Multiple Alternative Sentence Compressions (MASC)

A Framework for Automatic Summarization

Nitin Madnani, David Zajic, Bonnie Dorr

Necip Fazil Ayan, Jimmy Lin

University of Maryland, College Park

Outline

- Problem Description
- MASC Architecture
- MASC Results
- Improving Candidate Selection
- Summary & Future Work

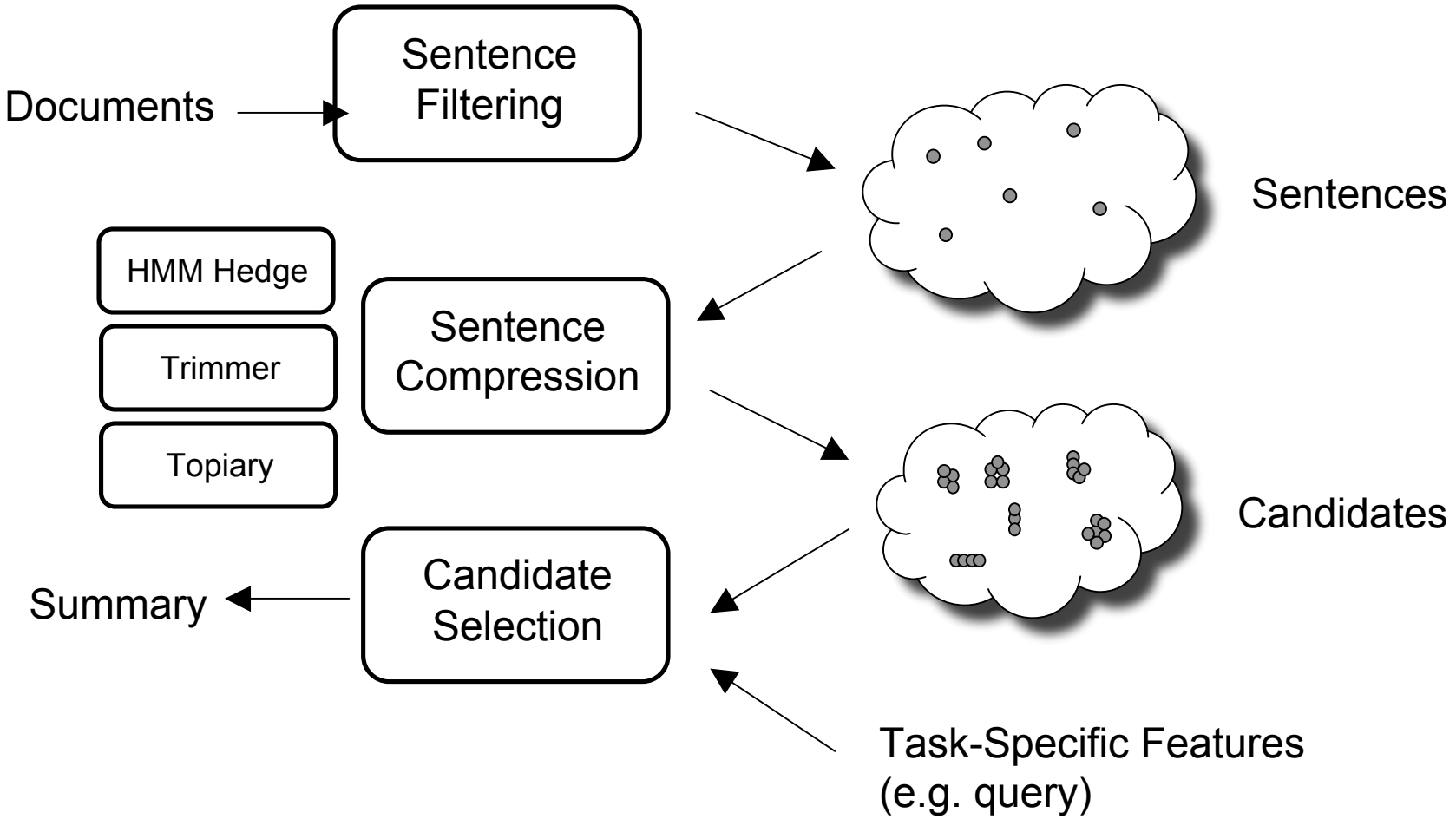
Problem Description

- Sentence-level extractive summarization
  - Source sentences contain mixture of relevant/non-relevant, novel/redundant information.
- Compression
  - Single output compression can't provide best compression of each sentence for every user need.
- Multiple Alternative Sentence Compression
  - Generation of multiple candidate compressions of source sentences.
  - Feature-based selection to choose among candidates.

