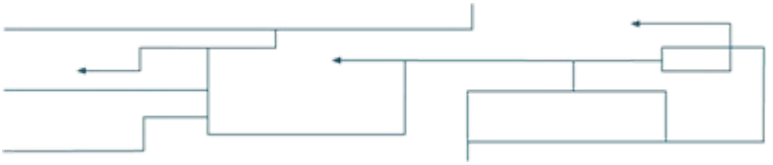
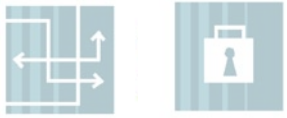
DAI-Labor
TU Berlin



Identifying Sentence-Level Semantic Content Units with Topic Models

Leonhard Hennig, Thomas Strecker, Sascha Narr,
Ernesto William De Luca, Sahin Albayrak
DAI-Labor, TU Berlin



CC IRML
Information Retrieval
& Machine Learning

Agenda

Content modeling for multi-document summarization

Gold-standard content units

A topic modeling approach to content unit discovery

Evaluation

- Word distribution similarity
- Sentence distribution similarity

Conclusion & Outlook

Summarize a set of documents related to a single event or topic

Challenge: Identification of similar information

- frequency indicates importance
- avoid redundancy

Typically solved by sentence extraction

- Compute sentence relevance based on features of interest
- Rank and select subset of important and non-redundant sentences

Identification of similar information

DAI-Labor TU Berlin

Previous Work:

Sentence clustering

Word frequency

Graph-based ranking

Maximum Marginal Relevance

Problems:

- Mostly lexical overlap, Vector Space Model
- Content unit granularity: sentence

But:

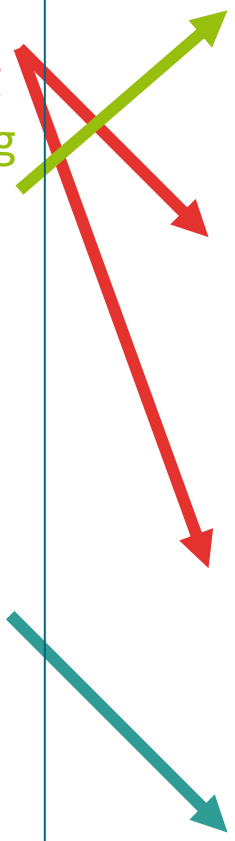
- Content units can be as 'small' as clauses (e.g. facts)
- Content units can be combined differently into sentences

Example: Similar content units

DAI-Labor
TU Berlin

NEW YORK, July 17 --
The U.S. Coast Guard and the Air National Guard are conducting a massive search off the coast of Long Island, N.Y. for a small plane carrying John F. Kennedy Jr., son of the 35th U.S. President, U.S. media reported Saturday.

The search began Saturday morning in an area covering some 1,000 square miles, presumably the flight path of Kennedy's plane, searchers said.

