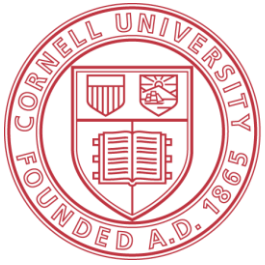
A Sentence Compression Based Framework to Query-focused Multi-document Summarization



Cornell University

IBM Research

Lu Wang[1], Hema Raghavan[2], Vittorio Castelli[2], Radu Florian[2], and Claire Cardie[1]

[1]Department of Computer Science
Cornell University

[2]IBM T. J. Watson Research Center

Problem

- **Query-focused multi-document summarization**
  - Given a complex query (or open-ended question) and a set of relevant documents, we aim to generate a fluent, well-organized, and compact summary that answers the query.

Related Work

- Document Understanding Conference (DUC) has fostered the task of query-focused multi-document summarization since 2004.
- Most top-performing systems for multi-document summarization **remain largely extractive**.
  - Topic signature words: Lin and Hovy (2000), Conroy et al. (2006)
  - Graph-based approach: Erkan and Radev (2004), Otterbacher et al. (2005)
  - Content (topic) model-based approach: Daume and Marcu (2006), Haghighi and Vanderwende (2009), Celikyilmaz and Hakkani-Tur (2011)
  - Discriminative ranking: Fuentes et al., (2007)
  - Submodular function: Lin and Bilmes (2011)

Unfortunately, ...

- Extractive summaries sometimes are not compact enough.

- In human written summaries, people tend to use:
  - Paraphrase
  - Abstraction
  - Reordering
  - Sentence Compression

Motivation

- ***Query**: Track the spread of the West Nile virus through the United States and the efforts taken to control it.*

  - The publicity about the West Nile virus and counties' intensified mosquito control efforts have prompted more residents to call about dead crows they have found, which has led to more testing, [said Dennis McGowan, a spokesman for the state Department of Health and Senior Services].

Motivation

- ***Query**: In what ways have stolen artworks been recovered? How often are suspects arrested or prosecuted for the thefts?*

  - A man suspected of stealing a million-dollar collection of [hundreds of ancient] Nepalese and Tibetan art objects in New York [11 years ago] was arrested [Thursday at his South Los Angeles home, where he had been hiding the antiquities, police said].

Contribution

- We present *learning-based sentence-compression framework* for query-focused multi-document summarization.
- We use a *beam search decoder* to find highly probable compressions in an efficient way.
- Under this framework, we show how to integrate various indicative metrics such as *linguistic motivation* and *query relevance* into the compression process.
- By evaluation on newswire articles, we show that sentence compression can provide significant improvements over pure extraction-based approaches in both automatic and human evaluation.

Related Work

- Sentence compression
  - Knight and Marcu (2000) use Noisy-channel model to generate compressions, and Galley and McKeown (2007) extend it via synchronous context-free grammars (SCFG).
  - Discriminative leaning is investigated for deciding if a term should be removed based on syntax information by McDonald (2006).
  - Clarke and Lapata (2008) integrate discourse structure by Integer Linear Programming.

Related Work

- Sentence compression to multi-document summarization
  - Heuristics-based compression:
    - Zajic et al., (2006) and Gillick and Favre (2009) use heuristics to generate multiple alternative compressions and rank all the sentences.
  - Learning-based compression
    - Martins and Smith (2009) present a learning model based on dependency tree which can determine if a node should be removed.
    - Berg-Kirkpatrick et al. (2011) use discriminative learning to model the extraction and compression together.

Framework

- **Step One**: Sentence Ranking
  - Determines the importance of each sentence given the query.
- **Step Two**: Sentence Compression
  - Iteratively generates the most likely succinct versions of the ranked sentences until a length limit is reached.
- **Step Three**: Post-processing
  - Applies coreference resolution and sentence ordering.

Framework

- **Step One**: Sentence Ranking
  - Determines the importance of each sentence given the query.
- **Step Two**: Sentence Compression
  - Iteratively generates the most likely succinct versions of the ranked sentences until a length limit is reached.
- **Step Three**: Post-processing
  - Applies coreference resolution and sentence ordering

Sentence Ranking

- Experiment with
  - Support Vector Regression (SVR) (Mozer et al., 1997)
  - LambdaMART (Burges et al., 2007).
- Training
  - 40 topics from the DUC 2005 corpus along with their manually generated abstracts.
  - ROUGE-2 score as objective

Sentence Ranking

- Sample features

Query-relevant Features	Query-independent Features
Unigram/Bigram TF similarity	Relative position
Unigram/Bigram TF-IDF similarity	Length
Mention similarity	Contains verb/web link?

Framework

- **Step One**: Sentence Ranking
  - Determines the importance of each sentence given the query.
- **Step Two**: Sentence Compression
  - Iteratively generates the most likely succinct versions of the ranked sentences until a length limit is reached.
- **Step Three**: Post-processing
  - Applies coreference resolution and sentence ordering

Framework

- **Step One**: Sentence Ranking
  - Determines the importance of each sentence given the query.
- **Step Two**: Sentence Compression
  - Iteratively generates the most likely succinct versions of the ranked sentences until a length limit is reached.
- **Step Three**: Post-processing
  - Applies coreference resolution and sentence ordering

Sentence Compression

- Challenges
  - How to remove the redundancy within sentences without producing ungrammatical compressions?
  - How to guarantee the compressions are query-relevant?
  - How to find a compression or multiple compressions efficiently?

Tree-based Compression

- Each sentence is represented by its parse tree.
- Operations are carried out on parse tree constituents based on an flexible scoring function.
  - in line with syntax-driven approaches (Galley and McKeown, 2007).
- We aim to learn how to identify the proper set of constituents to be removed.

Tree-based Compression

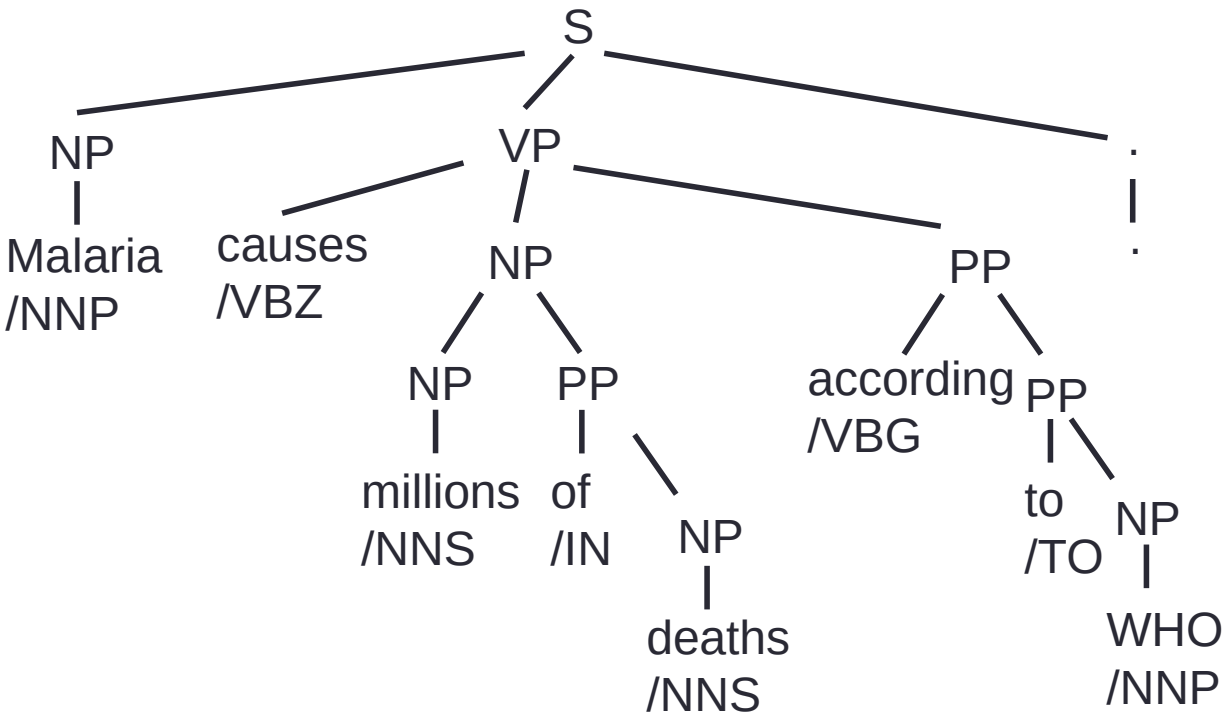
- Malaria causes millions of deaths according to WHO.

