Ranking Documents by Answer-Passage Quality

Evi Yulianti, Ruey-Cheng Chen[†], Falk Scholer, W. Bruce Croft, Mark Sanderson

RMIT University (Melbourne, Australia)

[†] SEEK Ltd. (Melbourne, Australia)

Introduction

Passage-Level Evidence for Ad Hoc Retrieval

- Combining document-level and passage-level evidence has been considered an effective retrieval approach[1].
- Combining evidence from the best-matching passage in retrieved documents leads to increased retrieval effectiveness[2].
- “Evidence” is however limited to the query.

[1] Callan, 1994; Wilkinson, 1994.
[2] Bendersky and Kurland, 2008.

Introduction

Passage-Level Evidence for Ad Hoc Retrieval

- Combining document-level and passage-level evidence has been considered an effective retrieval approach[1].
- Combining evidence from the best-matching passage in retrieved documents leads to increased retrieval effectiveness[2].
- "Evidence" is however limited to the query.

Further Question

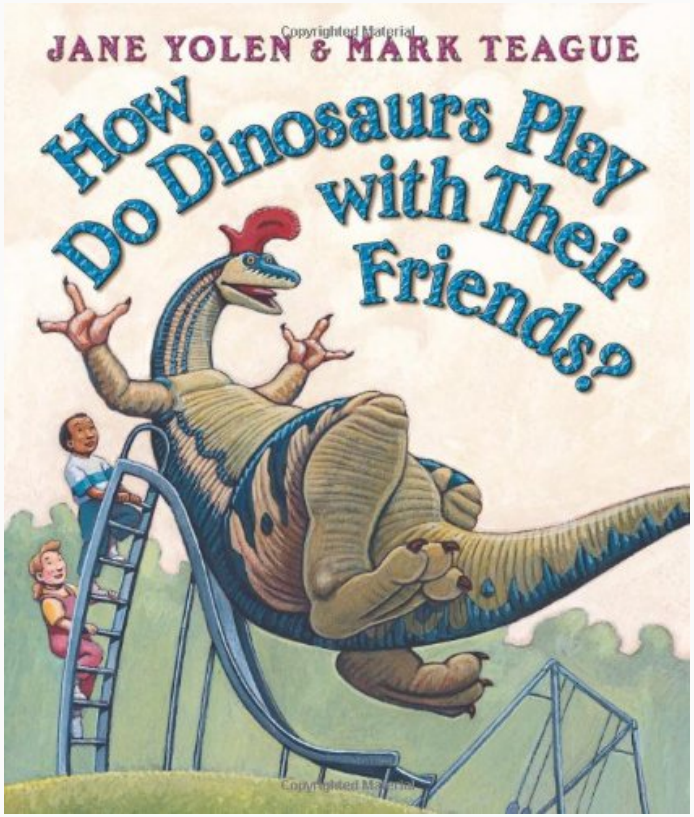
*Will selecting passages that are **more likely to report an answer** lead to further improvements?*

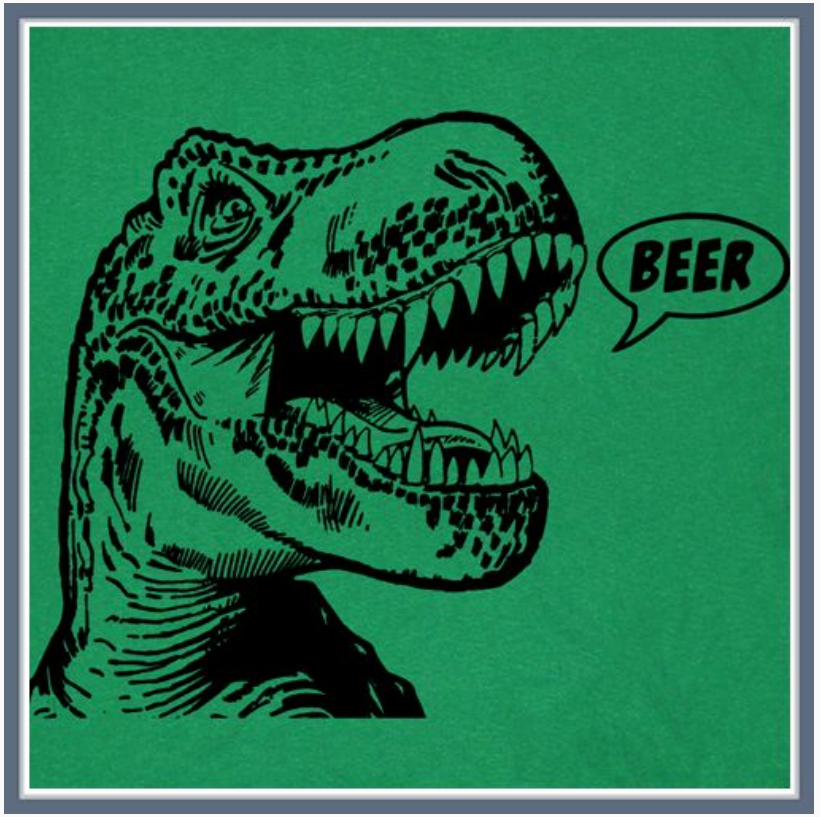
---

[1] Callan, 1994; Wilkinson, 1994.

[2] Bendersky and Kurland, 2008.