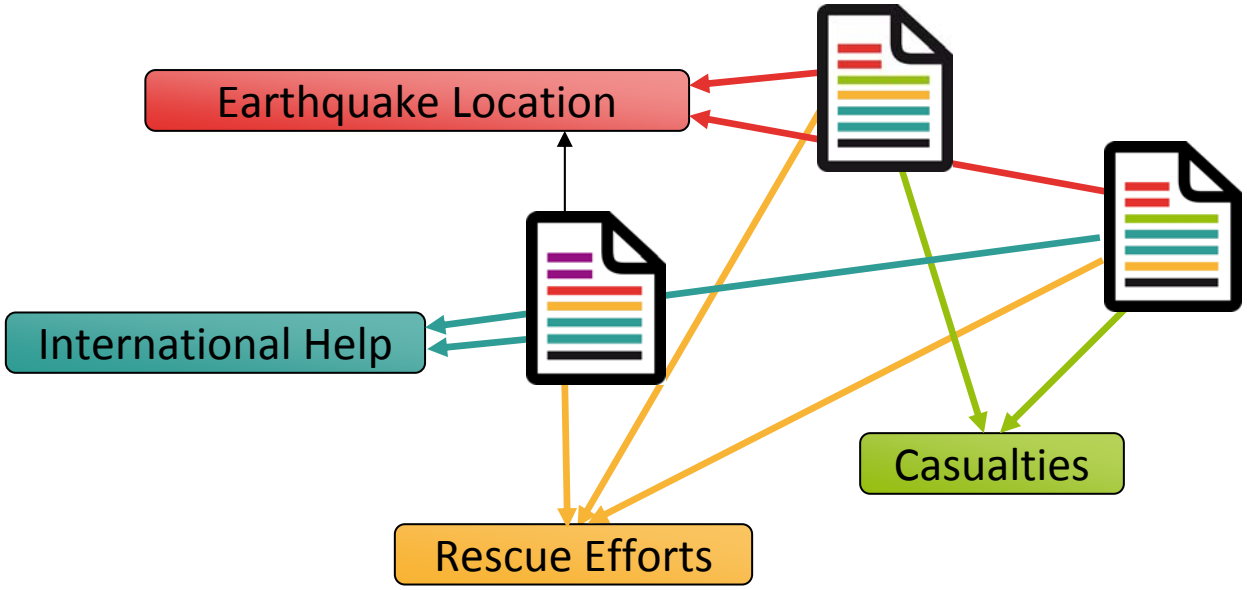


WASHINGTON, July 17 --
A small plane carrying John F. Kennedy Jr., son of the former U.S. president, was reported missing early Saturday, and a search was under way off the coast of New York's Long Island, official sources said.

The U.S. Coast Guard confirmed it was searching for the plane with help from the Air National Guard.
The search was being conducted in water off eastern tip of Long Island, along the presumed flight path of Kennedy's plane.

Observations

- Documents from same domain often exhibit similar sub-topic structuring and similar word patterns (Barzilay 04)



- Similar content units in human reference summaries are often expressed using similar word patterns, but with variations in word choice (e.g. synonymy) and word order (Nenkova 04, Harnly 05)

Discovery of similar content units

DAI-Labor TU Berlin

Can we automatically discover these similar content units?

Challenge 1: Semantic similarity

- Synonymy and polysemy
- Variations in word order and word choice

Challenge 2: (Sub-) sentential content units

- Granularity: sentence or clause
- Combined differently into sentences

Agenda

DAI-Labor
TU Berlin

Content modeling for multi-document summarization

Gold-standard content units

A topic modeling approach to content unit discovery

Evaluation

- Word distribution similarity
- Sentence distribution similarity

Conclusion & Outlook

Types of gold-standard content units

DAI-Labor TU Berlin

Copies

- Sentences are verbatim copies of each other

Similar

- Sentences express the same content

(a) The supreme court struck down as unconstitutional a law giving the president a line-item veto which lets him cancel specific items in tax and spending measures.

(b) The U.S. supreme court Thursday struck down as unconstitutional the line-item veto law that lets the U.S. president strike out specific items in tax and spending measures.

Types of gold-standard content units

DAI-Labor TU Berlin

Clause

- Sentence clauses are repeated verbatim or with similar word usage as clauses of other sentences

(a) Germany, Azerbajan, Greece, France, the Netherlands, Kazakhstan, Ukraine and Russia have been participating in the fight against the blaze that threatened to engulf the entire field of 30 storage tanks containing 1 million tons of crude oil.

(b) However, he said the strong fire had destroyed seven storage tanks and damaged two other ones in the refinery which held 30 storage tanks containing 1 million tons of crude oil.

(c) Germany, Azerbaijan, Greece, France, the Netherlands, Kazakistan, Ukraine and Russia participated in the fight against the blaze.