Tree-based Compression

- Malaria causes millions of deaths [according to WHO].

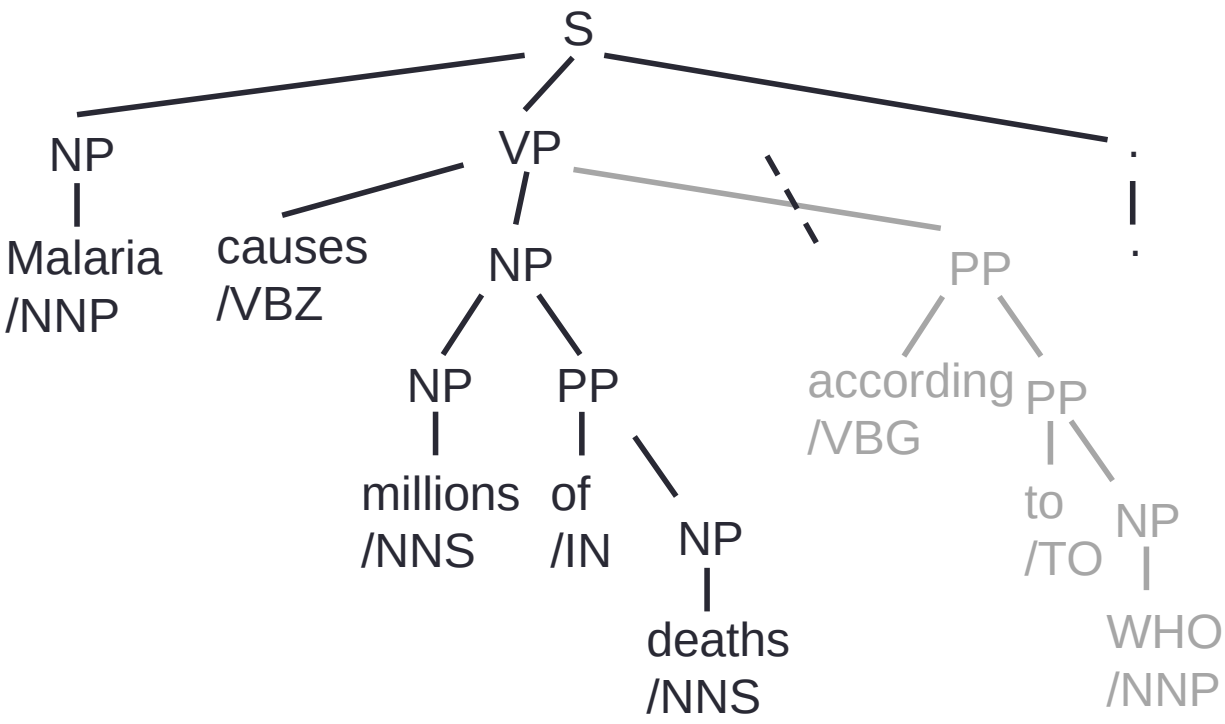
Tree-based Compression

- Malaria causes millions of deaths [according to WHO].



Tree-based Compression

- Malaria causes millions of deaths [according to WHO].



Problem Definition

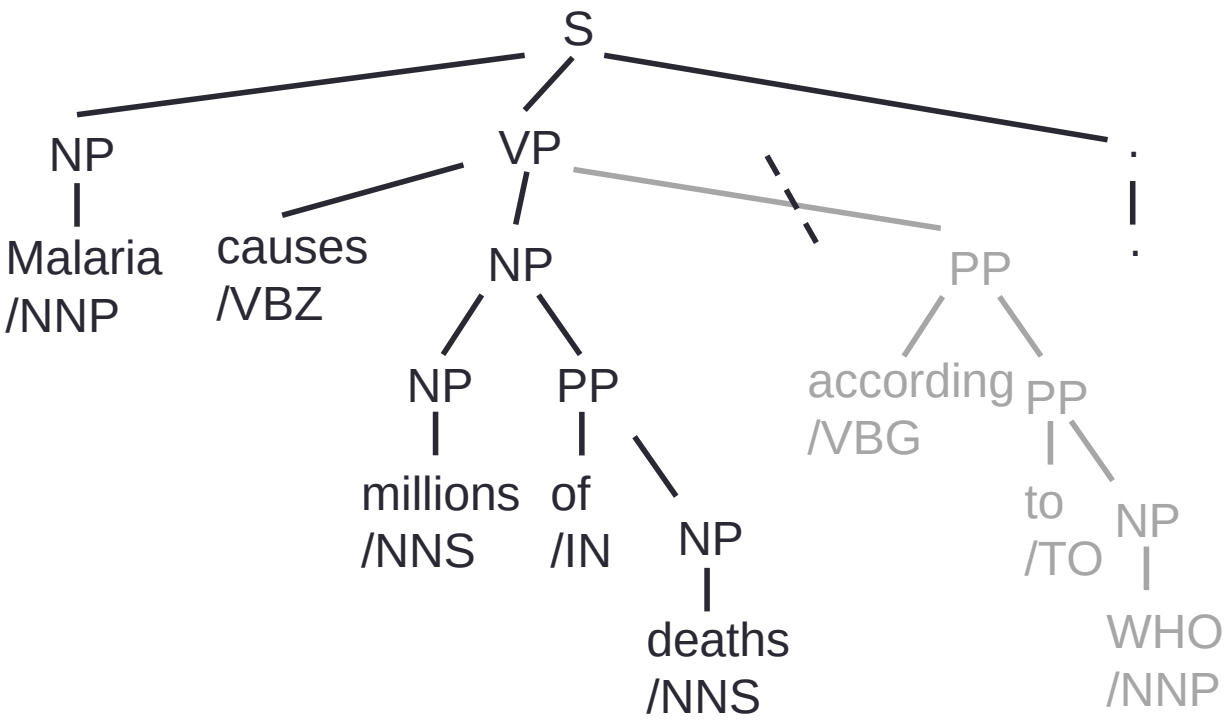
- Input:
  - A parse tree *T* of the sentence to be compressed;
  - *T* is represented as a list of ordered constituent nodes:

