Contextualizing Citations for Scientific Summarization using Word Embeddings and Domain Knowledge

Arman Cohan and Nazli Goharian

Information Retrieval Lab, Dept. of Computer Science

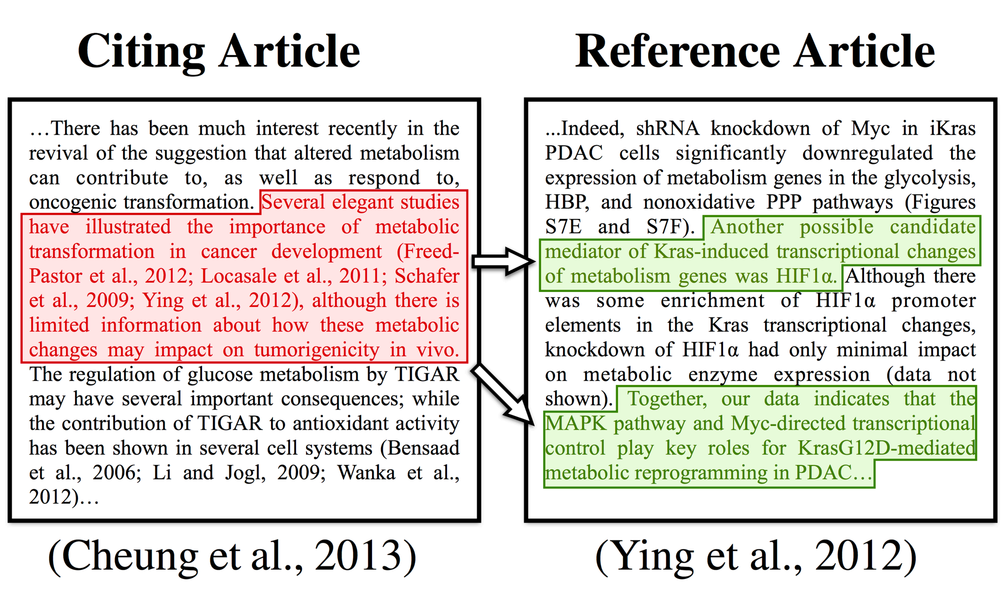
Georgetown University

{arman,nazli}@ir.cs.georgetown.edu

Motivation & Background

- Citation texts are not always accurate
- They lack the context from the reference paper
- Major issue in medical domain
    - Author stated: "Drug can cure cancer"
    - Citation says: "The drug cures cancer"

- Solution?
- Adding context of the reference paper to the citations
    - Verifying the claim of the citation text



- Using a set of citation texts to summarize a reference paper
    - Adding context to citations improves summarization performance
- Challenges:
    - Terminology variations
    - Paraphrasing

Contextualizing Citations

- Extend the Language Modeling for IR by incorporating *word embeddings* and *domain specific knowledge*

$$(*)\quad p(q_i|d) = \frac{f(q_i,d) + \mu p(q_i|C)}{\sum_{w \in V} f(w,d) + \mu}$$

f is the frequency function
problems: *d* is short, terminology variation

- ***word embeddings***: replace *f* in ($*$) with a function that captures semantic relatedness between the query (citation text) and document (reference text).

$$f(q_i,d) = \sum_{d_j \in d} s(q_i,d_j)$$

$$s(q_i,d_j) = \begin{cases} \overset{3}{\phi}(\overset{1}{e(q_i).e(d_j)}); & \text{if } e(q_i).e(d_j) > \overset{2}{\tau} \\ 0; & \text{otherwise} \end{cases}$$

1 Captures semantic similarity based on word embeddings: $e(q_i).e(d_j)$: similarity based on dot product of embeddings. ϕ: transformation function (see 3)

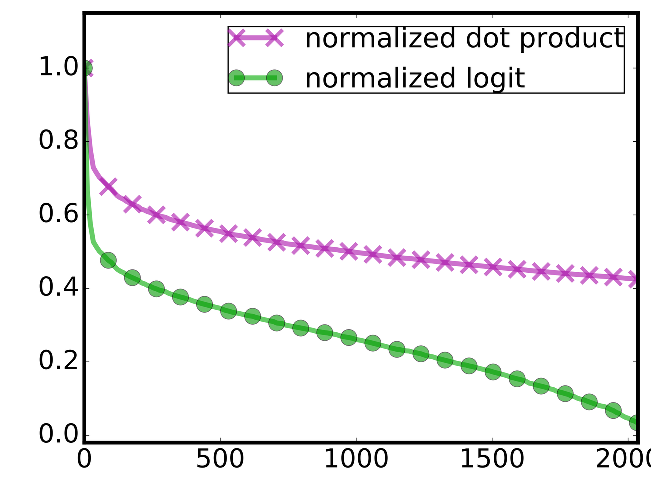
word 1	word 2	Similarity
marker	mint	0.11
notebook	sky	0.07
capture	promotion	0.12
blue	sky	0.31
produce	make	0.43