Unique

- Sentences express information not repeated in any other sentence

Data set

11 closely related document pairs selected from DUC 2007 multi-document summarization data set

3 annotators

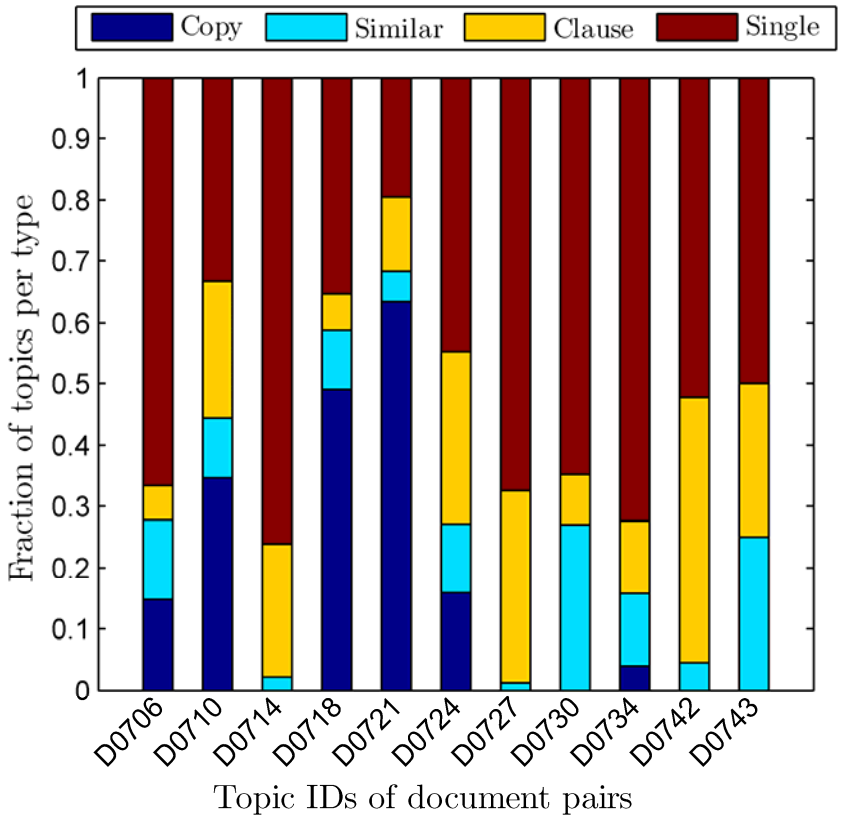
$\hat{\Theta}^{(a)} =$ Content Units

	s_1	s_2	...	s_{m-1}	s_m
z'_1	1	1		0	0
z'_2	0	1		1	0
...					
z'_{k-1}	0	0		1	0
z'_k	0	0		0	1

Sentences

- Inter-annotator agreement on number of content units: **0.97**
- Inter-annotator agreement on content units: **0.69**
  (i.e. agreement on "identical" content units)

Distribution of gold-standard content units



Distribution of types

53 % unique

20% clause

11% similar

ML estimate for content unit word likelihood

$$\hat{\Phi} = P(w \mid \hat{z}^{(a)}), where$$

$$p(w_i \mid \hat{z}_k^{(a)}) = \frac{w_{is}}{\sum_{s \in \hat{z}_k^{(a)}} \sum_i w_{is}}$$

Agenda

DAI-Labor
TU Berlin

Content modeling for multi-document summarization

Gold-standard content units

A topic modeling approach to content unit discovery

Evaluation

- Word distribution similarity
- Sentence distribution similarity

Conclusion & Outlook

DAI-Labor TU Berlin

Utilize Latent Dirichlet Allocation (LDA, Blei et al. 2003)

- Exploit word-document co-occurrence patterns
- Model each document as a distribution over topics, each topic as a distribution over words

For each document pair (d_1,d_2):

- Create Term-**Sentence** Matrix $\mathbf{A}$
- $\mathbf{A}_{ij}$ = frequency of term *i* in sentence *j*
- $M = \left| s \in d_1 \cup s \in d_2 \right|$

$\mathbf{A} =$ Terms

	s_1	s_2	...	s_{m-1}	s_m
w_1	1	2		0	0
w_2	1	0		1	0
...					
w_{n-1}	0	0		0	3
w_n	0	1		0	1

