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

* Joint work with IRLab and MPII

Contribution: Contextualized PACRR

- We start with the PACRR neural ranking architecture (Hui et al. 2017)
- Include task-specific contextual vectors

Kai Hui, Andrew Yates, Klaus Berberich and Gerard de Melo. “PACRR: A Position-Aware Neural IR Model for Relevance Matching.” *EMNLP* (2017).

PACRR overview

1. Retrieve candidate documents to re-rank using basic unsupervised ranking approach (e.g. BM25)

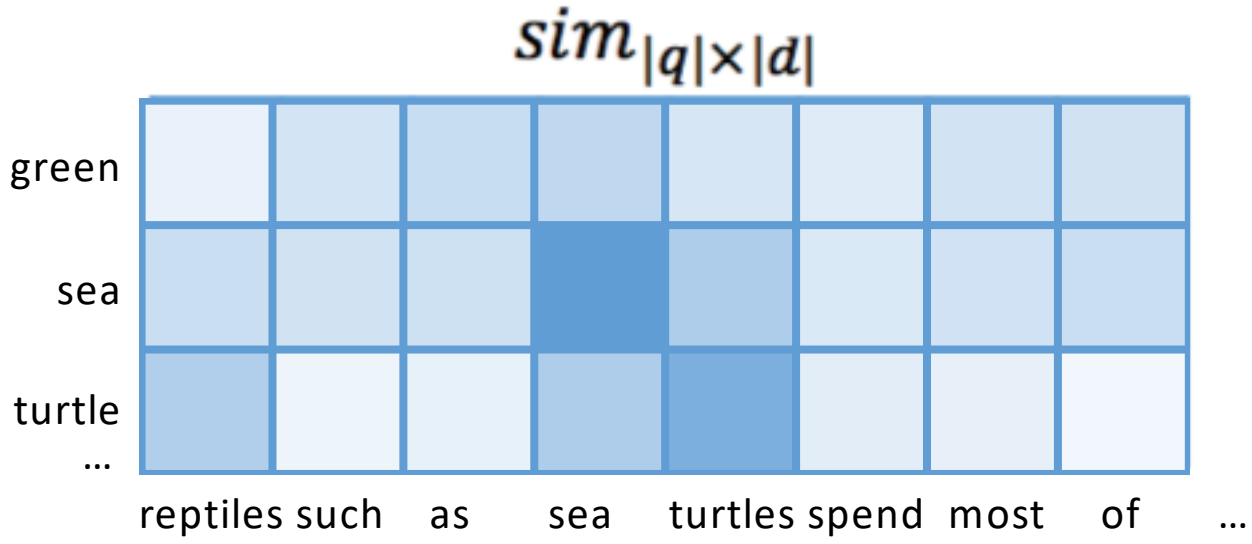
Query: green sea turtle ecology and behavior life cycle

Results:

1. As one of the first sea turtle species studied much of what is known of sea turtle ecology…
2. Reptiles such as sea turtles spend most of their lives in the ocean. However their life cycle…
3. Jesús A. Rivas (born in Caracas Venezuela) is a Venezuelan herpetologist tropical ecologist…
4. As of 2015 Mote employs over 200 staff members conducting research on 25 different…
5. Galápagos green turtles lifestyle is similar to other populations of Chelonians. The behavior…
6. Within the sea turtles E. imbricata has several unique anatomical and ecological traits. It is…