Outline

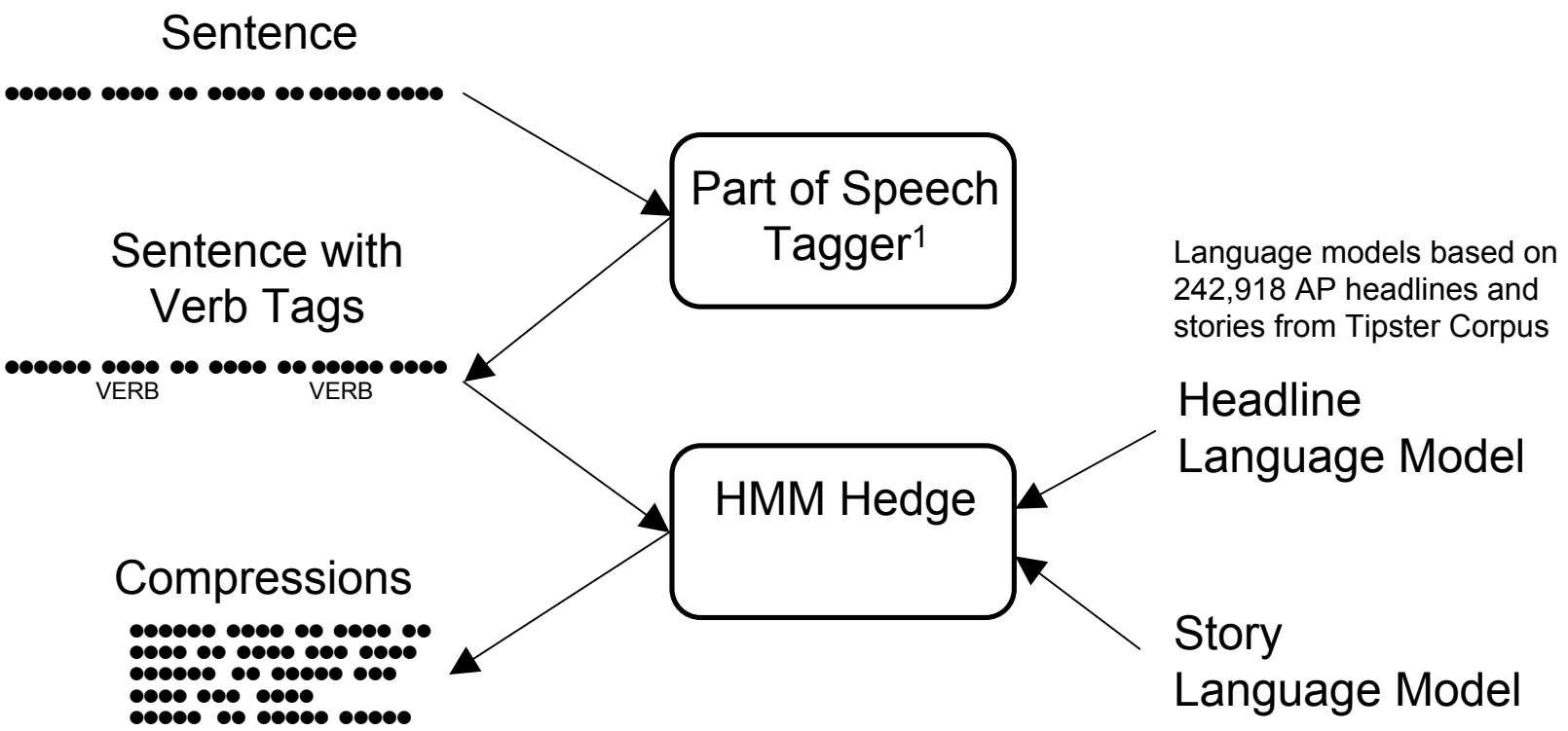
- Problem Description
- MASC Architecture
- MASC Results
- Improving Candidate Selection
- Summary & Future Work

MASC Architecture



(Zajic et al., 2005) (Zajic et al., 2006)