Sentences

Gibbs sampling, 2000 iterations

Compute p($w|z$) and p($z|s$) from single sample

$\Phi =$ Terms

	z_1	z_2	...	z_{k-1}	z_k
w_1	0.02	0.03		0.0	0.0
w_2	0.02	0.0		0.018	0.0
...					
w_{n-1}	0.0	0.0		0.0	0.04
w_n	0.0	0.015		0.0	0.017

Topics

$\Theta =$ Topics

	s_1	s_2	...	s_{m-1}	s_m
z_1	0.0	0.6		0.0	0.51
z_2	0.98	0.38		0.0	0.0
...					
z_{k-1}	0.0	0.0		0.98	0.0
z_k	0.0	0.0		0.0	0.48

Sentences

Agenda

Content modeling for multi-document summarization

Gold-standard content units

A topic modeling approach to content unit discovery

Evaluation

- Word distribution similarity
- Sentence distribution similarity

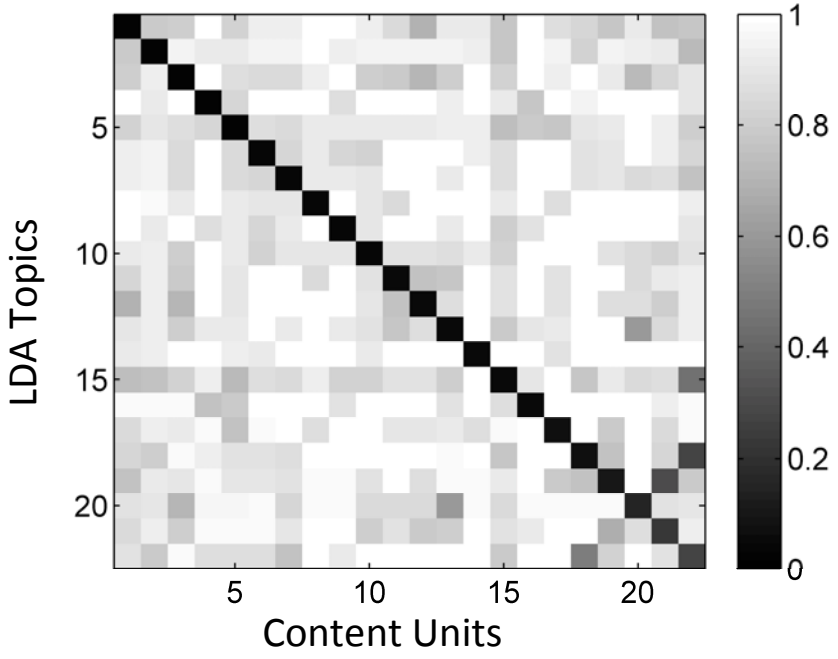
Conclusion & Outlook

Word distribution similarity 1/2

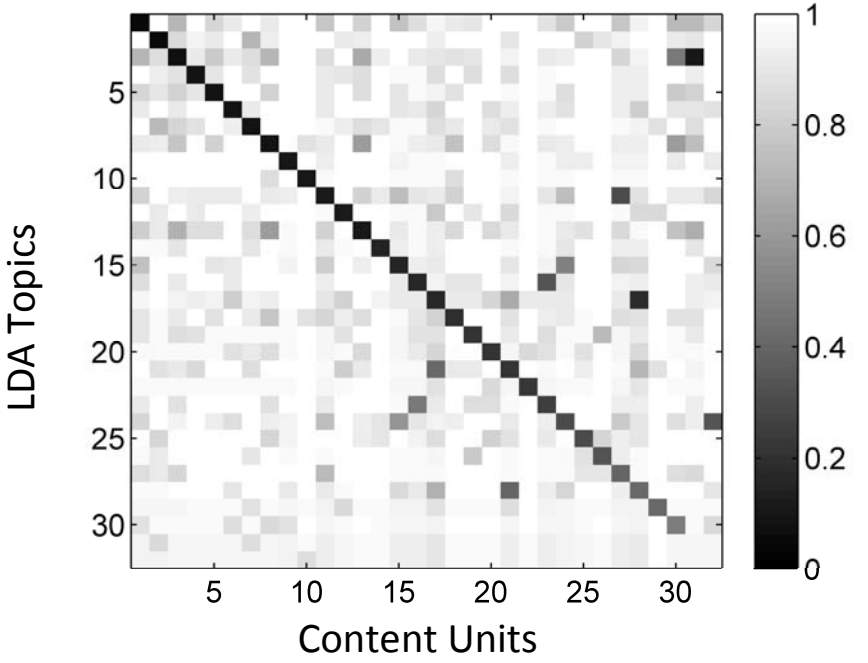
Pair-wise comparison of word distributions $\hat{\Phi}, \Phi$ of annotated content units and latent topics using Jensen-Shannon divergence