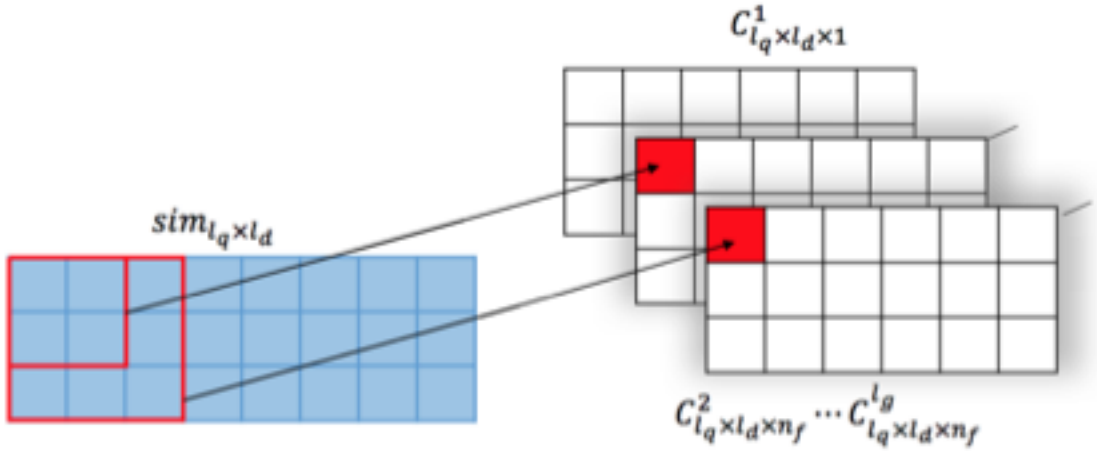
PACRR overview

2. Build similarity matrix for each document



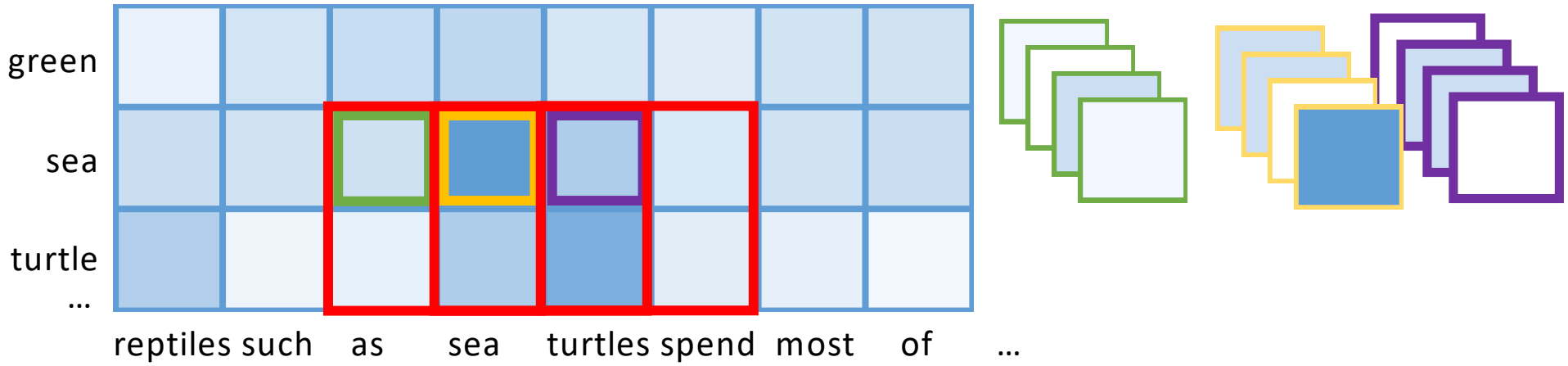
PACRR overview

3. Convolution over similarity matrix to find matching patterns (e.g. ngrams)



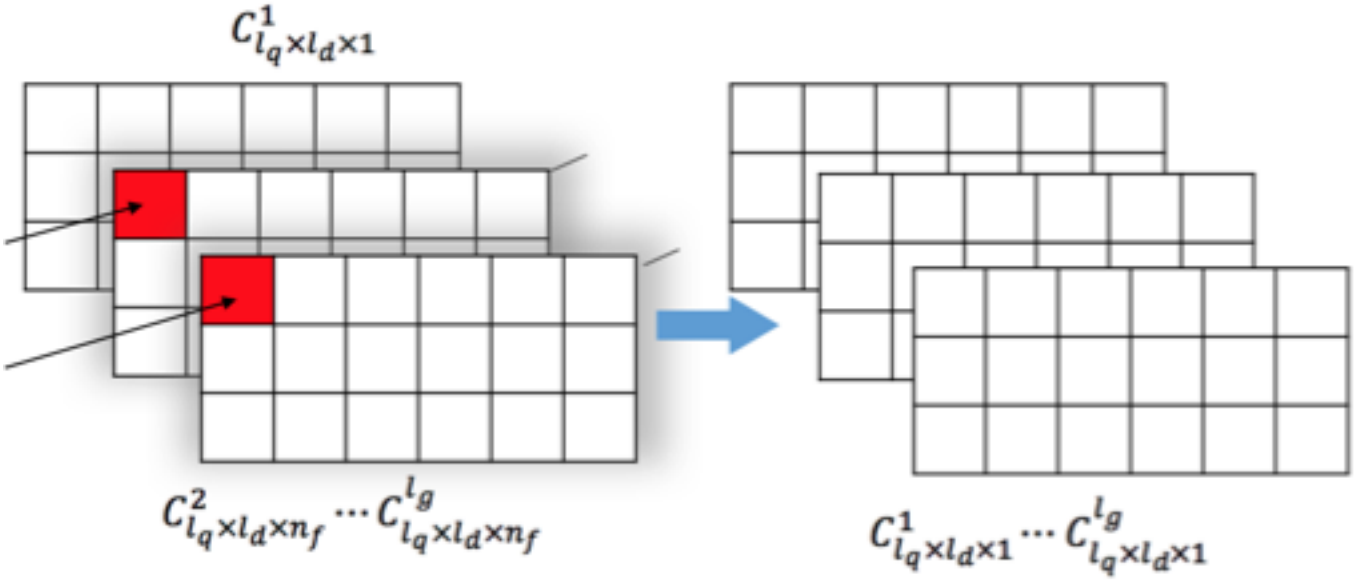
PACRR overview

3. Convolution over similarity matrix to find matching patterns (e.g. ngrams)



