Fine Grained Citation Span for References in Wikipedia

Besnik Fetahu, Katja Markert, and Avishek Anand

emnlp2017

Leibniz Universität Hannover

Citations in Wikipedia

Citation Span Cases

Sub-Sentence

Obama was born on August 4, 1961,[5] at Kapi‘olani Maternity & Gynecological Hospital in Honolulu, Hawaii.[6][7][8]

Sentence

On February 10, 2007, Obama announced his candidacy for President of the United States in front of the Old State Capitol building in Springfield, Illinois.[158][159] […] Obama emphasized issues of rapidly ending the Iraq War, increasing energy independence, and reforming the health care system,[161] in a campaign that projected themes of hope and change.[162]

Multi-Sentence

At the Democratic National Convention in Charlotte, North Carolina, Obama and Joe Biden were formally nominated by former President Bill Clinton as the Democratic Party candidates for president and vice president in the general election. Their main opponents were Republicans Mitt Romney, the former governor of Massachusetts, and Representative Paul Ryan of Wisconsin.[183]

Motivation and Problem

- Citations in Wikipedia do not have an explicit span as to what textual fragment they cover
- Text in Wikipedia constantly changes and as such the citation span is subject to that change
- V*erifiability* principle in Wikipedia states that statements/facts should point to external references
- Manual labor by editors is required to flag something as requiring a citation

Citation Span Properties

- What are some of the properties of this problem?
    - Popular entities have higher density of citations
    - Citation span varies heavily w.r.t the popularity of entities
    - Citations span can be from a sub-sentence, sentence, or entire paragraph
    - Citations do not span across paragraphs in Wikipedia

Citation Span Task

Citation Span Task

Citing Paragraph

He was reelected to the Illinois Senate in 1998, defeating Republican Yesse Yehudah in the general election, and was re-elected again in 2002.[117] In 2000, he lost a Democratic primary race for Illinois's 1st congressional district in the United States House of Representatives to four-term incumbent Bobby Rush by a margin of two to one.[118]

Chunk Paragraph

Textual Fragments

He was reelected to the Illinois Senate in 1998, defeating Republican Yesse Yehudah in the general election, and was re-elected again in 2002.[117] In 2000, he lost a Democratic primary race for Illinois's 1st congressional district in the United States House of Representatives to four-term incumbent Bobby Rush by a margin of two to one.[118]