$$T = t_0 t_1 \ldots t_m,$$

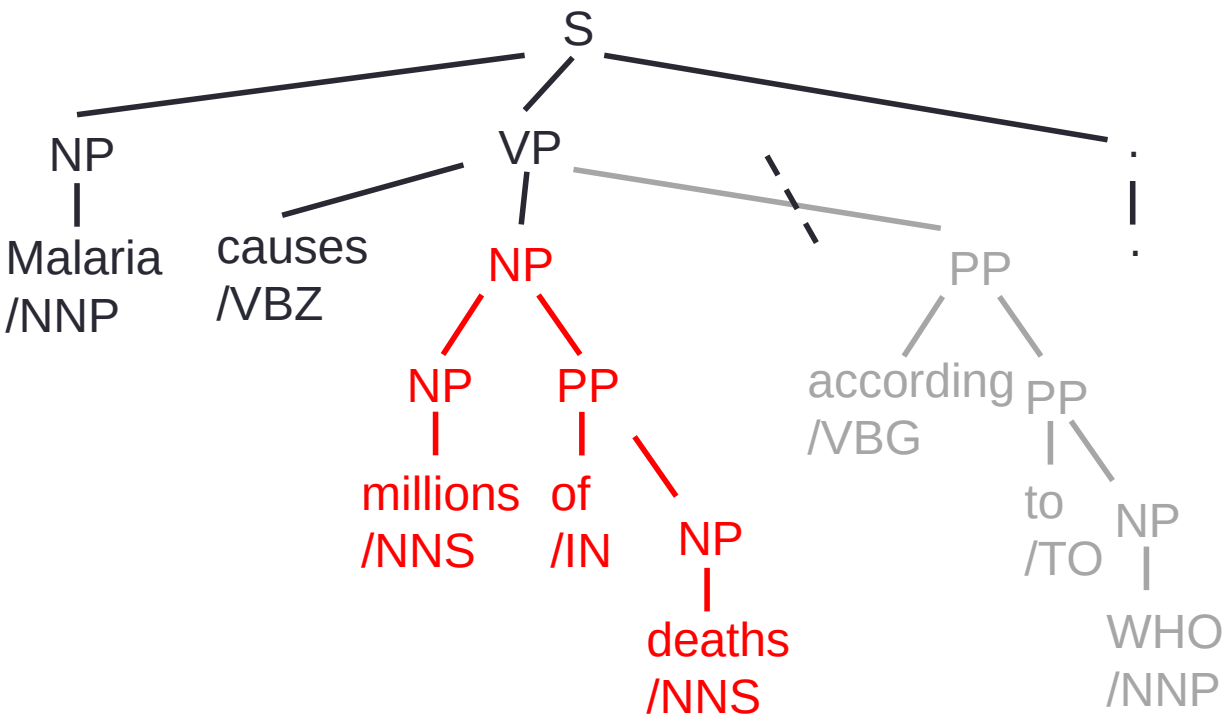
  according to a given tree traversal algorithm.

- Output:
  - A set of labels $L = l_0 l_1 \ldots l_m$, where $l_i \in \{RETAIN, REMOVE, PARTIAL\_REMOVE\}$