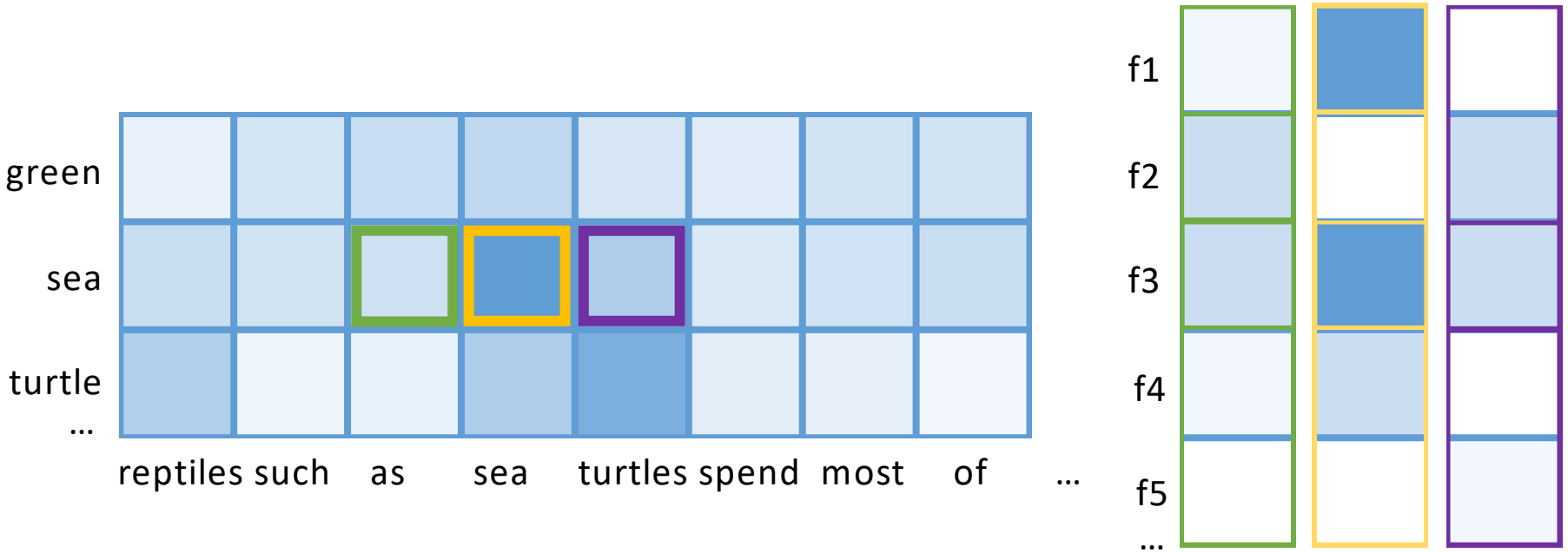
PACRR overview

4. Perform max pooling over filters



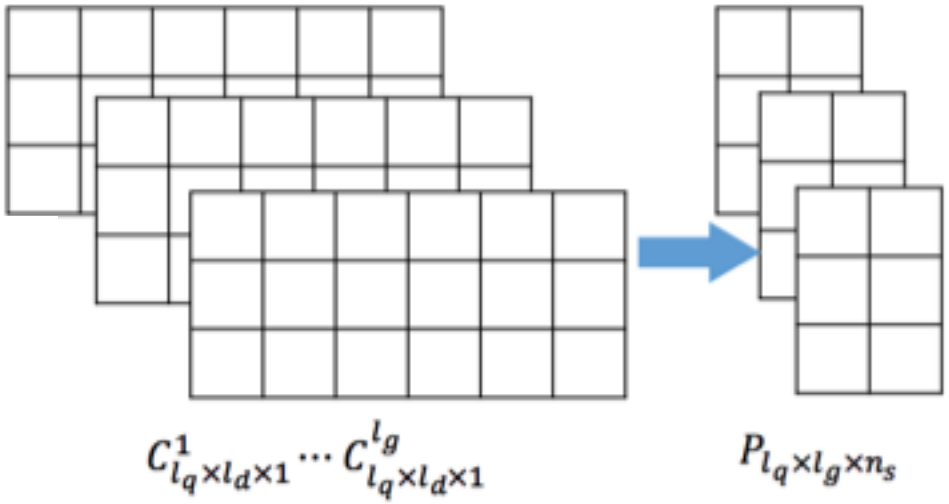
PACRR overview

4. Perform max pooling over filters



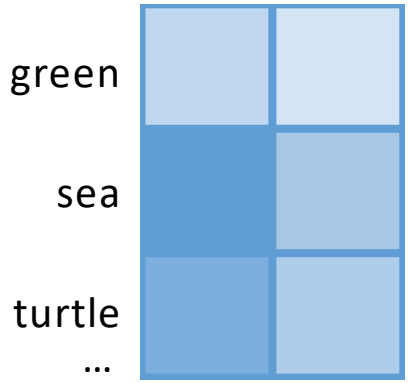
PACRR overview

5. Perform k-max pooling over query terms



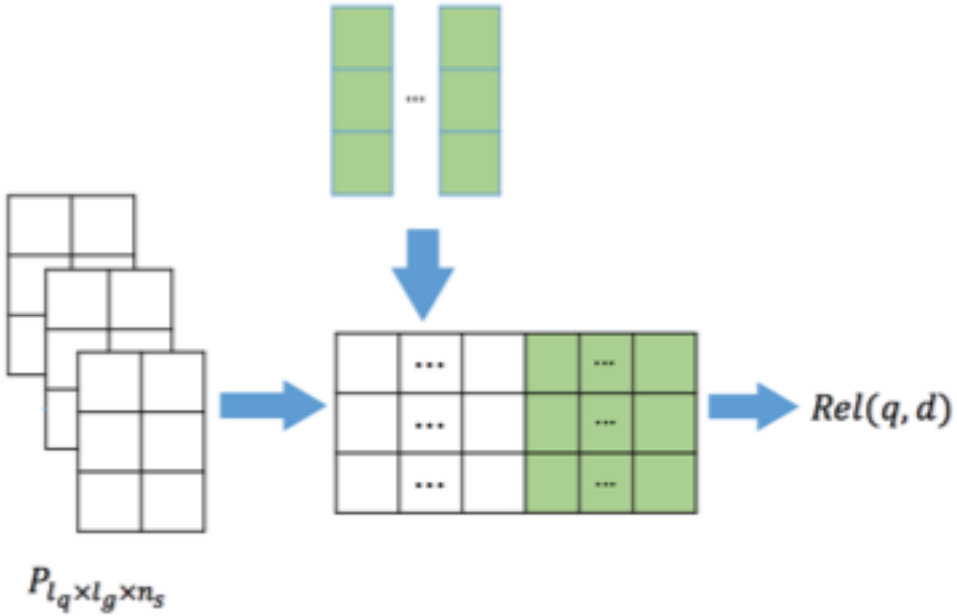