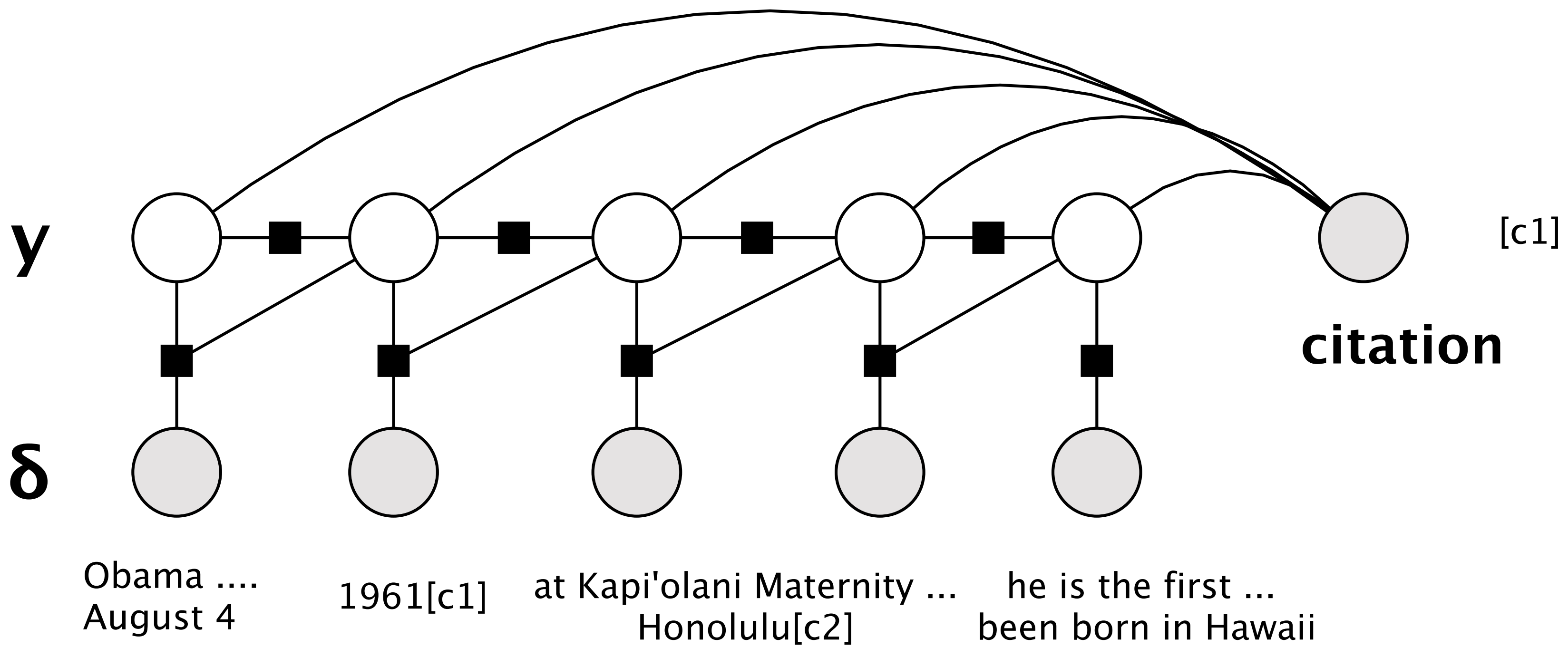
Citation Span for **[117]**

Citation Span

He was reelected to the Illinois Senate in 1998, defeating Republican Yesse Yehudah in the general election, and was re-elected again in 2002.[117] In 2000, he lost a Democratic primary race for Illinois's 1st congressional district in the United States House of Representatives to four-term incumbent Bobby Rush by a margin of two to one.[118]

Citation Span Approach

Sequence Classification — model the citation span detection as a sequence classification problem, in this case sequences are text fragments from a citing paragraph.



Structural Features

Structural Features

Chunked Citing Paragraph

He was reelected to the Illinois Senate in 1998, defeating Republican Yesse Yehudah in the general election, and was re-elected again in 2002.[117] In 2000, he lost a Democratic primary race for Illinois's 1st congressional district in the United States House of Representatives to four-term incumbent Bobby Rush by a margin of two to one.[118]

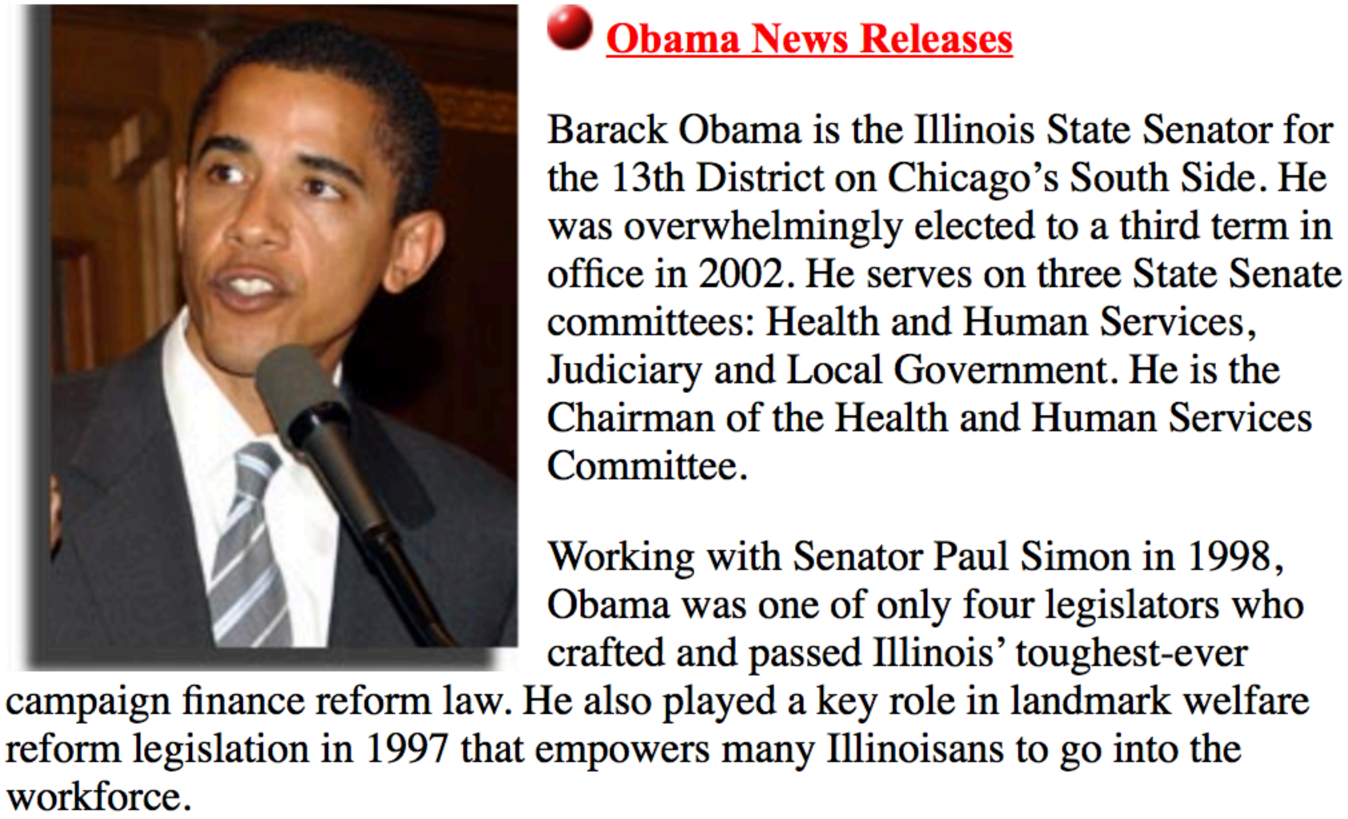
structural features for ***c*****=[117]**