HMM Hedge Architecture

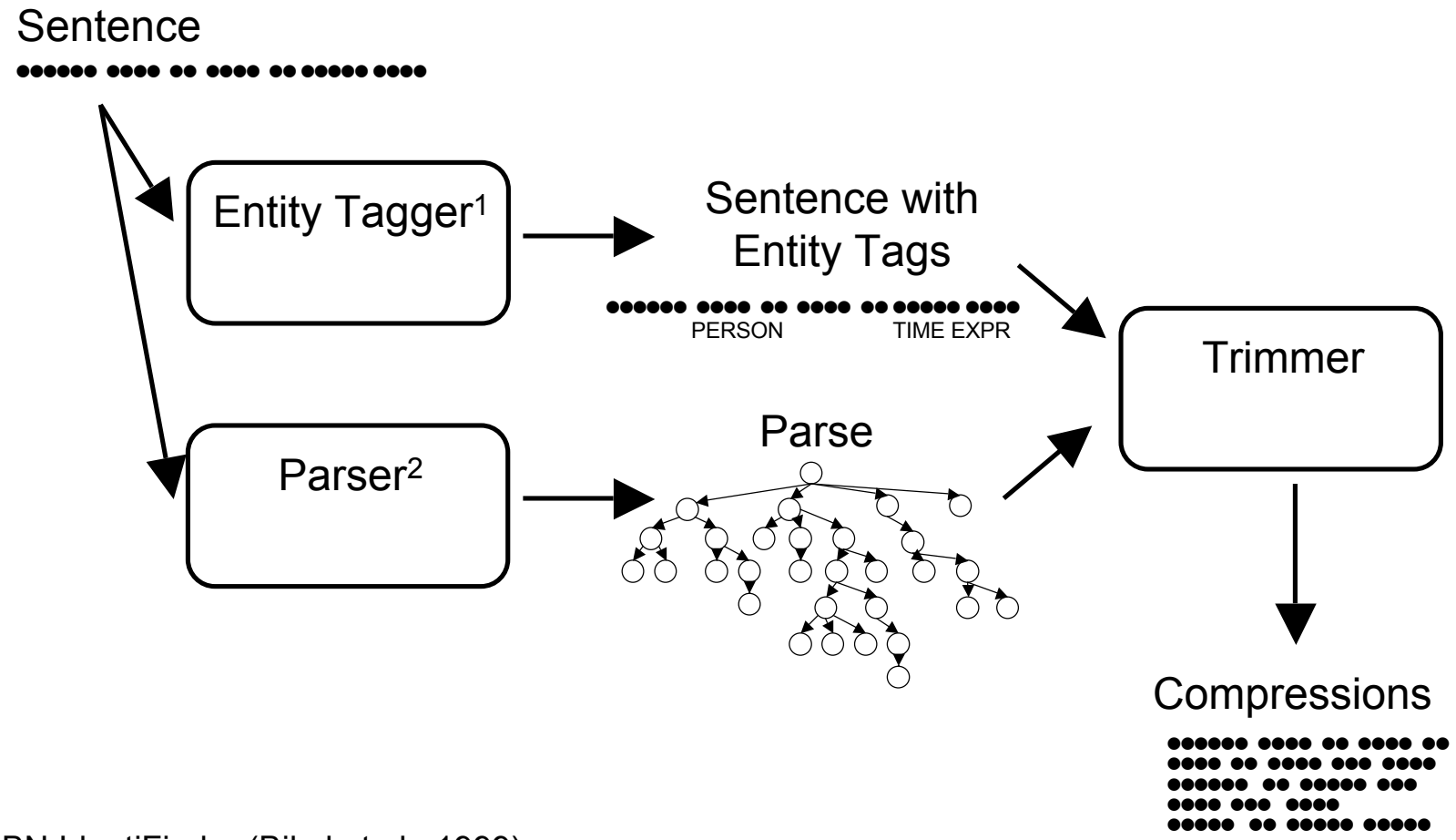


[1]TreeTagger (Schmid, 1994)

HMM Hedge
Multiple Alternative Compressions

- Calculate best compression at each word-length from 5 to 15 words
- Calculate 5 best compressions at each word length

Trimmer Architecture

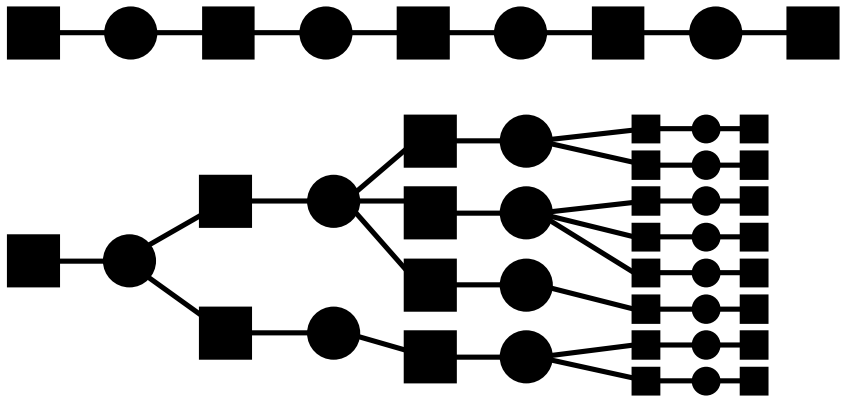


[1]BBN IdentiFinder (Bikel et al., 1999)

[2]Charniak Parser (Charniak, 2000)

