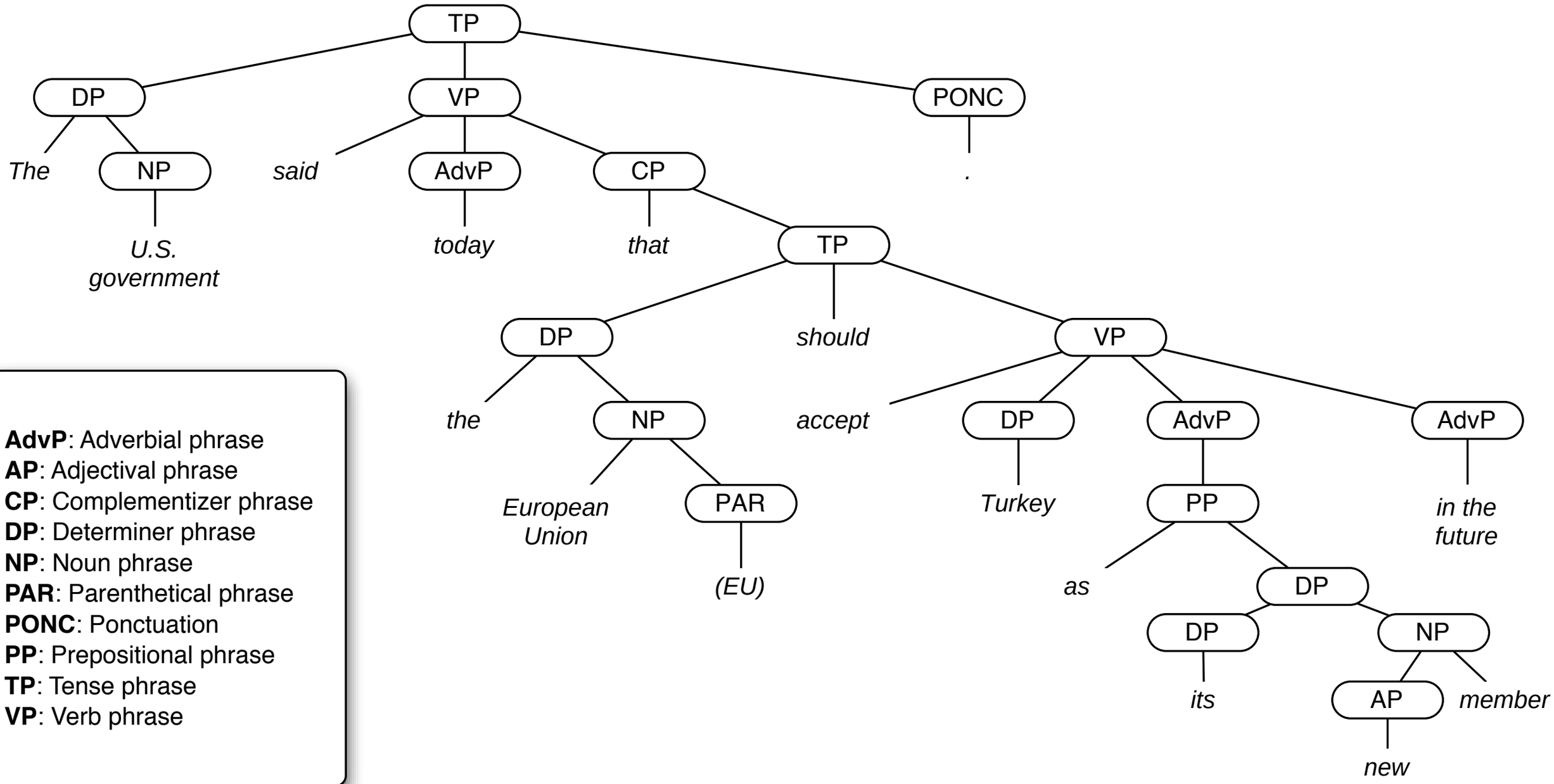
Originality of our approach

- Symbolic approach
  - Syntactic parser that produces an XML file
  - Tree transformations using only XSLT rules (700 lines) no Java, no C++, no Perl, no Python...
- No *outside* language information
  - no gazeeter, no Wordnet
  - ROUGE only used for evaluation not within the system itself