δ	*other **c'***	#chars	same sentence as ***c***	same sentence as previous δ	distance to ***c***
1	0	48	1	0	2
2	0	60	1	1	1
3	0	35	1	1	0
4	0	7	0	0	-1
5	1	230	0	1	-2

Citation Features

Citation Features

c_{117}=[https://web.archive.org/web/20040802233730/http://www.senatedem.state.il.us/obama/index.html]



Obama News Releases

Barack Obama is the Illinois State Senator for the 13th District on Chicago's South Side. He was overwhelmingly elected to a third term in office in 2002. He serves on three State Senate committees: Health and Human Services, Judiciary and Local Government. He is the Chairman of the Health and Human Services Committee.

Working with Senator Paul Simon in 1998, Obama was one of only four legislators who crafted and passed Illinois' toughest-ever campaign finance reform law. He also played a key role in landmark welfare reform legislation in 1997 that empowers many Illinoisans to go into the workforce.

1. He was reelected to the Illinois Senate in 1998
2. defeating Republican Yesse Yehudah in the general election
3. and was re-elected again in 2002.
4. In 2000
5. he lost a Democratic primary race for Illinois's 1st congressional district in the United States House of Representatives to four-term incumbent Bobby Rush by a margin of two to one.

$$P(w|M_{w_i}) = \frac{tf_{w,\phi(w_i)}}{\sum_{w' \in \phi(w_i)} tf_{w',\phi(w_i)}}$$

score a word based on moving language model taking into account words appearing in a window of +/- 3 words

$$f_i^{LM} = \min_{p \in c} \left[- \sum_{w \in \delta} P(w|M_\delta) \log \frac{P(w|M_\delta)}{P(w|M_p)} \right]$$

score the LM of a sequence against the LM of a paragraph from ***c***

Bilyana Taneva, Gerhard Weikum: **Gem-based entity-knowledge maintenance. CIKM 2013**: 149-158

Discourse Features

Discourse Features

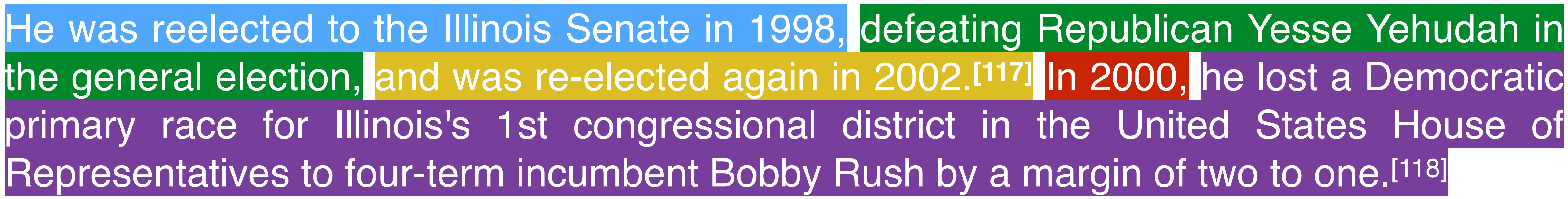
- Discourse sense annotation of sentences in a paragraph (*Pitler and Nenkova, ACL 2009*):
    - *temporal*
    - *expansion*
    - *contingency*
    - *comparison*