PACRR overview

5. Perform k-max pooling over query terms



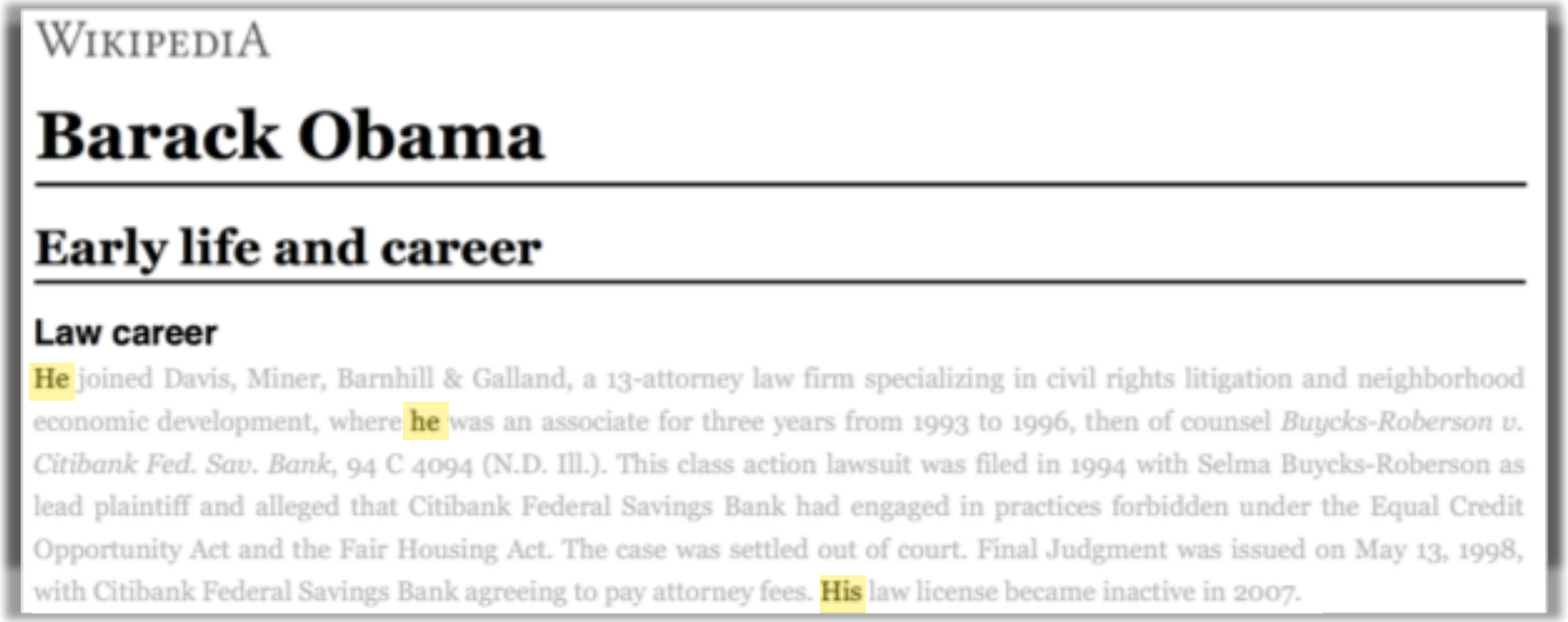
PACRR overview

6. Concatenation with query term contextual vectors
7. Combination (dense layers) to produce relevance score



Challenge: Discourse context

Well-written articles will often avoid overusing terms within a document because the context is clear.