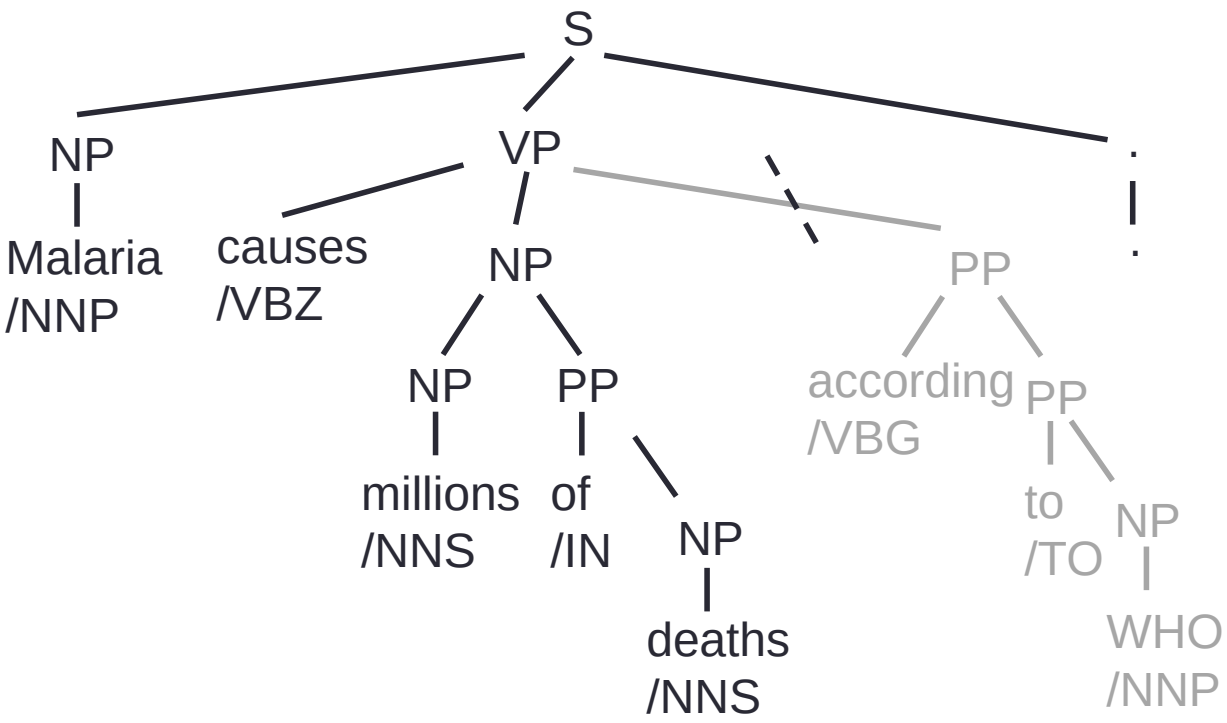
Prediction Labels -- RETAIN



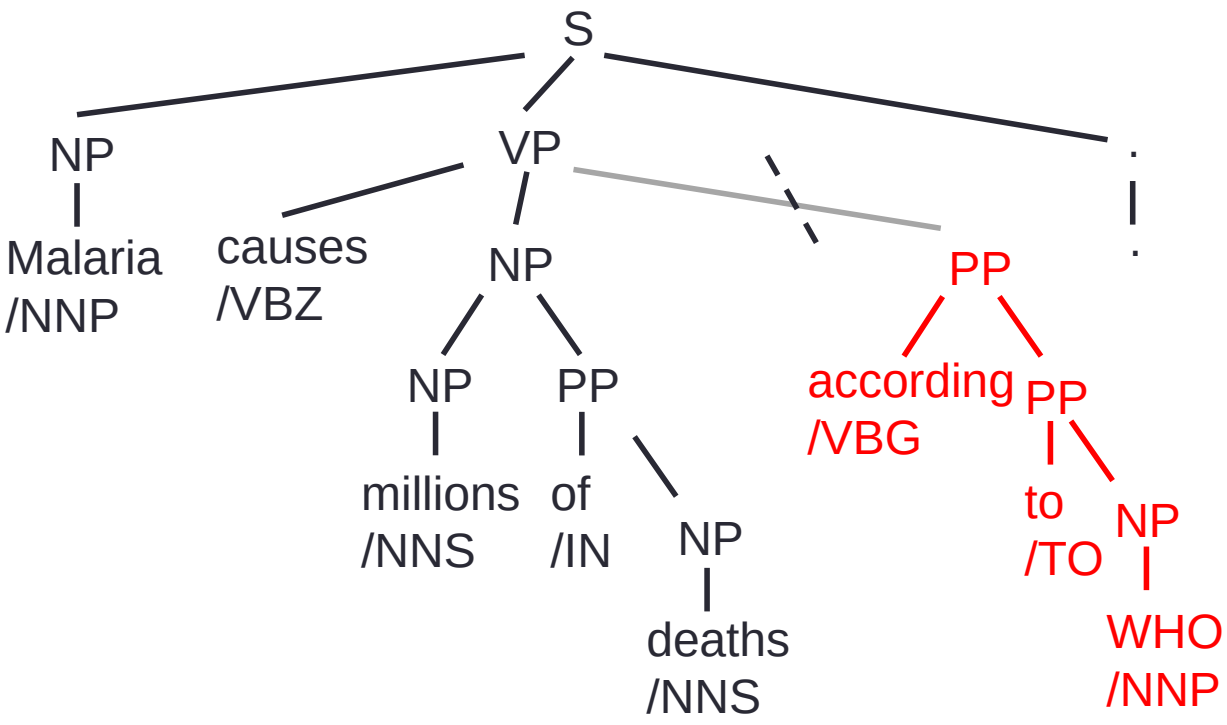
Prediction Labels -- RETAIN



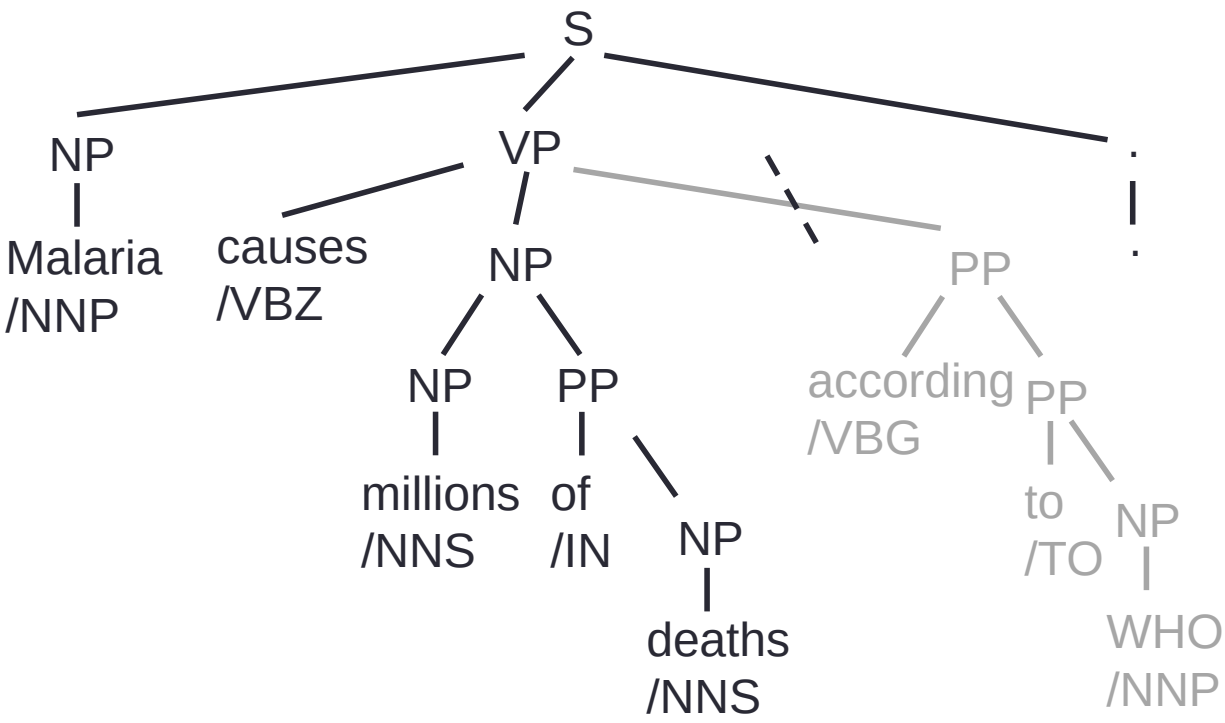
Prediction Labels -- REMOVE



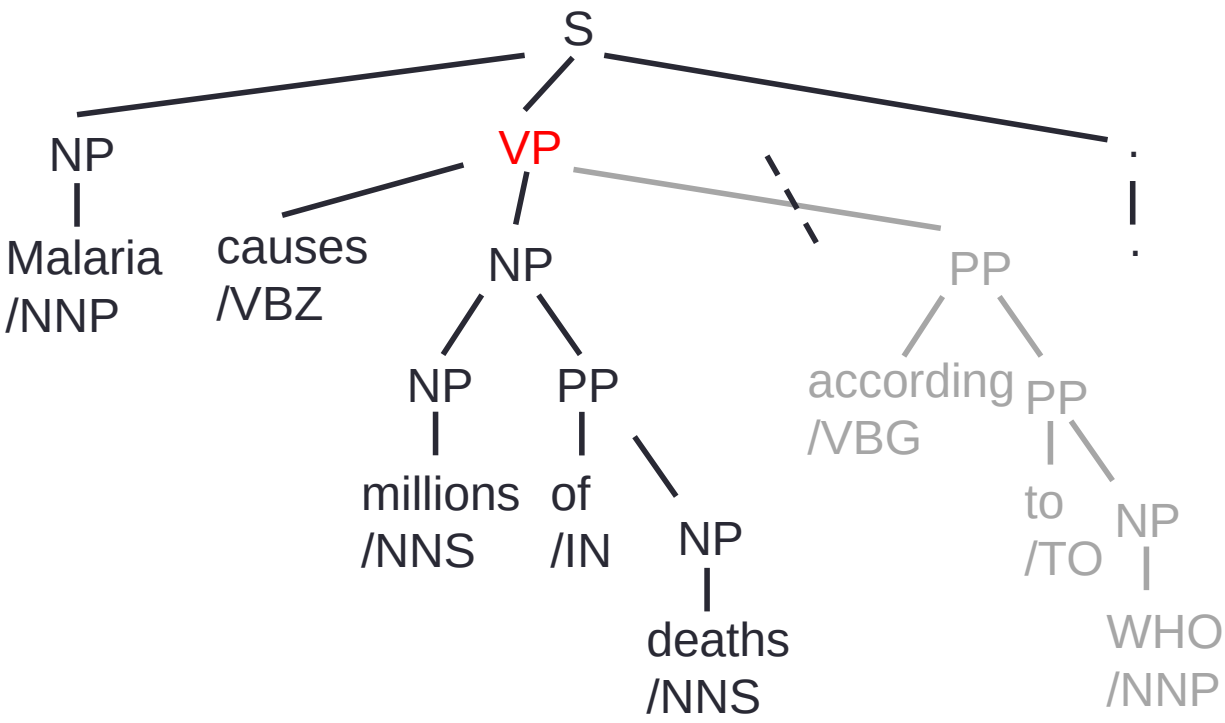
Prediction Labels -- REMOVE



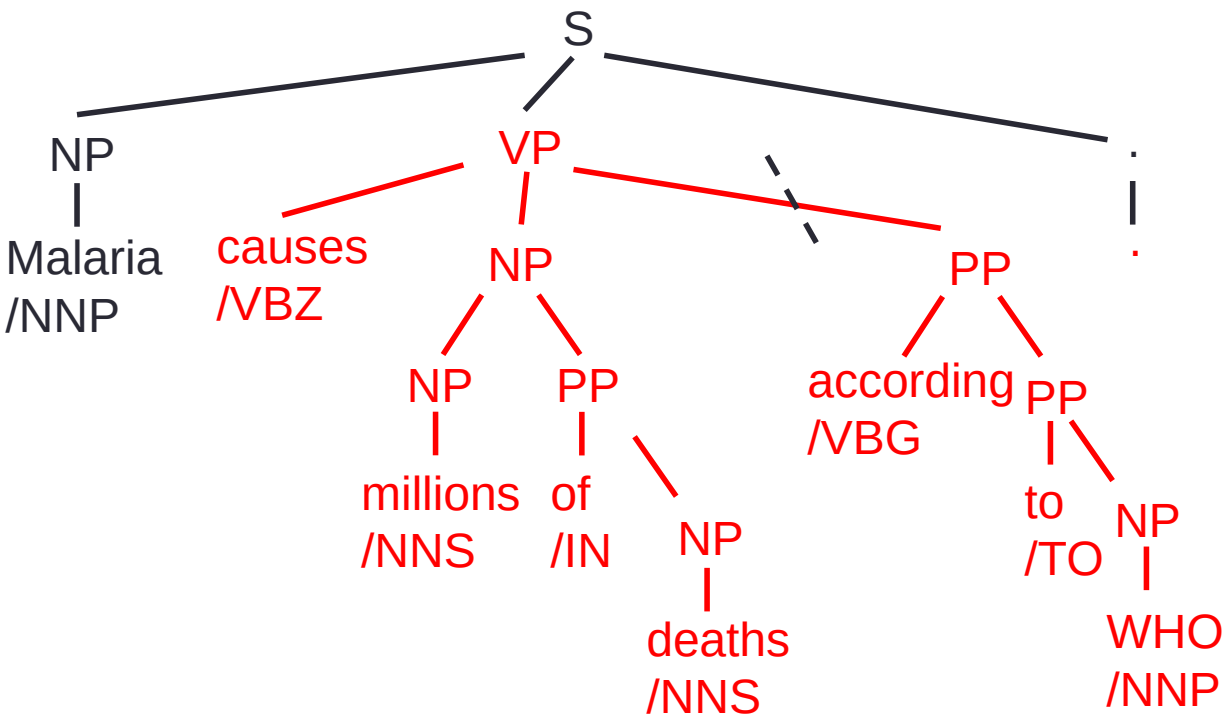
Prediction Labels – PARTIAL_REMOVE



Prediction Labels – PARTIAL_REMOVE



Prediction Labels – PARTIAL_REMOVE



Beam Search Decoder

- *T* is represented as a list of ordered constituent nodes:

$$T = t_0 t_1 \ldots t_m$$

Beam Search Decoder

- *T* is represented as a list of ordered constituent nodes:

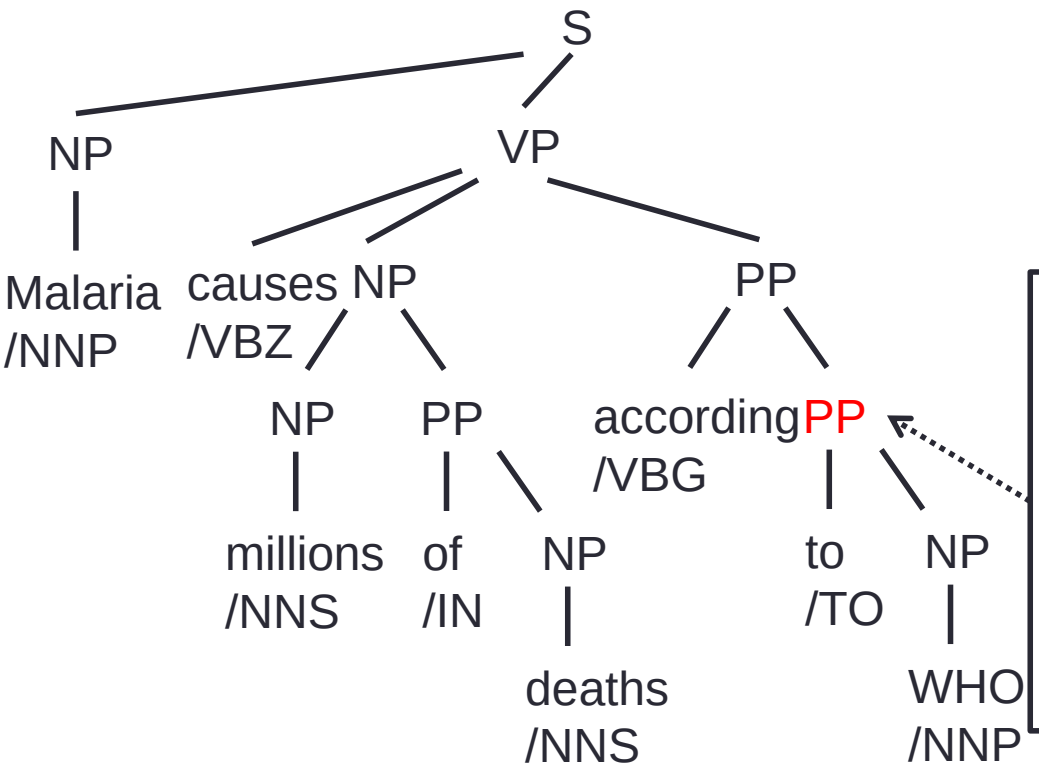
$$T = t_0 t_1 \ldots t_m$$

- Scorer S_{Basic}

$$S_{Basic}\,(L = \{l_0, l_1, \ldots, l_n\}) = \sum_{i=0}^{n} \log P(l_i | t_i)$$

for a sub-sequence $t_0 t_1 \ldots t_n$.

Beam Search Decoder



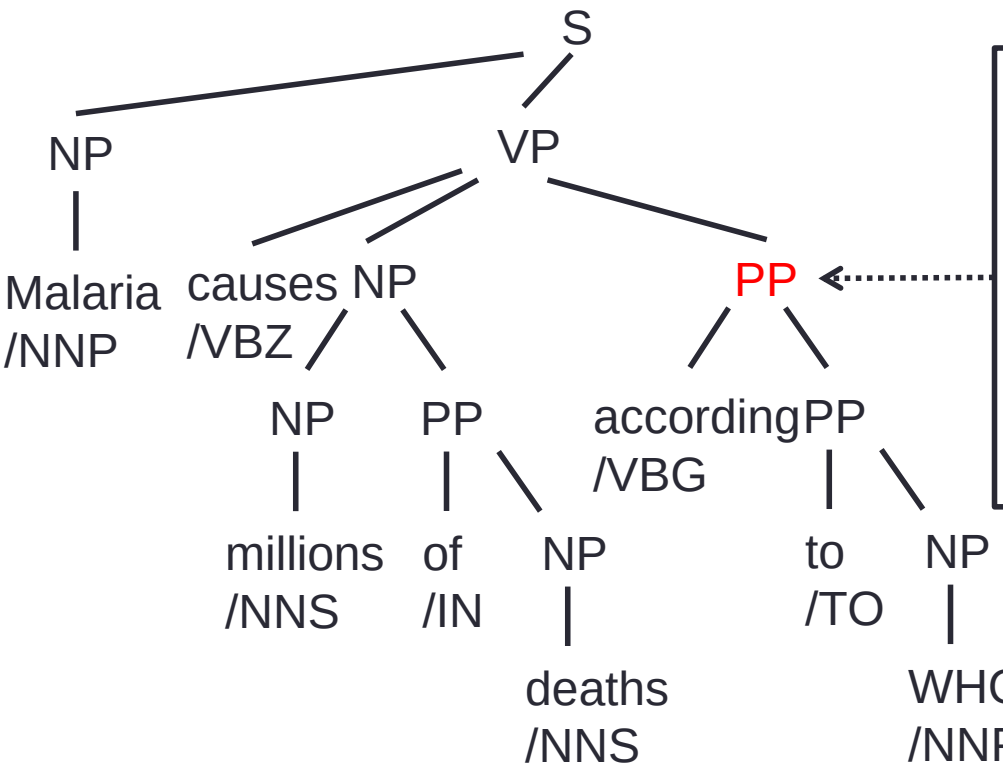
Maria causes millions of deaths according to WHO (-18.60)

Maria causes deaths according to WHO (-18.92)

Maria causes millions of deaths (-19.09)