- Discourse helps tying together different *fragments* and can serve as indicators for transition probabilities between different states (*covered*/*uncovered*) in the linear chain.

Temporal Features

Temporal Features

- Extract temporal expressions from text fragments
- Sequences that point to far apart time points, there is a higher likelihood to have different states in the linear chain CRF
- E.g. "*and was re-elected again in 2002*" and "*In 2000*" are *unlikely* to be *covered* by the same citation



He was reelected to the Illinois Senate in 1998, defeating Republican Yesse Yehudah in the general election, and was re-elected again in 2002.[117] In 2000, he lost a Democratic primary race for Illinois's 1st congressional district in the United States House of Representatives to four-term incumbent Bobby Rush by a margin of two to one.[118]

Experimental Setup

Dataset & Ground-truth

- We randomly sample a set of ~500 citing paragraphs (pointing at *web* or *news* references)

- Manually annotate (single author of the paper) each sub-sentence whether it is *explicitly* or *implied/inferred* by the content in the citation

- On a sample of 10% of the dataset we assess the annotator agreement (with a second annotator), yielding κ=0.84

Ground-truth Statistics

Citation Span Dataset Attributes						
span	**news**			**web**		
	dist.	*skip frag.*	*skip sent.*	*dist.*	*skip frag.*	*skip sent.*
<= 0.5	11%	6%	-	6.8%	-	-
(.5,1]	63%	-	-	63%	-	1%
(1,2]	17%	-	8%	14%	-	19%
(2,5]	7%	5%	18%	13.1%	-	21%
>5	1.8%	-	20%	3.1%	-	67%

Evaluation Metrics

- Mean Average Precision — **MAP**:

$$MAP = \frac{1}{|N|} \sum_{p \in N} \frac{|S' \cap S^t|}{|S'|}$$

Sequence marked as covered by our approach (baselines)

Sequence marked as covered in the *ground-truth*

- Recall — **R:**

$$R = \frac{1}{|N|} \sum_{p \in N} \frac{|S' \cap S^t|}{|S^t|}$$

- Erroneous Span (*word*, and *sequence* level) — **Δ**:

$$\Delta_w = \frac{1}{|N|} \sum_{p \in N} \frac{\sum_{\delta \in \mathcal{S}' \setminus \mathcal{S}^t} words(\delta)}{\sum_{\delta \in \mathcal{S}^t} words(\delta)}$$

Baselines

Citing Sentence — **CS**:

At the Democratic National Convention in Charlotte, North Carolina, Obama and Joe Biden were formally nominated by former President Bill Clinton as the Democratic Party candidates for president and vice president in the general election. Their main opponents were Republicans Mitt Romney, the former governor of Massachusetts, and Representative Paul Ryan of Wisconsin.[183]

Inter-Citation Text — **IC:**

He was reelected to the Illinois Senate in 1998, defeating Republican Yesse Yehudah in the general election, and was re-elected again in 2002.[117] In 2000, he lost a Democratic primary race for Illinois's 1st congressional district in the United States House of Representatives to four-term incumbent Bobby Rush by a margin of two to one.[118]

Citation Sentence Window — **CSW**: include sentences in a window of +/- 2 sentences if certain cue words (e.g. *this, these, above-mentioned*) appear in a specific position in the sentences.

Markov Random Fields — **MRF**: compute sentence similarity in a citing paragraph, then measures their *compatibility* (or similarity) to sentences in the cited reference. (*Qazvinian and Radev, ACL 2010*).

Approaches

- Citation Span Classification — **CSPC**: Random Forest classifier trained based on the proposed features.
- Citation Span Structured — **CSPS**: Linear-chain conditional random fields trained based on the proposed features.