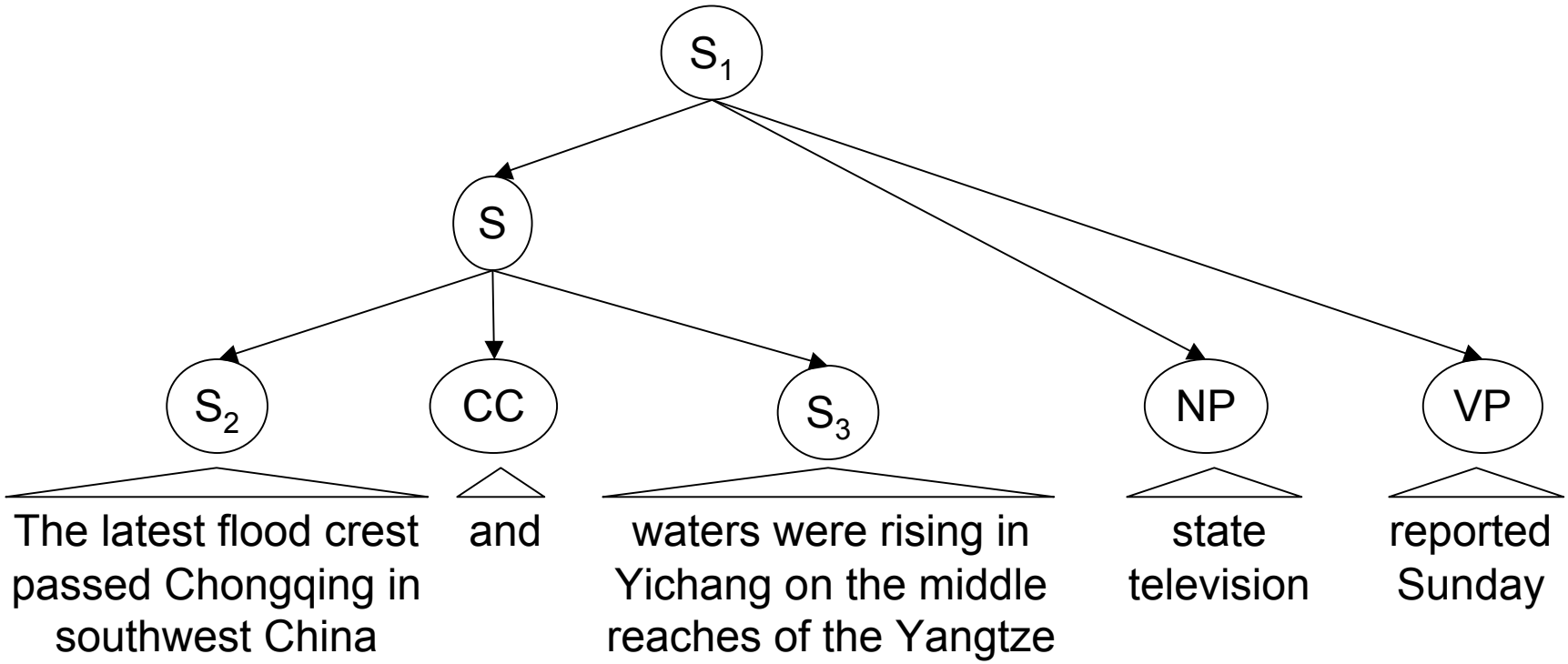
Multi-candidate Trimmer

- How to generate multiple candidate compressions?
  - Use the state of the parse tree after each rule application as a candidate
  - Use rules that generate multiple candidates
  - 9 single-output rules, 3 multi-output rules
    - Zajic et al, 2005, 2006; Zajic 2007



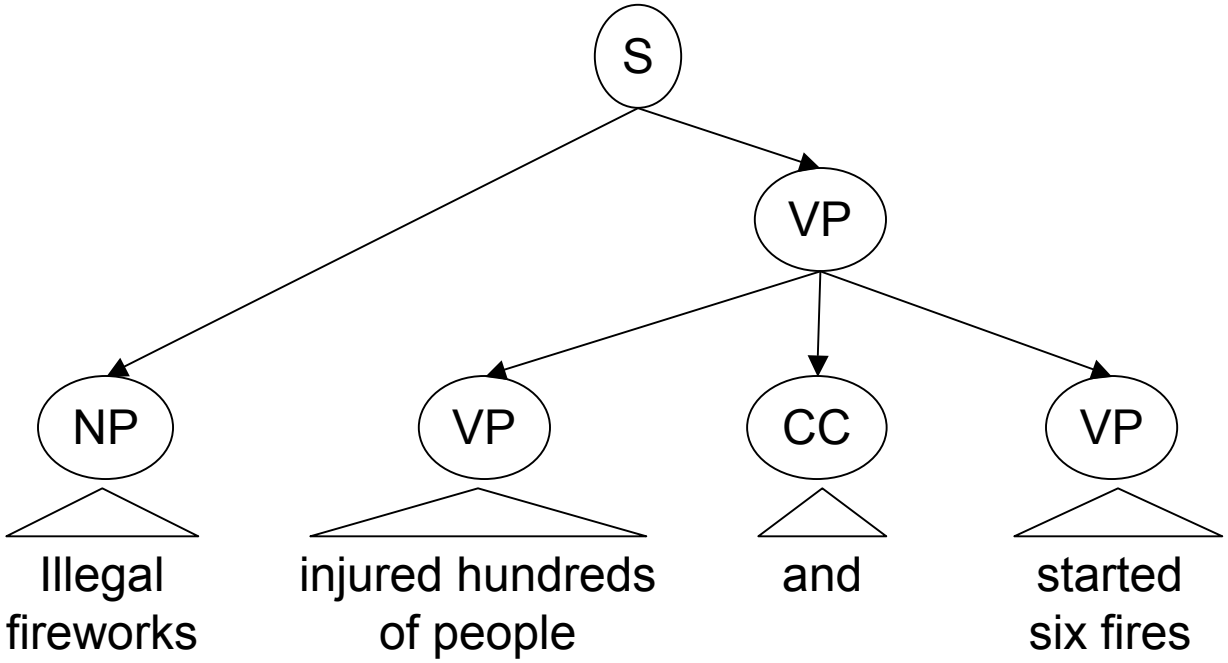
Trimmer Rule: Root-S

- Select node to be root of compression
- Consider any S node with NP,VP children