WIKIPEDIA

Barack Obama

Early life and career

Law career

He joined Davis, Miner, Barnhill & Galland, a 13-attorney law firm specializing in civil rights litigation and neighborhood economic development, where he was an associate for three years from 1993 to 1996, then of counsel *Buycks-Roberson v. Citibank Fed. Sav. Bank*, 94 C 4094 (N.D. Ill.). This class action lawsuit was filed in 1994 with Selma Buycks-Roberson as lead plaintiff and alleged that Citibank Federal Savings Bank had engaged in practices forbidden under the Equal Credit Opportunity Act and the Fair Housing Act. The case was settled out of court. Final Judgment was issued on May 13, 1998, with Citibank Federal Savings Bank agreeing to pay attorney fees. His law license became inactive in 2007.

Challenge: Query term utility

WIKIPEDIA

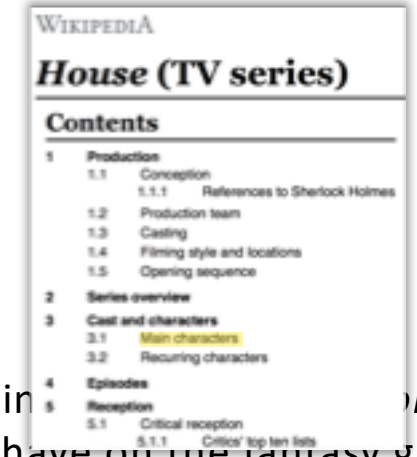
The Lord of the Rings

Contents

Some **query components** exhibit a specific structure

Other **query components** exhibit a specific to the particular topic

Who are the main characters in *The Lord of the Rings*?


WIKIPEDIA

***House* (TV series)**

Contents

1 Production
- 1.1 Conception
  - 1.1.1 References to Sherlock Holmes
- 1.2 Production team
- 1.3 Casting
- 1.4 Filming style and locations
- 1.5 Opening sequence

2 Series overview

3 Cast and characters
- 3.1 Main characters
- 3.2 Recurring characters

4 Episodes

5 Reception
- 5.1 Critical reception
  - 5.1.1 Critics' top ten lists

What in[illegible]*ord of the Rings* have on the fantasy genre?

Who are the main characters in House?

Heading position vectors

- Indicator of whether query term came from heading, intermediate, or main heading