Evaluation Results

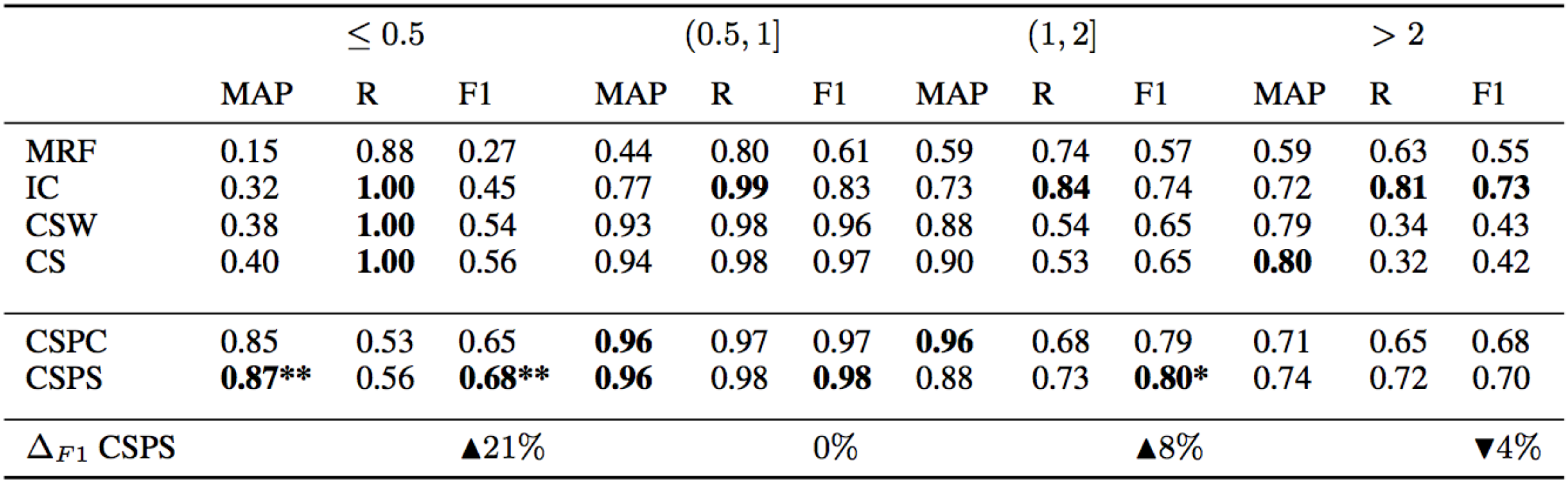
Citation Span Robustness

Citation span evaluation results across all citation span classes

	MAP	R	F1	Δ_w	Δ_δ
MRF	0.45	0.78	0.56	308%	278%
IC	0.72	**0.94**	0.77	113%	115%
CSW	0.85	0.84	0.82	38%	31%
CS	**0.86**	0.84	0.82	35%	27%
CSPC	0.83	0.66	0.74	**12%**	**12%**
CSPS	0.83	0.69	0.75	15%	14%