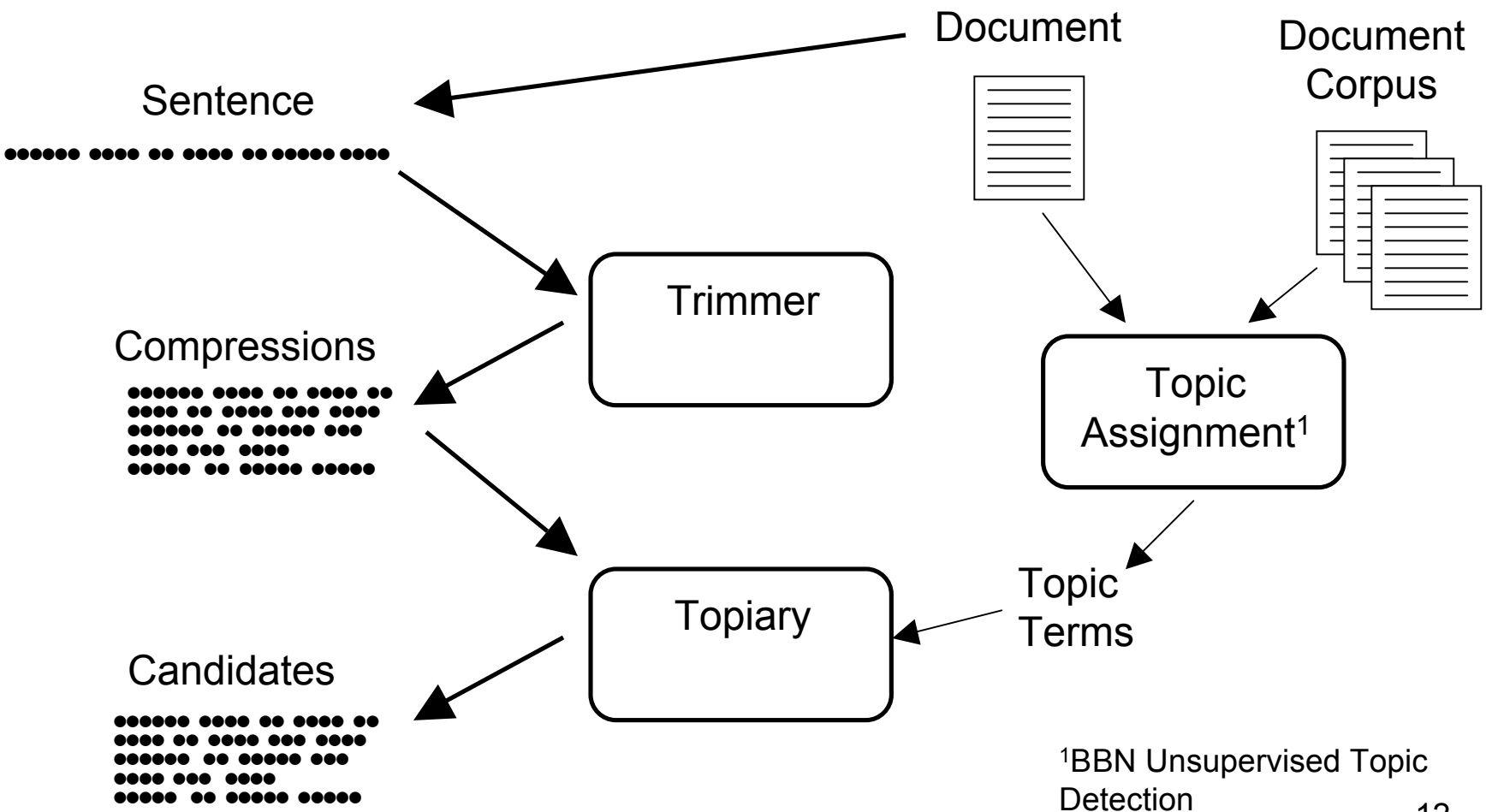


Trimmer Rule: Conjunction

- Conjunction rule removes right, left or neither child.



