Contextualized PACRR for Complex Answer Retrieval

Sean MacAvaney[1*], Andrew Yates[2], Kai Hui[2]

[1] IRLab, Georgetown University [2] Max Planck Institute for Informatics

* Work completed during internship at MPII

Challenges

Discourse Context: Difficulty due to paragraph independence & omission of key words/phrases that would otherwise be clear from document context

Missing "green"

Green sea turtle » Ecology and behavior » Breathing and sleep

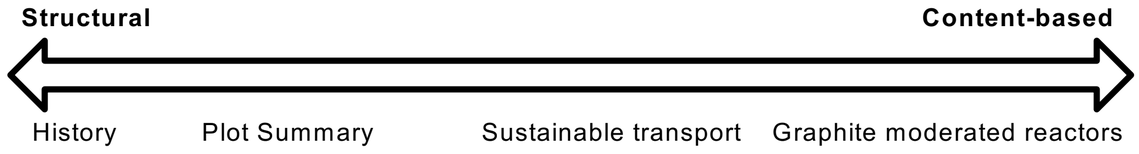
Sea turtles spend almost all their lives submerged, but must breathe air for the oxygen needed to meet the demands of vigorous activity. With a single explosive…

Very few matching terms

The Lord of the Rings » Plot Summary » Prologue

Thousands of years before the events of the novel, the Dark Lord Sauron had forged the One Ring to rule the other Rings of Power and corrupt those who wore them: the leaders of Men, Elves and Dwarves. Sauron was…

Heading Utility: Some headings provide common document structure, while others are specific to the content itself



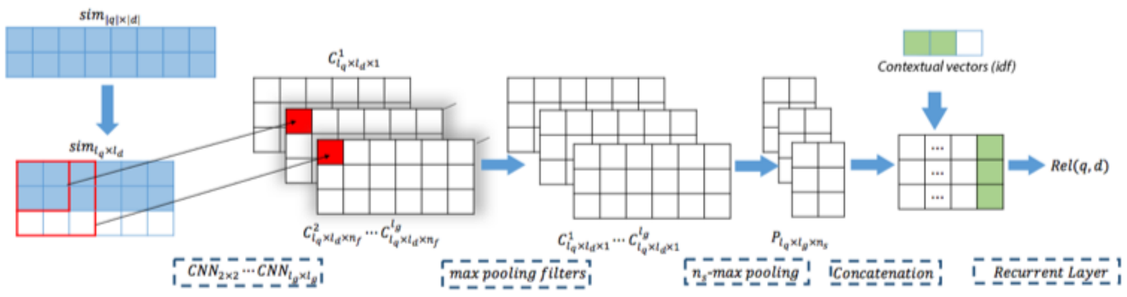
Sustainable biofuel » Sustainable transport

Biofuels have a limited ability to replace fossil fuels and should not be regarded as a 'silver bullet' to deal with transport emissions. Biofuels on their own cannot deliver a sustainable transport system and so must be…

Contextualized PACRR Model

PACRR: "Position-Aware Convolutional Recurrent Relevance" ranking model

- Add additional contextual vectors to combination



Heading position: Indicates whether query term came from heading, intermediate, or main heading

	Green	sea	turtle	Ecology	and	behavior	Life	cycle
title	1	1	1	0	0	0	0	0
inter.	0	0	0	1	1	1	0	0
main	0	0	0	0	0	0	1	1

Heading usage frequency: Estimate heading utility by calculating the probability of a given heading (h) occurring in the corpus (C)

$$frq(h) = \frac{\sum_{a \in C} I(h \in a)}{|C|}$$

- Stratified by percentile: 60th (1), 90th (2), 99th (3)

Green	sea	turtle	Ecology	and	behavior	Life	cycle
0	0	0	3	3	3	3	3

Term occurrence: Calculate the probability that a given term occurs verbatim in relevant paragraphs

$$occ(t) = \frac{\sum_{h \in C} \sum_{p \in rel(h)} [I(t \in h \wedge t \in p)]}{\sum_{h \in C} I(I \in h)}$$

- Distinguishes between headings like "History" and "Graphite moderated reactors"
- Extension (exp): Perform query expansion and allow added terms to also act as matches

	Green	sea	turtle	Ecology	and	behavior	Life	cycle
	0.6	0.5	0.6	0.1	0.8	0.2	0.1	0.3
(exp)	0.6	0.5	0.7	0.1	0.8	0.3	0.9	0.3

Results

Run	MAP	R-Prec	MRR
Automatic			
nn6-pos	0.144	0.111	0.216
nn4_pos_hperc	**0.148**	**0.116**	**0.224**
nn6_pos_tprob	0.140	0.107	0.213
Manual			
nn6_pos	0.197	0.209	0.417
nn4_pos_hperc	**0.201**	**0.213**	0.418
nn6_pos_tprob	0.198	0.206	**0.419**
Manual (lenient)			
nn6_pos	0.235	0.311	0.499
nn4_pos_hperc	0.234	0.311	0.505
nn6_pos_tprob	**0.241**	**0.321**	**0.520**

No statistically significant difference between each pair of runs (within environment)

Heading usage frequency generally performs best for automatic and manual

Term occurrence performs the best for lenient evaluation

Run	# Neg. train.	Heading position	Heading usage frq.	Term occ. (exp)
nn6-pos	6	✓		
nn4_pos_hperc	4	✓	✓	✓
nn6_pos_tprob	6	✓		✓