- Diagonal shows matches of most similar topic-content unit pairs
- Lower (darker) is better
- Reordered so that best matches are in upper left

Document pair from D0710



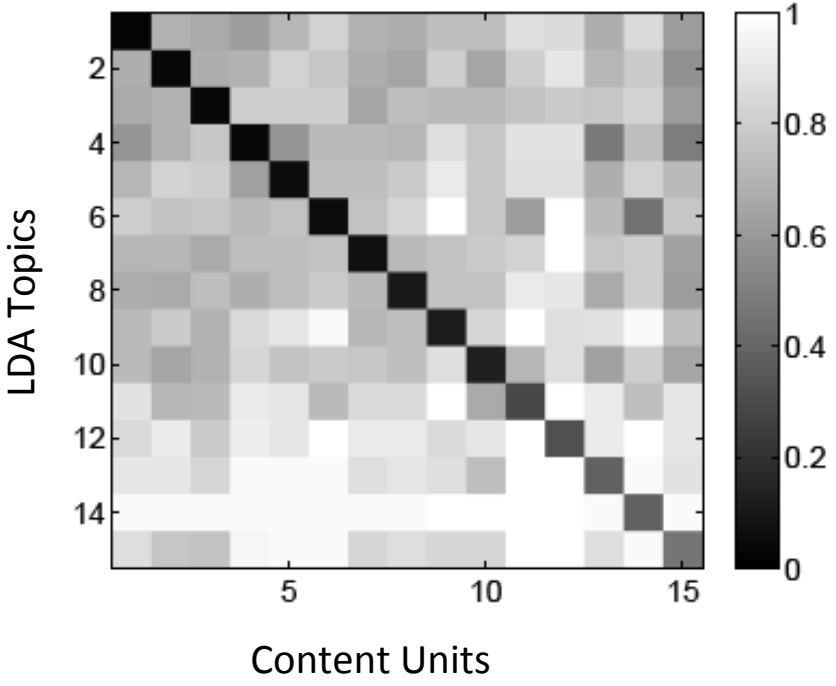
Document pair from D0734



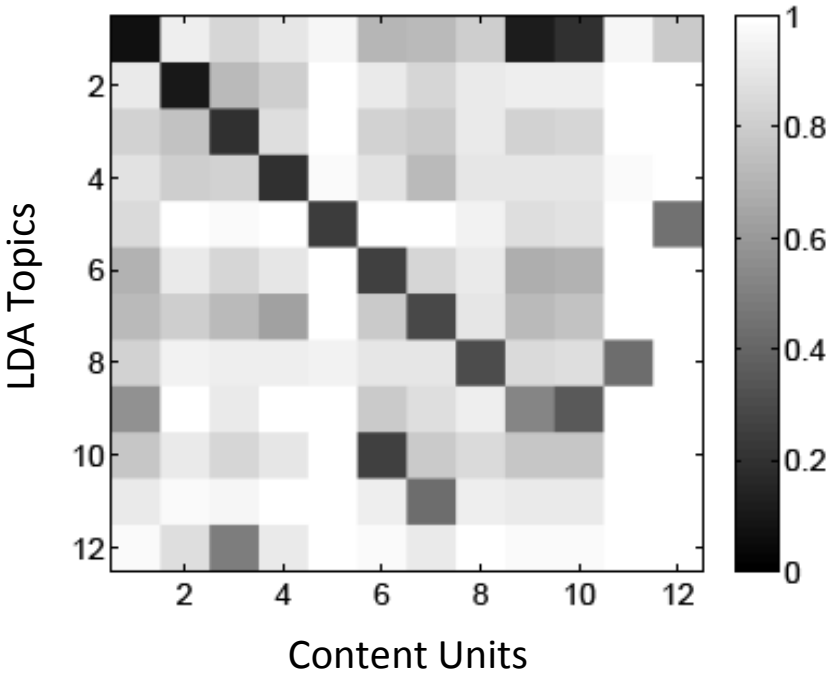
Word distribution similarity 2/2

LDA does not always result in topics matching well with manually annotated content units

Document pair from D0730



Document pair from D0742



Agenda

Content modeling for multi-document summarization

Gold-standard content units

A topic modeling approach to content unit discovery

Evaluation

- Word distribution similarity