Topiary Architecture



[1]BBN Unsupervised Topic Detection

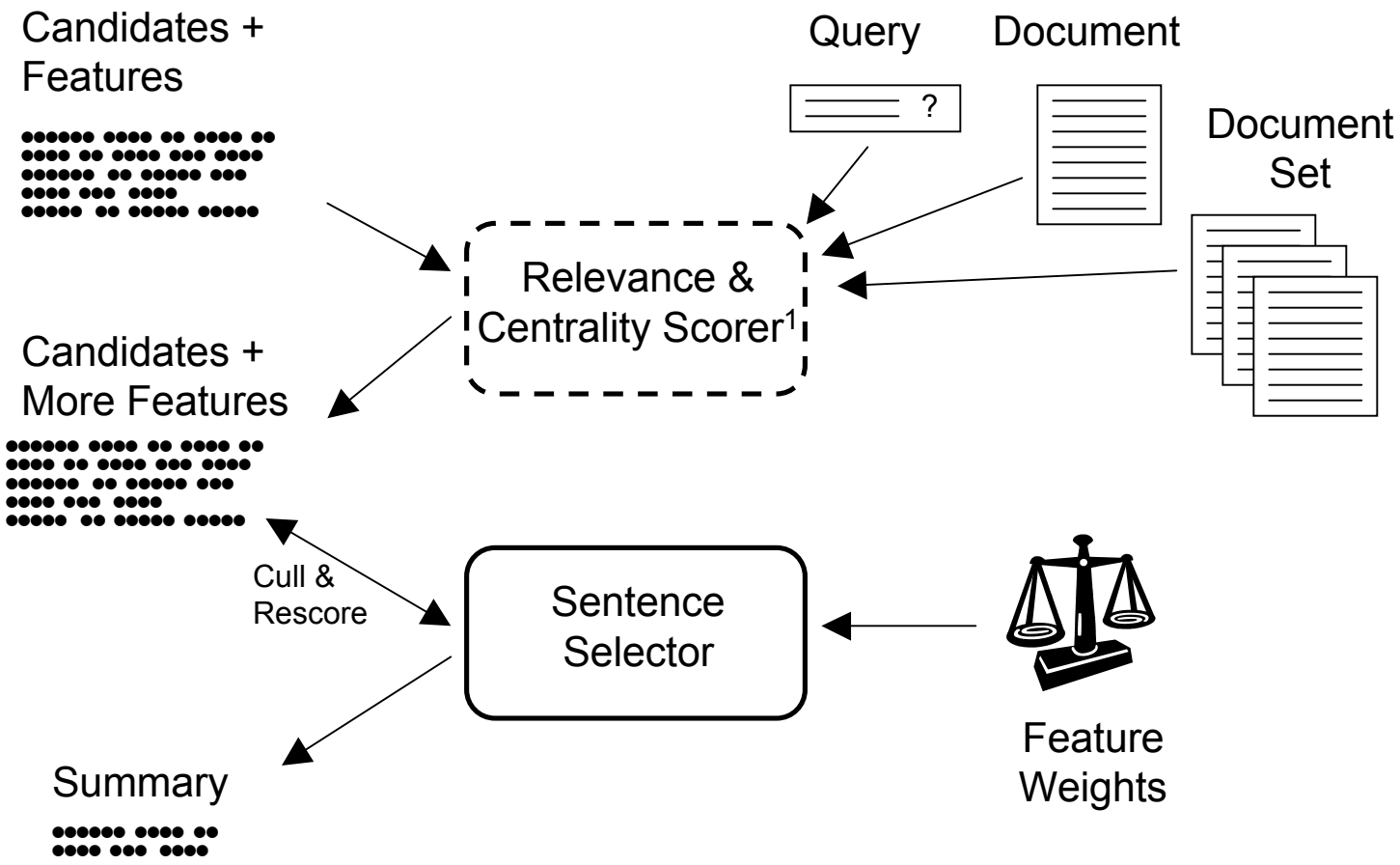
Topiary Examples
DUC2004

PINOCHET: wife appealed saying he too sick to be extradited to face charges

MAHATHIR ANWAR_IBRAHIM: Lawyers went to court to demand client's release

- **Mahathir Mohamad** is the former Prime Minister of Malaysia
- **Anwar bin Ibrahim** is a former deputy prime minister and finance minister of Malaysia, convicted of corruption in 1998