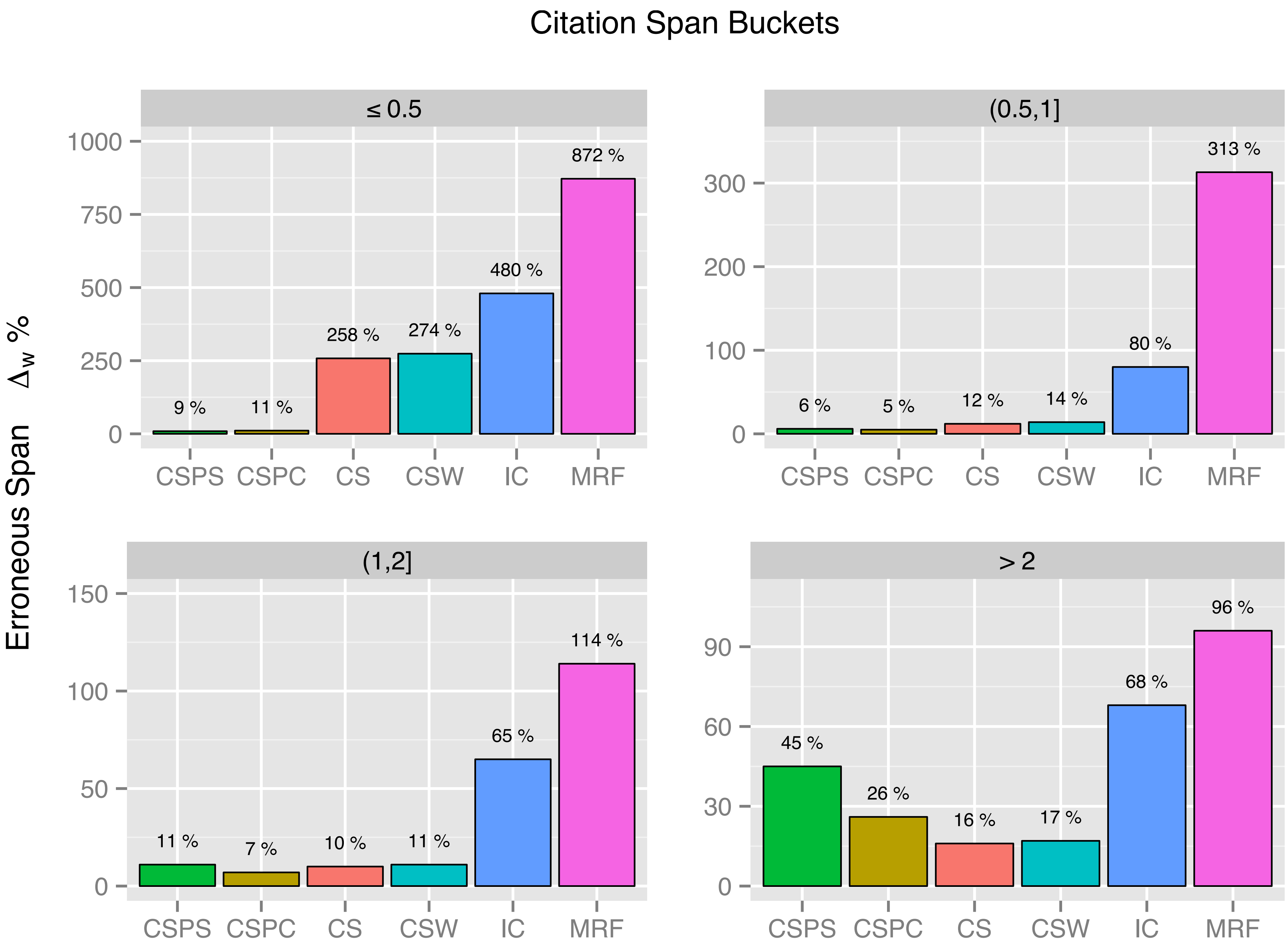
Citation Span Analysis — Accuracy

Citation Span results decomposed across the different span cases.



	≤ 0.5			$(0.5, 1]$			$(1, 2]$			> 2		
	MAP	R	F1	MAP	R	F1	MAP	R	F1	MAP	R	F1
MRF	0.15	0.88	0.27	0.44	0.80	0.61	0.59	0.74	0.57	0.59	0.63	0.55
IC	0.32	**1.00**	0.45	0.77	**0.99**	0.83	0.73	**0.84**	0.74	0.72	**0.81**	**0.73**
CSW	0.38	**1.00**	0.54	0.93	0.98	0.96	0.88	0.54	0.65	0.79	0.34	0.43
CS	0.40	**1.00**	0.56	0.94	0.98	0.97	0.90	0.53	0.65	**0.80**	0.32	0.42
CSPC	0.85	0.53	0.65	**0.96**	0.97	0.97	**0.96**	0.68	0.79	0.71	0.65	0.68
CSPS	**0.87****	0.56	**0.68****	**0.96**	0.98	**0.98**	0.88	0.73	**0.80***	0.74	0.72	0.70
Δ_{F1} CSPS			▲21%			0%			▲8%			▼4%

Citation Span Analysis — Erroneous Span



Conclusion

Conclusion and Future Work

- Citation span can be accurately determined for *web* and *news* references in Wikipedia.
- Optimal performance is achieved for citation with a span of up to 2 sentences.
- Sequence classification outperforms plain classification approaches in this task.
- An initial attempt on solving citation span for references in Wikipedia, which needs to be expanded to other reference types
- A ground-truth dataset for similar experiments, however, a larger ground-truth (highly time consuming process) is necessary.

Thank you!

@FetahuBesnik
http://l3s.de/~fetahu/emnlp17