WIKIPEDIA

Green sea turtle

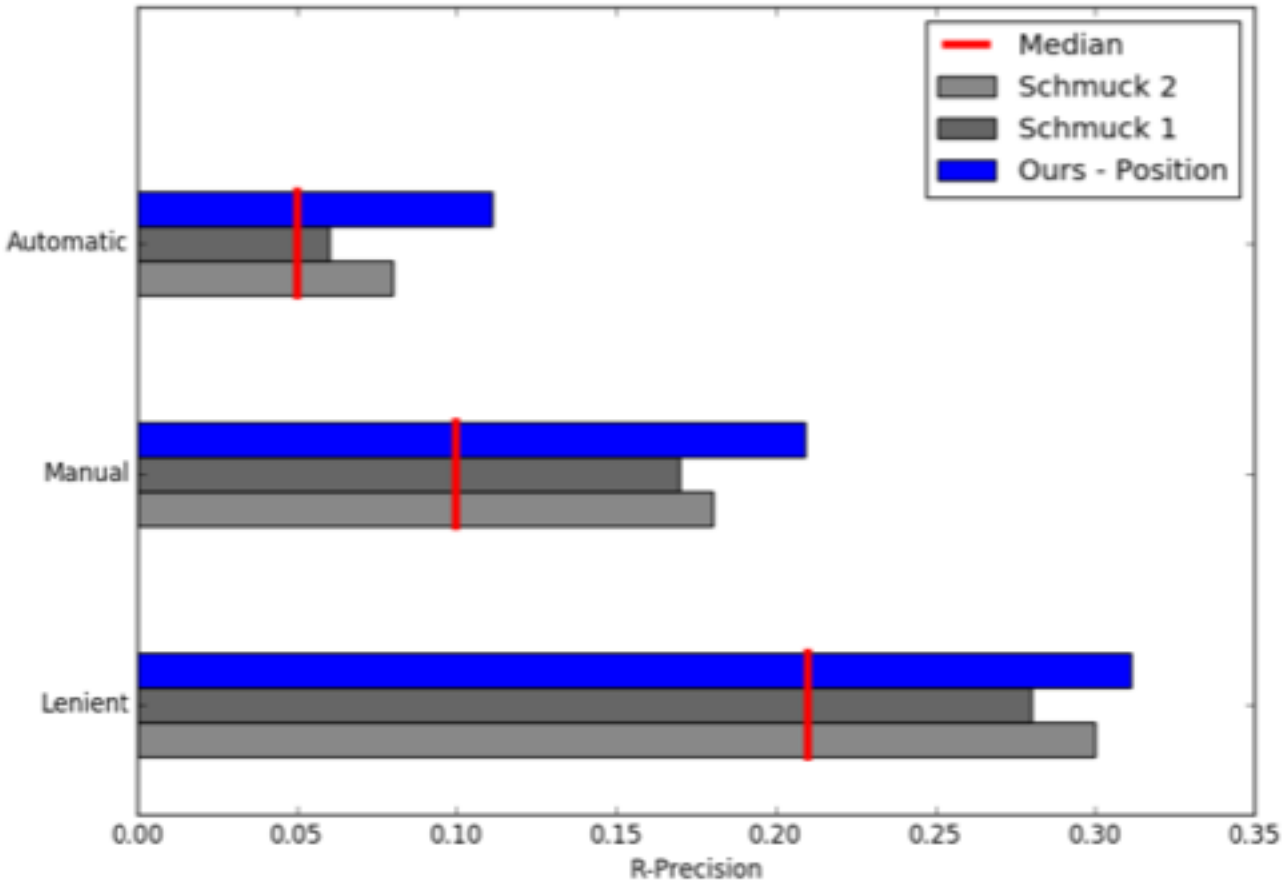
6 Ecology and behavior

6.3 Life cycle

	green	sea	turtle	ecology	and	behavior	life	cycle
title	1			0			0	
intermediate	0			1			0	
main	0			0			1	

- Intuition: **Title is important to keep intact**

Our approach performs well



Heading usage frequency

Content; does not occur frequently in training data

Structural; occurs frequently in training data

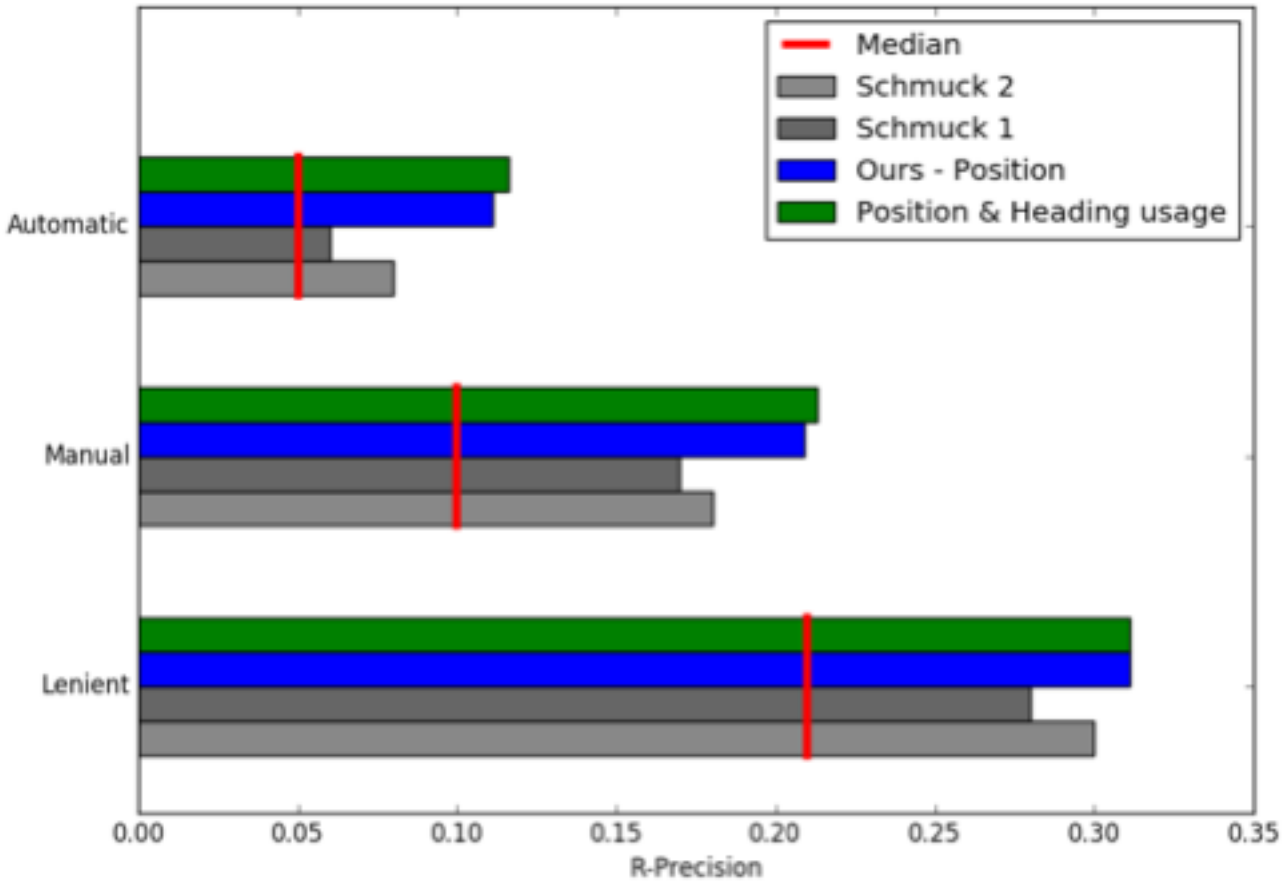
Heading usage frequency vector

- Calculate how frequently each heading occurs in training data
- Stratify by percentile (60^{th}=1, 90^{th}=2; 99^{th}=3)