rali

FIPS output for

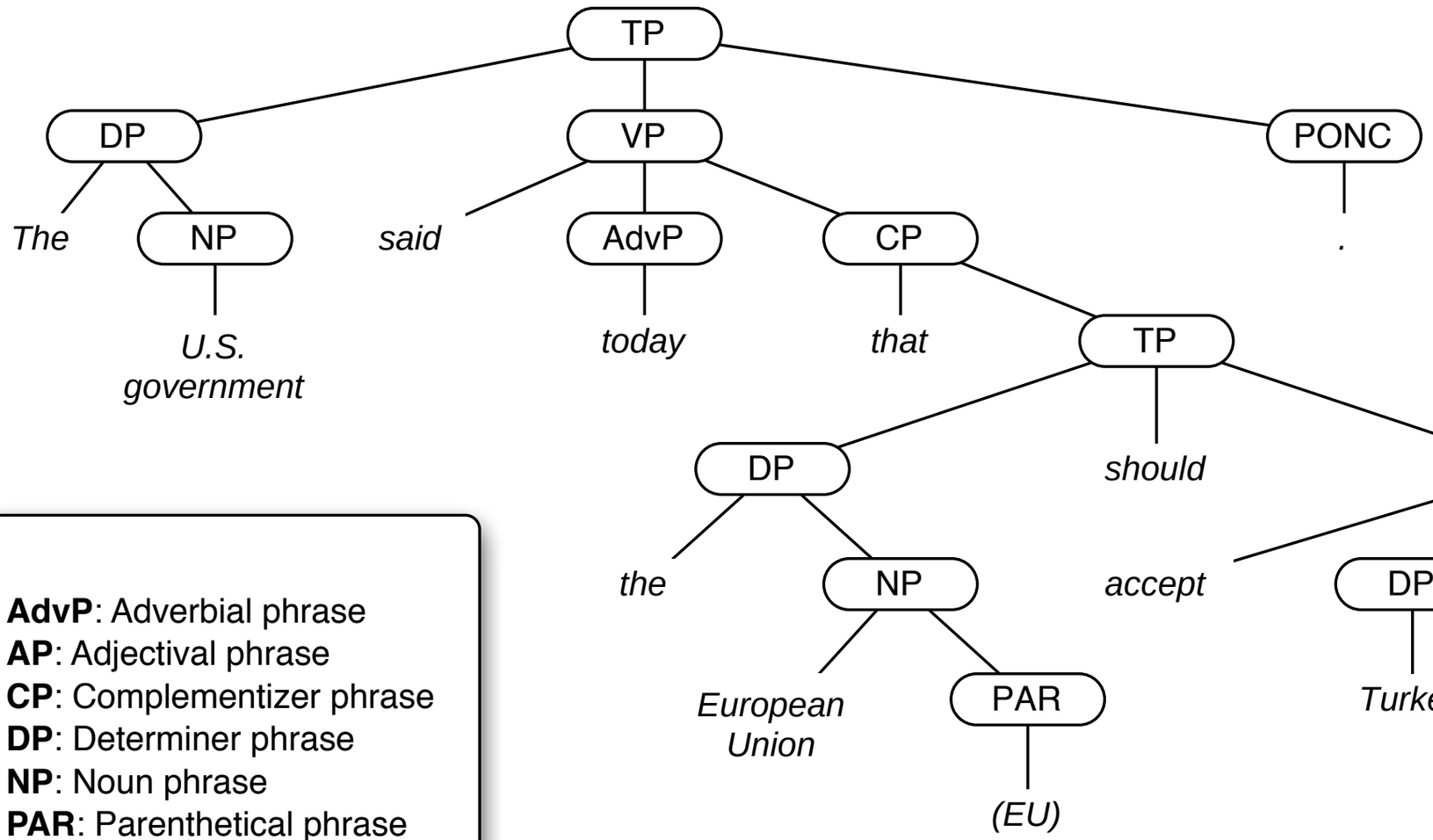
The U.S. government said today that the European Union (EU) should accept Turkey as its new member in the future.



AdvP: Adverbial phrase
AP: Adjectival phrase
CP: Complementizer phrase
DP: Determiner phrase
NP: Noun phrase
PAR: Parenthetical phrase
PONC: Ponctuation
PP: Prepositional phrase
TP: Tense phrase
VP: Verb phrase