…

Beam Search Decoder



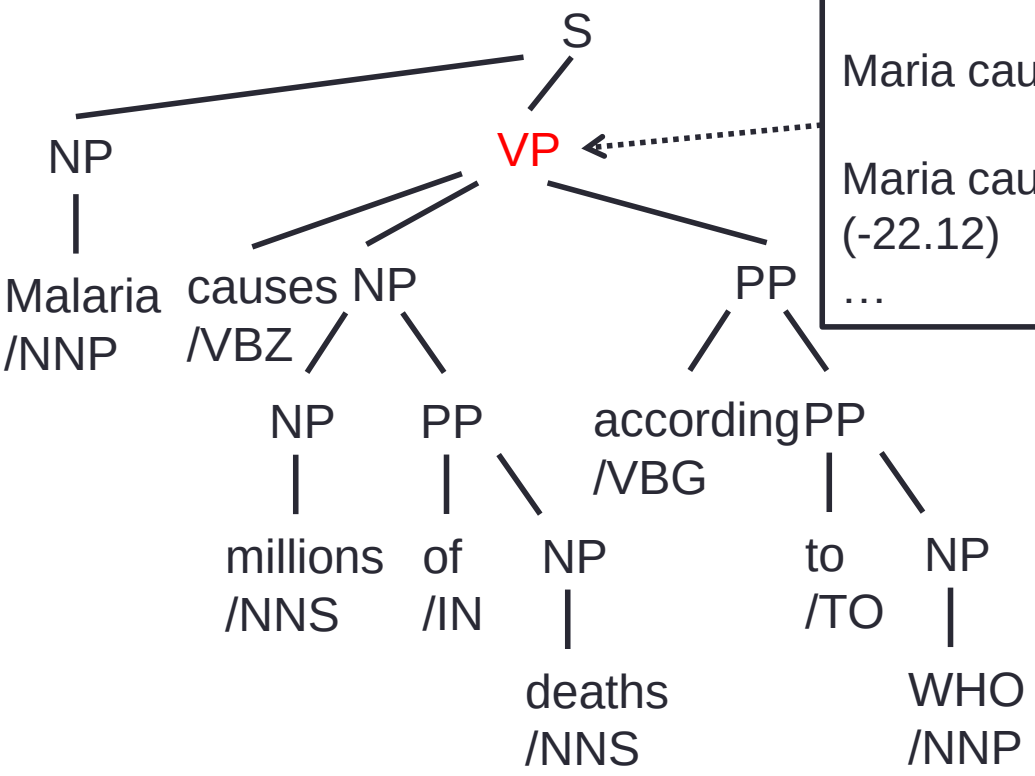
Maria causes deaths according to WHO (-19.90)

Maria causes millions of deaths (-20.10)

Maria causes millions of deaths according to WHO (-20.17)

…

Beam Search Decoder



Maria causes millions of deaths (-21.16)

Maria causes deaths according to WHO (-21.26)

Maria causes millions of deaths according to WHO (-22.12)

…

Learning to Compress

- $S_{Basic}\,(L = \{l_0, l_1, \ldots, l_n\}) = \sum_{i=0}^{n} \log P(l_i|t_i)$

- Maximum Entropy classifier is trained to produce the probability distribution on the labels for each node.

Learning to Compress

- Data:
  - Clarke and Lapata (2008)
  - 82 newswire articles with one manually produced compression aligned to each sentence.

- Sample features:

Constituent tag
Dependency relation
Is head node?
Semantic role of its head node

Linguistically-motivated Compression Rules

- Turner and Charniak (2005) have shown that applying hand-crafted rules for trimming sentences can improve both content and linguistic quality.

Linguistically-motivated Compression Rules

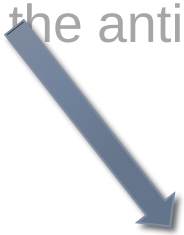
- ***Query***: *"In what ways have stolen artworks been recovered? How often are suspects arrested or prosecuted for the thefts?"*[1]

  - A man suspected of stealing a million-dollar collection of [hundreds of ancient] Nepalese and Tibetan art objects in New York [11 years ago] was arrested [Thursday at his South Los Angeles home, where he had been hiding the antiquities, police said].

Linguistically-motivated Compression Rules

- ***Query***: *"In what ways have stolen artworks been recovered? How often are suspects arrested or prosecuted for the thefts?"*[1]

  - A man suspected of stealing a million-dollar collection of [hundreds of ancient] Nepalese and Tibetan art objects in New York [11 years ago] was arrested [Thursday at his South Los Angeles home, where he had been hiding the antiquities, police said].

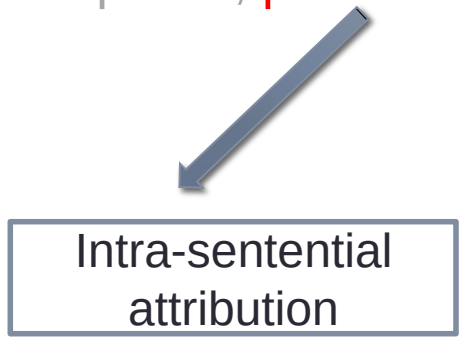


Relative date

Linguistically-motivated Compression Rules

- ***Query**: "In what ways have stolen artworks been recovered? How often are suspects arrested or prosecuted for the thefts?"*[1]

  - A man suspected of stealing a million-dollar collection of [hundreds of ancient] Nepalese and Tibetan art objects in New York [11 years ago] was arrested [Thursday at his South Los Angeles home, where he had been hiding the antiquities, police said].



Linguistically-motivated Compression Rules

Rule	Example
Header	[~~MOSCOW , October 19 (Xinhua) –~~] Russian federal troops Tuesday continued...
Relative dates	...Centers for Disease Control confirmed [~~Tuesday~~] that there was...
Intra-sentential attribution	...fueling the La Nina weather phenomenon, [~~the U.N. weather agency said~~].
Lead adverbials	[~~Interestingly~~], while the Democrats tend to talk about...
Noun appositives	Wayne County Prosecutor [~~John O'Hara~~] wanted to send a message...
Nonrestrictive relative clause	Putin, [~~who was born on October 7, 1952 in Leningrad~~], was elected in the presidential election...
Adverbial clausal modifiers	[~~Given the short time~~], car makers see electric vehicles as...
Within Parentheses	...to Christian home schoolers in the early 1990s [~~(www.homecomputermarket.com)~~].

Flexibility of the Model

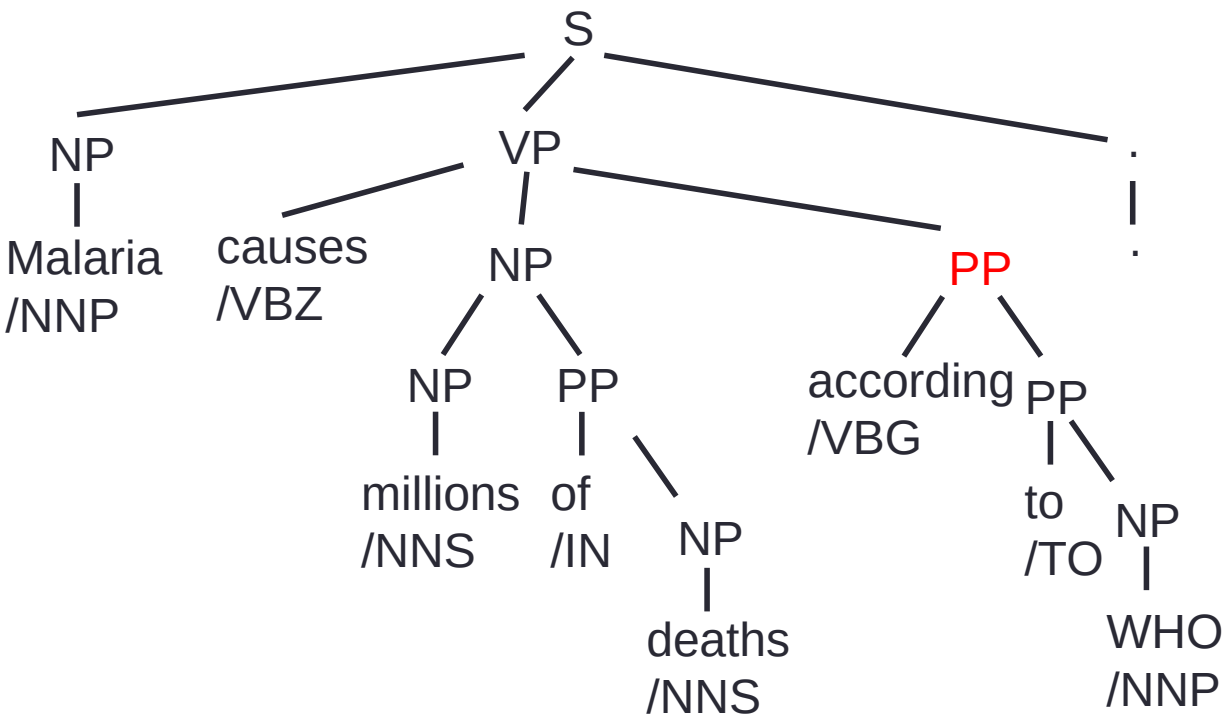
- Tree traversal algorithm
- Scoring function