green	sea	turtle	ecology	and	behavior	life	cycle
0	0	0	3	3	3	3	3

- Intuition: **Common headings are less likely to match verbatim**

Heading usage frequency performs well, too



Term occurrence vector

- For each term in each heading, calculate the frequency that it appears in relevant paragraphs

green	sea	turtle	ecology	and	behavior	life	cycle
0.6	0.5	0.6	0.1	0.8	0.2	0.1	0.3

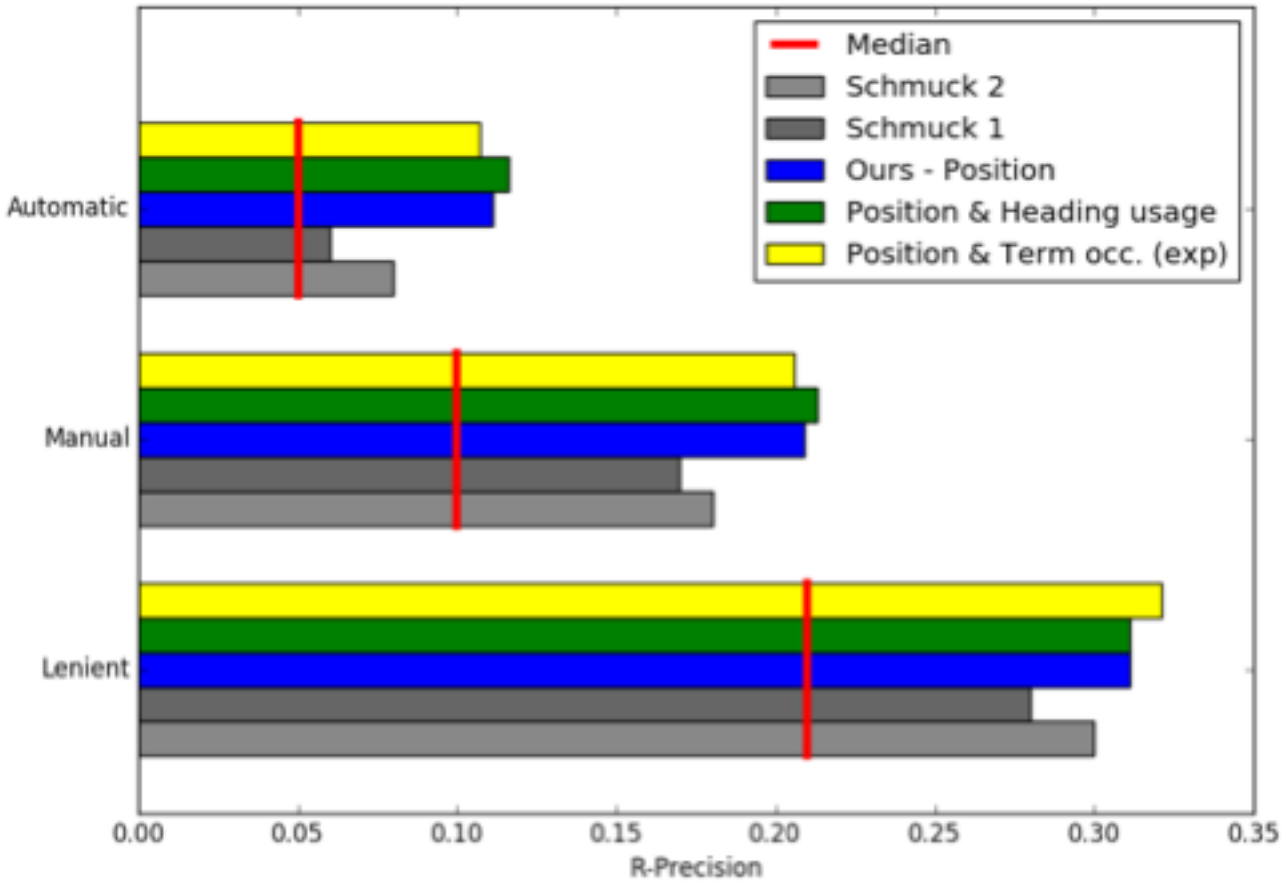
- Intuition: **Inform the model directly about which terms are likely to match**