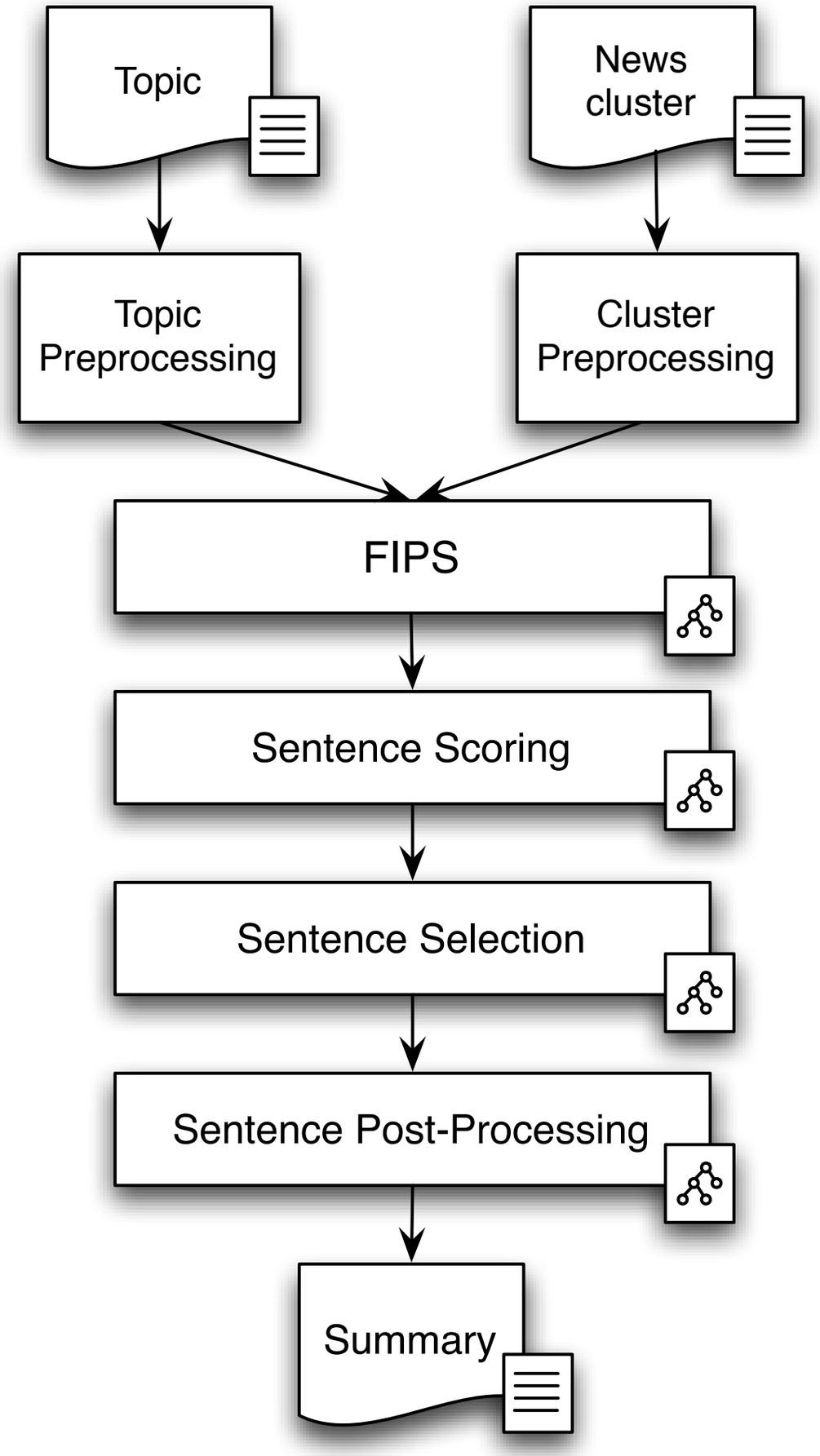
FIPS output for

The U.S. government said today that the European Union (EU) should accept Turkey as its new member in the future.