Flexibility of the Model

- Tree traversal algorithm
  - Basic search (post-order traversal)
  - Context-aware search
  - Head-driven search

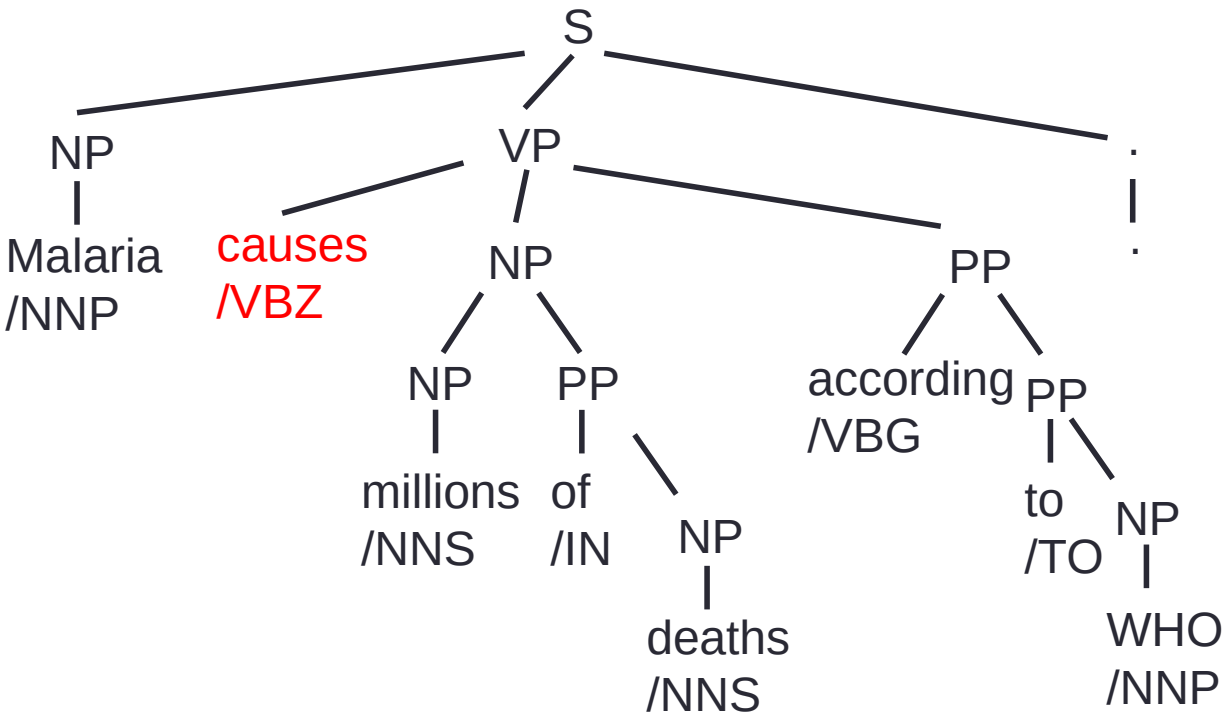
Flexibility of the Model

- Context-aware search



Flexibility of the Model

- Head-driven search



Flexibility of the Model

- Scoring function
- Scorer S_{Basic}

$$S_{Basic}\,(L = \{l_0, l_1, \ldots, l_n\}) = \sum_{i=0}^{n} \log P(l_i | t_i)$$

for a sub-sequence $t_0 t_1 \ldots t_n$.

Flexibility of the Model

- Given the current hypothesis W, query Q

- Query-relevance
  - $S_{query} = |W \cap Q|/Q$
- Importance
  - $S_{importance} = \sum_{i=0}^{|W|} SumBasic(w_i)/|W|$
  - SumBasic score is adopted from Toutanova et al. (2007)
- Language Model
  - $S_{LM} = P(W; \theta_{LM})$
- Cross-Sentence Redundancy
  - $S_{query} = 1 - |W \cap C|/|W|$
- Multi-scorer

$$S_{Multi} = \alpha_0 S_{Basic} + \alpha_1 S_{query} + \alpha_2 S_{importance} + \alpha_3 S_{LM} + \alpha_4 S_{query}$$

Framework

- **Step One**: Sentence Ranking
  - Determines the importance of each sentence given the query.
- **Step Two**: Sentence Compression
  - Iteratively generates the most likely succinct versions of the ranked sentences until a length limit is reached.
- **Step Three**: Post-processing
  - Applies coreference resolution and sentence ordering

Post-processing

- Coreference resolution
  - We replace each pronoun with its referent unless they appear in the same sentence.
- Sentence ordering
  - The sentences are sorted based first on the time stamp, and then the position in the source document.

Summarization Evaluation

- Data
  - Training: Document Understanding Conference (DUC) 2005
  - Test: DUC 2006 and DUC 2007

- Evaluation:
  - Automatic evaluation
    - ROUGE (Lin and Hovy, 2003)
  - Human evaluation
    - Pyramid (Nenkova and Passonneau, 2004)
    - Linguistic quality

Automatic Evaluation

	DUC 2006			DUC 2007		
System	C Rate	R-2	R-SU4	C Rate	R-2	R-SU4
Best DUC system	-	9.56	15.53	-	12.62	17.90

Automatic Evaluation

	DUC 2006			DUC 2007		
System	C Rate	R-2	R-SU4	C Rate	R-2	R-SU4
Best DUC system	-	9.56	15.53	-	12.62	17.90
Davis et al. (2012)	100%	10.2	15.2	100%	12.8	17.5

Automatic Evaluation

	DUC 2006			DUC 2007		
System	C Rate	R-2	R-SU4	C Rate	R-2	R-SU4
Best DUC system	-	9.56	15.53	-	12.62	17.90
Davis et al. (2012)	100%	10.2	15.2	100%	12.8	17.5
SVR	100%	7.78	13.02	100%	9.53	14.69

Automatic Evaluation

	DUC 2006			DUC 2007		
System	C Rate	R-2	R-SU4	C Rate	R-2	R-SU4
Best DUC system	-	9.56	15.53	-	12.62	17.90
Davis et al. (2012)	100%	10.2	15.2	100%	12.8	17.5
SVR	100%	7.78	13.02	100%	9.53	14.69
LambdaMART	100%	9.84	14.63	100%	12.34	15.62

Automatic Evaluation

	DUC 2006			DUC 2007		
System	C Rate	R-2	R-SU4	C Rate	R-2	R-SU4
Best DUC system	-	9.56	15.53	-	12.62	17.90
Davis et al. (2012)	100%	10.2	15.2	100%	12.8	17.5
SVR	100%	7.78	13.02	100%	9.53	14.69
LambdaMART	100%	9.84	14.63	100%	12.34	15.62
Rule-based	78.99%	**10.62**	**15.73**	78.11%	**13.18**	**18.15**