Term occurrence vector - expansion

- Since the PACRR model uses word similarity scores (not exact term matches), find words that are similar and include these as matches as well
- Perform relevance feedback on most common headings (e.g. history, life cycle, etc.)

	green	sea	turtle	ecology	and	behavior	life	cycle
Without expansion	0.6	0.5	0.6	0.1	0.8	0.2	0.1	0.3
With expansion	0.6	0.5	0.7	0.1	0.8	0.3	0.9	0.3

Expanded term occurrence is better for lenient



Failure case – focus not antioxidants

Antioxidant » Health effects » Relation to diet

As with the minerals discussed above some vitamins are recognized as organic essential nutrients necessary in the diet for good health… Moreover thousands of different phytochemicals have recently been discovered in food (particularly in fresh vegetables) which may have desirable properties including antioxidant activity…

	Position	Pos. + Heading Usage	Pos. + Term Occ.
Rank	1	2	9

Failure case – No mention of diet

Query: Antioxidant » Health effects » Relation to diet

According to a March 2015 Scientific American Special Report on Aging article laboratory mice at a University of Washington laboratory who produced more catalase which is an antioxidant lived longer. Research on the topic both supports and cautions against the benefit of antioxidants for health effects on aging.

	Position	Pos. + Heading Usage	Pos. + Term Occ.
Rank	3	5	10

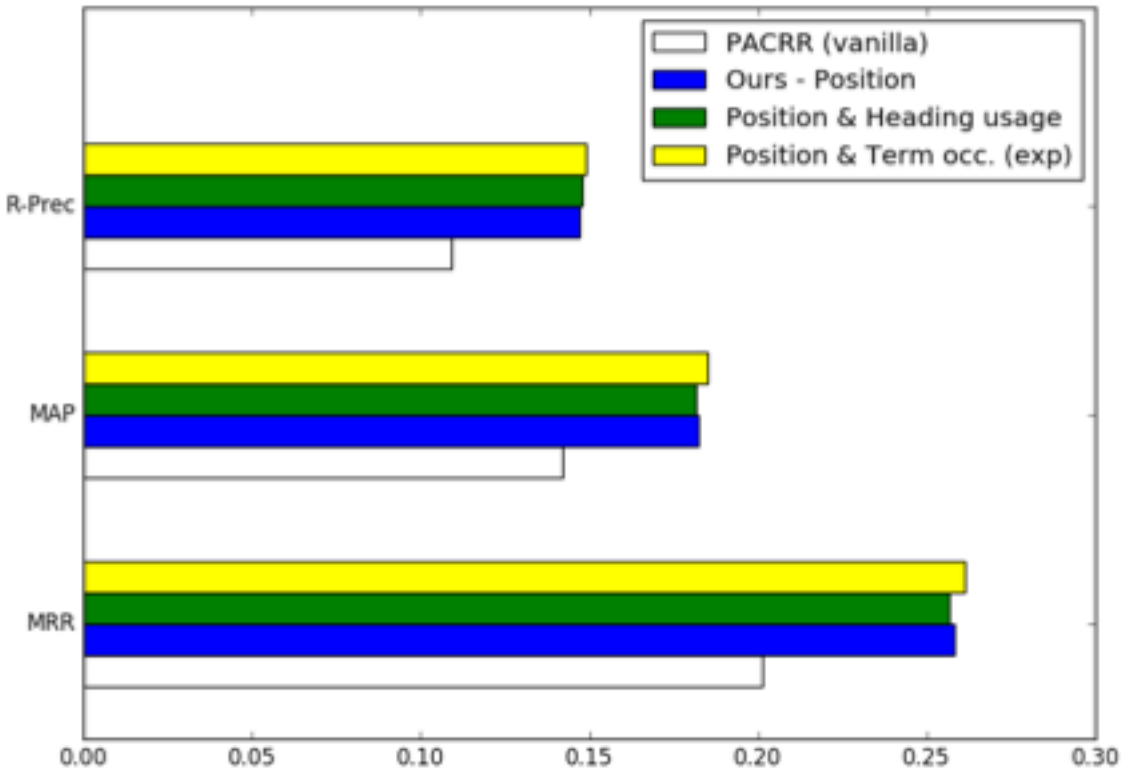
Positive case – Patterns look similar

Antioxidant » Health effects » Relation to diet

…As a result of animal experimental studies antioxidant and anti-inflammation are expected to be effective countermeasures for CNS risks from space radiation. Diets of blueberries and strawberries were shown to reduce CNS risks after heavy-ion exposure. Estimating the effects of diet and nutritional supplementation will be a primary goal of CNS research on countermeasures…

	Position	Pos. + Heading Usage	Pos. + Term Occ.
Rank	19	1	3

All three configurations outperform “vanilla” PACRR



(Vanilla PACRR was not an official run. These values are based on training automatic judgments.)

Summary

- Adding contextual vectors to the PACRR architecture performs well on CAR
- Simply informing the model of the information hierarchy does well. Adding more specialized contextual vectors doesn't help much.

Special thanks to the members of the IRLab for constructive comments.