Selector Architecture

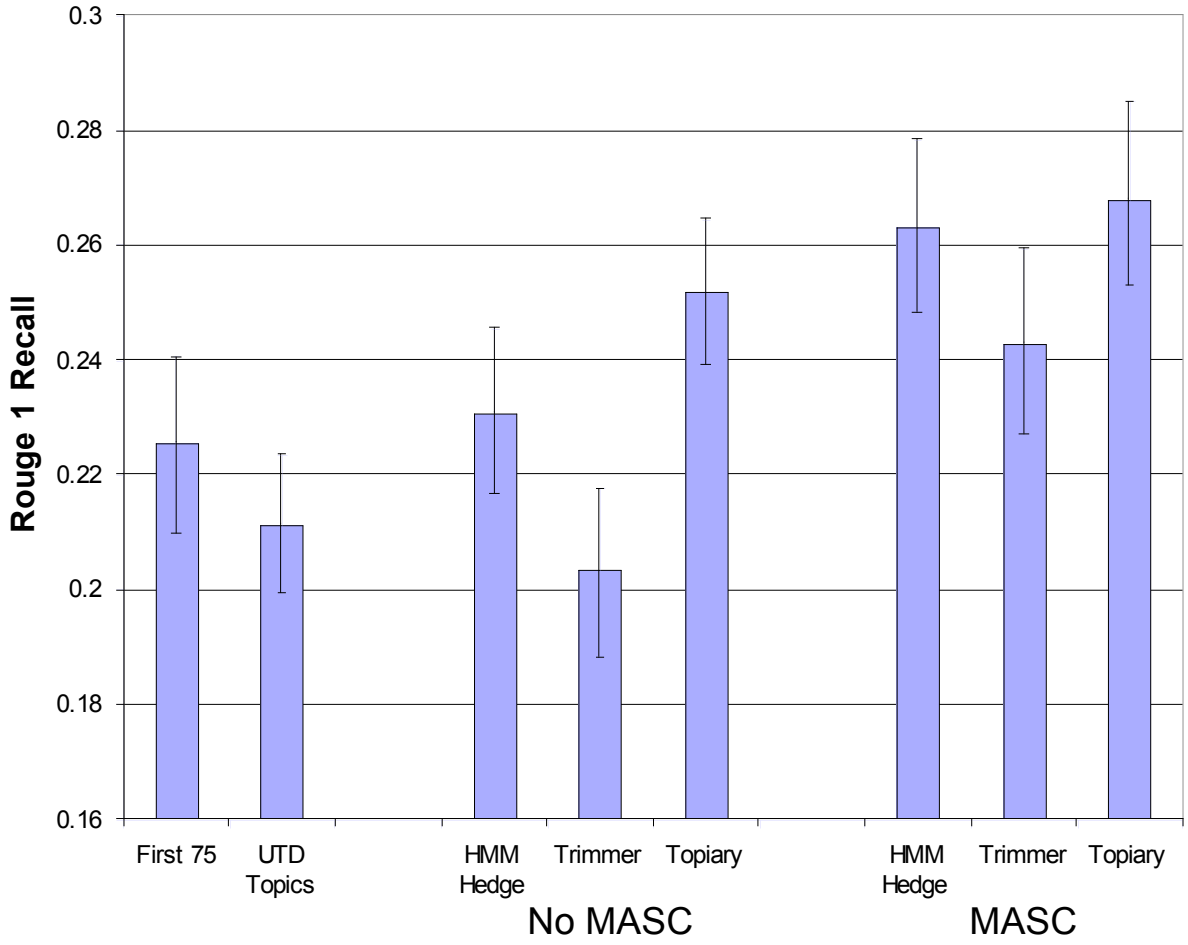


[1]Uniform Retrieval Architecture (URA), UMD's software infrastructure for IR tasks.

Outline

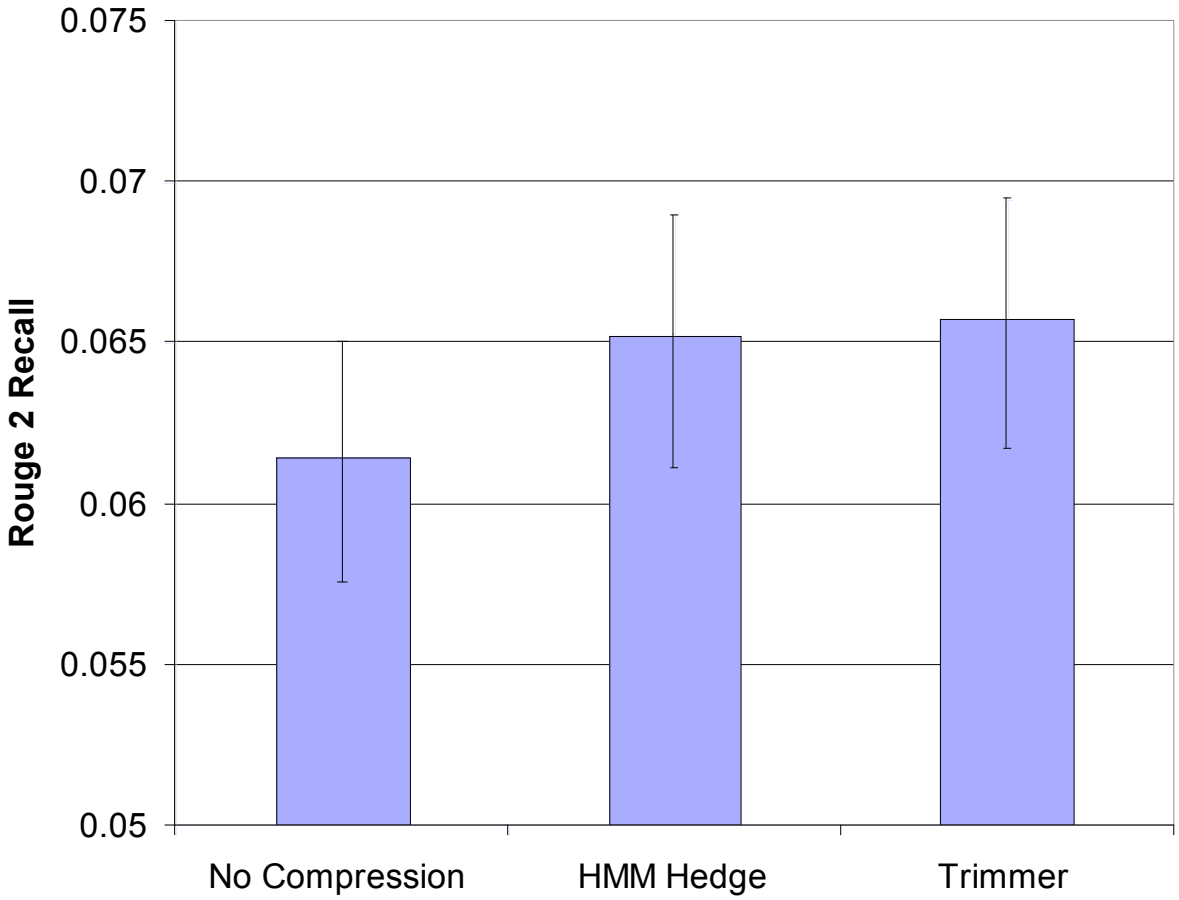
- Problem Description
- MASC Architecture
- MASC Results
- Improving Candidate Selection
- Summary & Future Work