Automatic Evaluation

	DUC 2006			DUC 2007		
System	C Rate	R-2	R-SU4	C Rate	R-2	R-SU4
Best DUC system	-	9.56	15.53	-	12.62	17.90
Davis et al. (2012)	100%	10.2	15.2	100%	12.8	17.5
SVR	100%	7.78	13.02	100%	9.53	14.69
LambdaMART	100%	9.84	14.63	100%	12.34	15.62
Rule-based	78.99%	10.62	15.73	78.11%	13.18	18.15
Tree (Basic+Scorer$_{Basic}$)	**70.48%**	10.49	**15.86**	**69.27%**	13.00	**18.29**

Automatic Evaluation

	DUC 2006			DUC 2007		
System	C Rate	R-2	R-SU4	C Rate	R-2	R-SU4
Best DUC system	-	9.56	15.53	-	12.62	17.90
Davis et al. (2012)	100%	10.2	15.2	100%	12.8	17.5
SVR	100%	7.78	13.02	100%	9.53	14.69
LambdaMART	100%	9.84	14.63	100%	12.34	15.62
Rule-based	78.99%	10.62	15.73	78.11%	13.18	18.15
Tree ($Basic+Scorer_{Basic}$)	70.48%	10.49	15.86	69.27%	13.00	18.29
Tree ($Context+Scorer_{Basic}$)	**65.21%**	10.55	**16.10**	**63.44%**	12.75	18.07

Automatic Evaluation

	DUC 2006			DUC 2007		
System	C Rate	R-2	R-SU4	C Rate	R-2	R-SU4
Best DUC system	-	9.56	15.53	-	12.62	17.90
Davis et al. (2012)	100%	10.2	15.2	100%	12.8	17.5
SVR	100%	7.78	13.02	100%	9.53	14.69
LambdaMART	100%	9.84	14.63	100%	12.34	15.62
Rule-based	78.99%	10.62	15.73	78.11%	13.18	18.15
Tree (Basic +$\text{Scorer}_{\text{Basic}}$)	70.48%	10.49	15.86	69.27%	13.00	18.29
Tree (Context+$\text{Scorer}_{\text{Basic}}$)	65.21%	10.55	16.10	63.44%	12.75	18.07
Tree (Head +$\text{Scorer}_{\text{Basic}}$)	**66.70%**	10.66	**16.18**	**65.05%**	12.93	**18.15**

Automatic Evaluation

	DUC 2006			DUC 2007		
System	C Rate	R-2	R-SU4	C Rate	R-2	R-SU4
Best DUC system	-	9.56	15.53	-	12.62	17.90
Davis et al. (2012)	100%	10.2	15.2	100%	12.8	17.5
SVR	100%	7.78	13.02	100%	9.53	14.69
LambdaMART	100%	9.84	14.63	100%	12.34	15.62
Rule-based	78.99%	10.62	15.73	78.11%	13.18	18.15
Tree (Basic +$\text{Scorer}_{\text{Basic}}$)	70.48%	10.49	15.86	69.27%	13.00	18.29
Tree (Basic +$\text{Scorer}_{\text{Basic}}$)	65.21%	10.55	16.10	63.44%	12.75	18.07
Tree (Context +$\text{Scorer}_{\text{Basic}}$)	66.70%	10.66	16.18	65.05%	12.93	18.15
Tree (Head +$\text{Scorer}_{\text{Multi}}$)	70.20%	**11.02**	**16.25**	73.40%	**13.49**	**18.46**

Human Evaluation

System	Gra	Non-Red	Ref	Foc	Coh
Best DUC system (ROUGE)	22.9±8.2	3.5±0.9	3.5±1.0	3.6±1.0	2.9±1.1
Best DUC system (LQ)	4.0±0.8	4.2±0.7	3.8±0.7	3.6±0.9	3.4±0.9
Our System	3.0±0.9	4.0±1.1	3.6±1.0	3.4±0.9	2.8±1.0

- Linguistic Quality (LQ)

Grammaticality (Gra)
Non-redundancy (Non-Red)
Referential clarity (Ref)
Focus (Foc)
Structure and Coherence (Coh)

Human Evaluation

System	Gra	Non-Red	Ref	Foc	Coh	Pyr
Best DUC system (ROUGE)	22.9±8.2	3.5±0.9	3.5±1.0	3.6±1.0	2.9±1.1	22.9±8.2
Best DUC system (LQ)	4.0±0.8	4.2±0.7	3.8±0.7	3.6±0.9	3.4±0.9	-
Our System	3.0±0.9	4.0±1.1	3.6±1.0	3.4±0.9	2.8±1.0	**26.4**±10.3

- Linguistic Quality (LQ)

Grammaticality (Gra)
Non-redundancy (Non-Red)
Referential clarity (Ref)
Focus (Foc)
Structure and Coherence (Coh)

Compression Evaluation

- 1188 sentences for training and 441 sentences for testing from the dataset (Clarke and Lapata, 2008).
- Compare with
  - Dorr et al. (2003), Hedge Trimmer
  - McDonald (2006), discriminative learning with soft syntactic evidence
  - Martins and Smith (2009), dependency-tree based compressor
- Evaluation metrics
  - Unigram precision/recall/F1 (Martins and Smith, 2009)
  - Dependency relations (Clarke and Lapata, 2008).

Compression Results

System	C Rate	Uni-Prec	Uni-Rec	Uni-F1	Rel-F1
HedgeTrimmer	57.64%	0.72	0.65	0.64	0.50
McDonald (2006)	70.95%	0.77	0.78	0.77	0.55
Martins and Smith (2009)	71.35%	0.77	0.78	0.77	0.56

Compression Results

System	C Rate	Uni-Prec	Uni-Rec	Uni-F1	Rel-F1
HedgeTrimmer	57.64%	0.72	0.65	0.64	0.50
McDonald (2006)	70.95%	0.77	0.78	0.77	0.55
Martins and Smith (2009)	71.35%	0.77	0.78	0.77	0.56
Rule-based	87.65%	0.74	0.91	0.80	0.63

Compression Results

System	C Rate	Uni-Prec	Uni-Rec	Uni-F1	Rel-F1
HedgeTrimmer	57.64%	0.72	0.65	0.64	0.50
McDonald (2006)	70.95%	0.77	0.78	0.77	0.55
Martins and Smith (2009)	71.35%	0.77	0.78	0.77	0.56
Rule-based	87.65%	0.74	0.91	0.80	0.63
Tree (BASIC)	69.65%	0.77	0.79	0.75	0.56
Tree (CONTEXT)	67.01%	**0.79**	0.78	0.76	0.57
Tree (HEAD)	68.06%	**0.79**	0.80	0.77	**0.59**

Sample System Output

- ***Query**: How were the bombings of the US embassies in Kenya and Tanzania conducted? What terrorist groups and individuals were responsible? How and where were the attacks planned?*

- WASHINGTON, August 13 (Xinhua) – President Bill Clinton Thursday condemned terrorist bomb attacks at U.S. embassies in Kenya and Tanzania and vowed to find the bombers and bring them to justice. Clinton met with his top aides Wednesday in the White House to assess the situation following the twin bombings at U.S. embassies in Kenya and Tanzania, which have killed more than 250 people and injured over 5,000, most of them Kenyans and Tanzanians. Local sources said the plan to bomb U.S. embassies in Kenya and Tanzania took three months to complete and bombers destined for Kenya were dispatched through Somali and Rwanda. …

Conclusion

- We have presented a framework for query-focused multi-document summarization based on sentence compression.

- We show substantial improvement over pure extraction-based methods and state-of-the-art systems in both automatic and human evaluation.

Thank you!