2 Filter out the noise (less similar words)

3 The similarity values do not differentiate well between highly related words and less related words

ϕ dampens the effect of less similar words (similarity should fall quickly as we move to less similar words):

$$\phi(x) = \log(\frac{x}{1-x})$$



Dot product of embeddings and its logit for a sample word and its top most similar words, x axis: n-th similar word, y axis: similarity value

- **Domain specific knowledge**:
    1) Retrofitting (Faruqui, et all 2015): Modify embeddings using ontology
        - Bring embeddings of similar words (according to an ontology) closer to each other in the embedding space
        - We use MESH and PO ontologies to capture relationships in the biomedical domain
    2) Directly interpolate into the language model

$$p(q_i|d) = \lambda p_1(q_i|d) + (1-\lambda)p_2(q_i|d)$$

p_1 and p_2 are according to ($*$)
Except that p_2 uses the following similarity function f_2

$$f_2(q_i,d) = \sum_{d_j \in d} s_2(q_i,d_j);\ s_2(q_i,d_j) = \begin{cases} 1, & \text{if } q_i = d_j \\ \gamma, & \text{if } q_i \approx d_j \\ 0, & \text{o.w.} \end{cases}$$

Experiments

- TAC 2014 Summarization dataset (20 reference articles, 313 citations)
- **Intrinsic evaluation**: Compare the retrieved references with gold annotations

Contextualization method	Character offset overlap				Similarity by ROUGE			Char offset precision for top K	
	c-P	c-R	c-F	nDCG	RG1	RG2	RG3	c-P@1	c-P@5
BM25 (Jones et al., 2000)	19.5	18.6	17.8	38.1	43.6	23.2	16.3	25.5	24.2
DESM (Mitra et al., 2016)	20.3	23.8	22.3	45.6	50.3	26.2	20.6	32.5	26.5
VSM (Cohan et al., 2015)	20.5	24.7	21.2	48.1	49.5	26.4	20	31.9	26.1
LMD-LDA (Jian et al., 2016)	22.6	24.8	22.3	46	48.3	26.4	20.1	31.4	27.7
QR (Cohan et al., 2015)	22.2	29.4	23.8	49.8	50.6	27.2	21.8	37.7	28.1
WE_{WIKI}	21.8	28.5	23.2	† 52.8	50	26.9	20.9	36.5	29.9
WE_{BIO}	23.9	† 31.2	† 25.5	† 57.1	51.9	† 29.2	† 23.1	† 46.2	† 34.1
WE_{BIO}+Rtrft	† 24.8	† **33.6**	† 26.4	† 58.3	52.4	† **30.7**	† 24.0	† 55.5	† 34.9
WE_{BIO}+Dmn	† **25.4**	† 33.0	† **27.0**	† **59.8**	† **53.0**	† 30.6	† **24.4**	† **56.1**	† **37.1**

† shows statistical significance (t-test, p<0.05) over the best baseline for the respective metric

- Human performance C-P@1: 56.7%, ours: 56.1%
- Our performance correlates with human performance

- **External evaluation**: How does the performance of citation-based summarization change if we contextualize citations?

Summarization method ➡	KLSUM		LexRank		LSA		SumBasic	
Contextualization method ⬇	RG1	RG2	RG1	RG1	RG2	RG1	RG1	RG2
No Context	36.0	8.3	41.3	10.8	34.7	6.5	38.7	8.7
BM25 (Jones et al., 2000)	35.5	8.0	39.8	9.9	33.5	6.2	39.5	9.4
DESM (Mitra et al., 2016)	36.3	8.7	40.2	10.4	32.6	6.5	38.3	7.9
VSM (Cohan et al., 2015)	35.3	7.9	40.0	9.9	33.5	6.2	39.5	9.4
LMD-LDA (Jian et al., 2016)	38.4	9.1	43.1	11.0	37.8	7.6	40.1	8.9
QR (Cohan et al., 2015)	39.9	10.2	43.8	11.7	38.9	8.0	40.1	8.6
WE_{WIKI}	39.7	10.2	42.7	11.8	38.0	8.0	40.2	9.2
WE_{BIO}	† 41.7	† 11.7	† 45.6	† **13.8**	† 40.3	† 9.1	† 42.4	† **12.6**
WE_{BIO}+Rtrft	† 42.9	† 12.2	† 46.2	11.6	† 40.0	8.9	† 41.3	9.7
WE_{BIO}+Dmn	† **44.0**	† **13.4**	† **47.3**	† 13.6	† **42.3**	† **10.4**	† **44.0**	† 11.7

† shows statistical significance (t-test, p<0.05) over the best baseline for the respective metric

- Summary:
    - Contextualization improves the quality of the citation texts (*up to 4.1 points in Rouge-1 and 3.2 points in Rouge-2 scores*)
    - Embeddings and domain knowledge provide improved semantic matching (*up to +12% improvement in character offset overlap F1 scores*)
    - Contextualization helps citation-based summarization