- Sentence distribution similarity

Conclusion & Outlook

DAI-Labor TU Berlin

Compare sentence-topic assignments

$\Theta =$ (Topics)

	s_1	s_2	...	s_{m-1}	s_m
z_1	0.0	0.6		0.0	0.51
z_2	0.98	0.38		0.0	0.0
...					
z_{k-1}	0.0	0.0		0.98	0.0
z_k	0.0	0.0		0.0	0.48

Sentences

$\hat{\Theta}^{(a)} =$ (Content Units)

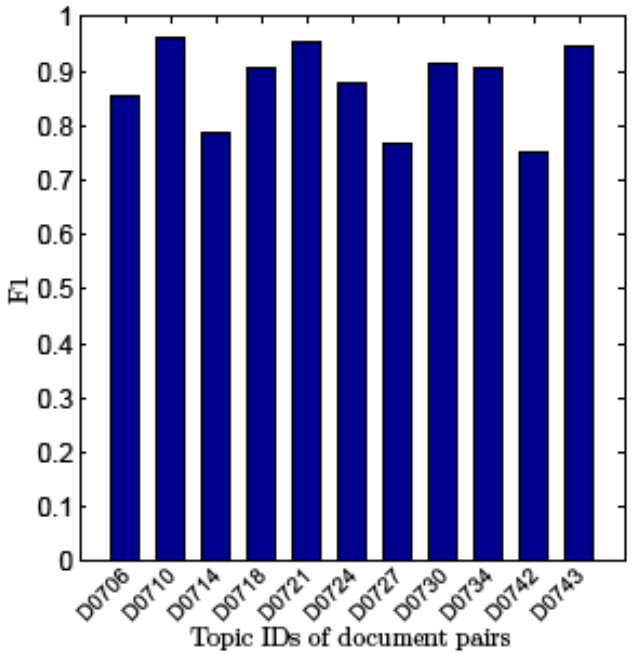
	s_1	s_2	...	s_{m-1}	s_m
z'_1	0	1		0	1
z'_2	1	0		0	0
...					
z'_{n-1}	0	0		1	0
z'_n	0	0		1	1

Sentences

Binarize Θ to compute per-topic precision and recall:

$$\Theta'_{ij} = \begin{cases} 1, & \Theta_{ij} \geq \varepsilon \\ 0, & \Theta_{ij} < \varepsilon \end{cases}$$