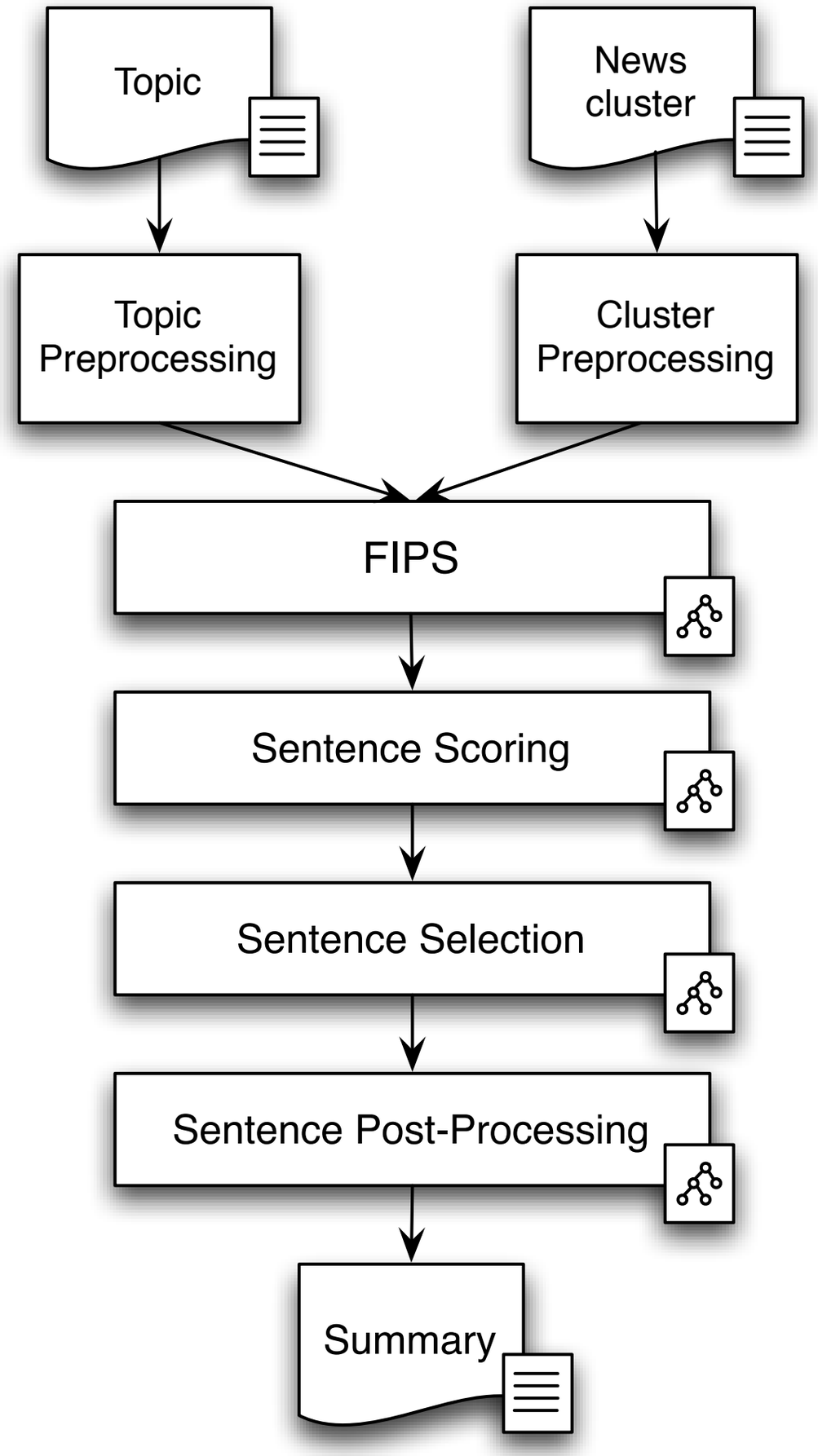


Full parse of sentences
71% for docs
93% for topics

AdvP: Adverbial phrase
AP: Adjectival phrase
CP: Complementizer phrase
DP: Determiner phrase
NP: Noun phrase
PAR: Parenthetical phrase
PONC: Ponctuation
PP: Prepositional phrase
TP: Tense phrase
VP: Verb phrase


Topic
News cluster
Topic Preprocessing
Cluster Preprocessing
FIPS
Sentence Scoring
Sentence Selection
Sentence Post-Processing
Summary

rali



Minutes per cluster (25 articles)

0.1

4.0

4.0

0.1

Total: 8.2

rali

Sentence scoring

- Word-based *tf·idf* similarity score (15%)
- Lemma-based *tf·idf* similarity score (50%)
- Lemma-based *tf·idf* similarity score with node depth (5%)
- Sentence weight (20%)
- Absolute sentence position (10%)

rali

Sentence selection

- Keep sentences with the highest scores
- Sentences are dismissed (regardless of score) if
  - they cannot be parsed by FIPS (29%)
  - duplicate from different documents (4%)
  - without a verb (5%)
  - with the « I » pronoun (3%)
  - ending with « : » or « ? » (2%)
  - with all upper case words or with less than 5 words (4%)

rali

Sentence post-processing

- Referential clarity
  - some pronouns are removed
    Climate is changing, he said ⇒ *Climate is changing*
  - ambiguous temporal references are fixed
    - Reference to the present day ⇒ date of document
    - Day of the week ⇒ month and year of document
    - No repetition of a date within a summary
- Sentence compression by pruning *non-essential* subtrees (e.g. parenthetical expressions)