Evaluation of Headline Generation Systems



DUC2004 Test Data, Rouge recall with unigrams

Evaluation of Multi-Document Summarization Systems

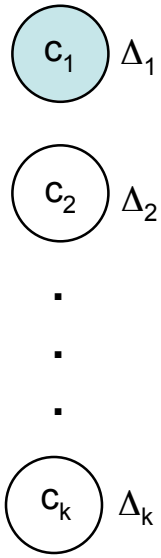


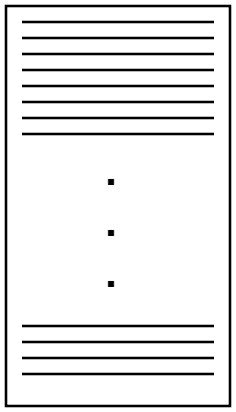
DUC2006 Test Data

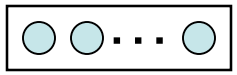
Outline

- Problem Description
- MASC Architecture
- MASC Results
- Improving Candidate Selection
- Summary & Future Work

Tuning Feature Weights with Δ_{ROUGE}




Hypotheses(**H**)


Summary(**S**)

Initialize: **S** = {}, **H** = {}

C ← current k-best candidates

for c ∈ **C**

$\Delta_{ROUGE}(c) = R_{2R}(\mathbf{S} \cup \{c\}) - R_{2R}(\mathbf{S})$

Add hypothesis to **H**

S ← **S** ∪ $\{c_1\}$

Update remaining candidates

Repeat unless |**S**| > L

$w_{opt} \leftarrow powell_{ROUGE}(\mathbf{H}, w_0)$

Optimization Results

ROUGE	Manual	Δ_{ROUGE} (k=10)
1	0.363	**0.403**
2	0.081	**0.104**
SU-4	0.126	**0.154**

Manual : Feature weights optimized **manually** to maximize ROUGE-2 Recall on the **final** system output

Key Insights for Δ_{ROUGE} optimization:

- Uses multiple alternative sentence compressions
- Directly optimizes candidate selection process.

DUC2007 data, all differences significant at ***p < 0.05***

Redundancy

- Candidate words can be emitted by two disparate word distributions

$$P(w) = \lambda P(w \mid S)\left(= n(w,S)/|S|\right) + (1\text{-}\lambda)\, P(w \mid L)\left(= n(w,L)/|L|\right)$$

REDUNDANT **NON-REDUNDANT**

S = Summary, *L* = General English language†

- Assuming candidate words are i.i.d., the redundancy feature for a given candidate is:

$$R(c) = \log\left(P(c)\right) = \log\left(\prod_{w \in c} \lambda P(w \mid S) + (1-\lambda) P(w \mid L)\right)$$

†Other documents in the same cluster are used to represent the general language

Incorporating Paraphrases

- Redundancy uses bags-of-words to compute *P(w|S)*

$$P(w \mid S) = \frac{n(w,S)}{|S|}$$

- Not useful if candidate word is a paraphrase of summary word (classified as non-redundant)

- Add another bag-of-words *P*, such that

$$P = \{ \text{ a paraphrase for } w, \quad \forall w \in S \}$$

- Use *n(w,P)* for redundancy computation if *n(w,S)* = 0

Generating Paraphrases

- Leverage phrase-based MT system
  - Use E-F correspondences extracted from word-aligned bi-text
  - Pivot each pair of E-F correspondence with common foreign side to get E-E correspondence
  - $c(e_1,e_2) = \sum_f c(e_1,f)c(f,e_2)$
- Example

```
上升 ||| climbed   ||| 1.0
上升 ||| increased ||| 2.0
上升 ||| uplifted  ||| 1.0
```

```
increased ||| climbed ||| 2.0
climbed ||| uplifted ||| 1.0
. . .
. . .
uplifted ||| increased ||| 2.0
```

- Pick most frequent correspondence for *w*

Paraphrase Results

- Using paraphrases yields no significant improvements
- Unrelated to the quality of the paraphrases
- Anomalous cases occur extremely rarely
  - The original bag-of-words is sufficient to capture candidate redundancy almost all the time

Outline

- Problem Description
- MASC Architecture
- MASC Results
- Improving Candidate Selection
- Summary & Future Work

DUC 2007 Results

- Systems 7, 36
- Main:
  - Responsiveness = 3.089 (4th)
  - ROUGE-2 = 0.108 (8th)
  - ROUGE-SU4 = 0.158 (11th)
- Update:
  - Responsiveness = 2.800 (2nd)
  - ROUGE-2 = 0.086 (9th)
  - ROUGE-SU4 = 0.124 (8th)

Summary

- MASC with feature-based candidate selection improves headline generation and shows promise for multi-document summarization.
- Optimizing for Δ_{ROUGE} provides significant improvements over previous approach
- Redundancy feature works at lexical as well as document-level
- Using paraphrases requires novel formulation

Future Work

- Fully explore Trimmer search space
- Split redundancy feature into its components and tune λ automatically
- Use an n-gram LM to estimate *P(w|L)*
- Continue to experiment with paraphrase-based approaches to redundancy
  - Scale up to phrase-level paraphrases
  - Use combination of high-coverage and high-quality paraphrases