Topic-sentence distribution similarity

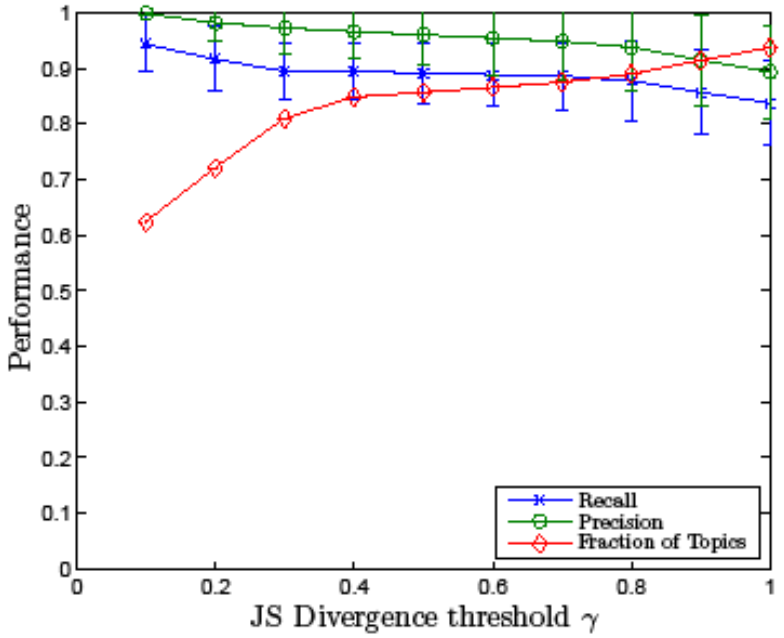


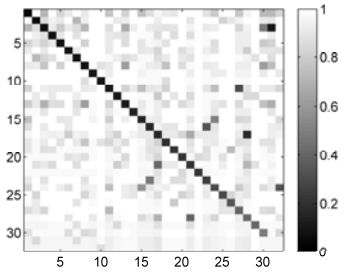
F1, averaged across annotators and using all topic-content unit matches

Mean F1 = 0.88

- Suggests that LDA identifies the same "associations" of topics and sentences as human annotators

Topic-sentence distribution similarity

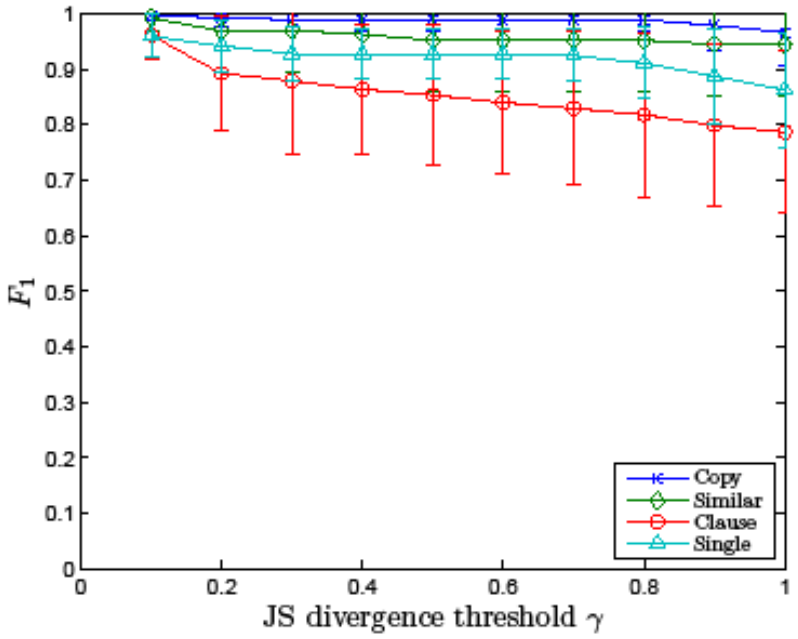




Precision and recall, averaged over all document pairs and annotators

- Many matches have low JS divergence (approx. 80% $\leq$ 0.3)
- Precision and recall decrease as matches of lower quality are added

Topic-sentence distribution per content unit type



F1, averaged over all document pairs and annotators, per content unit type

Copies and ***Similar*** are easiest to identify

Clause is hardest due to word "noise"

Overall F1 quite high

Conclusion & Future Work

DAI-Labor TU Berlin

LDA learns topics similar to manually annotated content units

- unsupervised
- domain- and language-independent
- benefits of probabilistic models

Initial results seem promising

- phrasal / sentence meaning similarity remains open (e.g. influence of word order, negation, ...)

Representation of sentences in terms of content units is a step towards bridging the gap between word and sentence importance modeling in multi-document summarization?

Thank you.

DAI-Labor
TU Berlin

Questions?