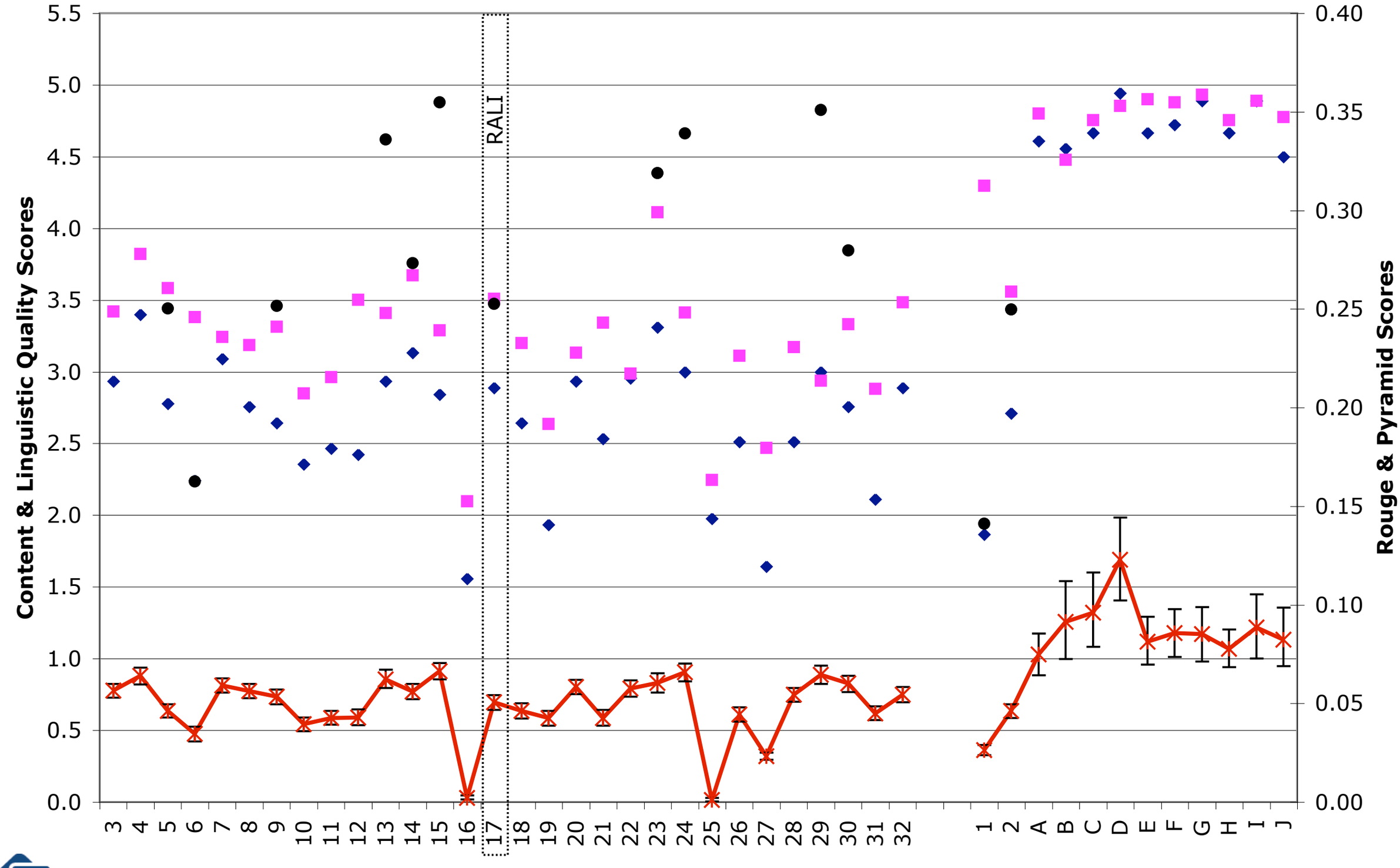
rali

Results

- Content (11th)
- Linguistic quality (5th)
- Bad non-redundancy (23rd)
- Pyramid: 8th over 11

rali

DUC 2007 Average Scores



rali

Possible Improvements

- Parsing : dedicated lexicons
- Anaphora resolution with pronoun resolutions
- Reduce redundancy with internal *tf·idf*
- Better pruning of subordinate clauses, adjectival and adverbial modifiers

rali

Possible Improvements

- Parsing : dedicated lexicons
- Anaphora resolution with pronoun resolutions
- Reduce redundancy with internal *tf·idf*
- Better pruning of subordinate clauses, adjectival and adverbial modifiers

Combine with Wordnet, Gazetteers, etc

rali

Conclusion

- Simple and powerful
- Back to the roots of AI
- Modern tools and reliable syntactic parsers open new possibilities for principled summarization

rali