“Dinosaurs”





https://www.amazon.com/gp/product/043985654X/

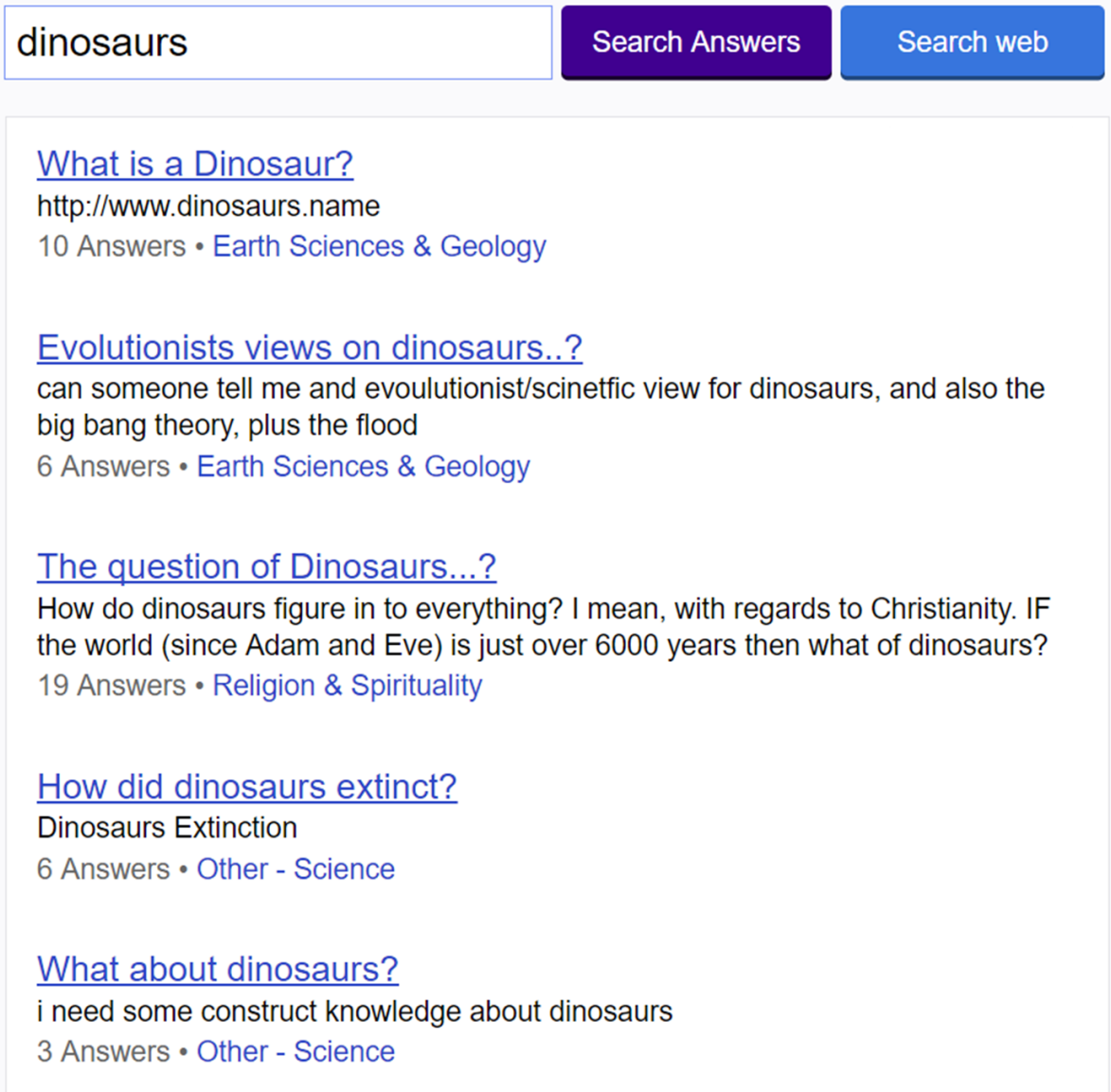
https://www.pinterest.com/pin/700309810777473872

“Dinosaurs” (TREC Web Topic 14)

“I want to find information about and pictures of dinosaurs.”

- Go to the Discovery Channel’s dinosaur site, which has pictures of dinosaurs and games.
- I’m looking for free pictures of dinosaurs.
- I want to find pictures of dinosaurs that I can color in, as in a coloring book.
- I’m looking for a list of all (or many of) the different kinds of dinosaurs, with pictures.
- Take me to the homepage for the BBC series, ”Walking with Dinosaurs”.

“Dinosaurs” (Yahoo! Answers)



The Answer-Bearingness Hypothesis

Documents that are likely to bear focused answers to the posed query should be ranked highly.

- Passages are scored by querying an oracle “answer source”
- Ranking takes best-scoring passage quality into account.

The Answer-Bearingness Hypothesis

Documents that are likely to bear focused answers to the posed query should be ranked highly.

- Passages are scored by querying an oracle "answer source"
- Ranking takes best-scoring passage quality into account.

The Answer-Bearingness Hypothesis

Documents that are likely to bear focused answers to the posed query should be ranked highly.

- Passages are scored by querying an oracle "answer source"
- Ranking takes best-scoring passage quality into account.

Key Ingredients

$$\textit{Answer} \Longrightarrow \textit{Passage} \Longrightarrow \textit{Quality}$$

Ad Hoc Retrieval Methodologies: A Breakdown

	Document	Passage
Local Collection	BM25, SDM, and DfR Pseudo relevance feedback Quality-biased ranking[3]	Passage-based LM[4]
External Resources	External expansion Weighted dependence model[5]	Answer-passage quality

[3] Rocchio, 1971; Lavrenko and Croft, 2001; Bendersky et al., 2011.
[4] Bendersky and Kurland, 2008; Krikon and Kurland, 2011; He et al., 2012.
[5] Diaz and Metzler, 2006; Weerkamp et al., 2012; Bendersky et al., 2010.

Methodology

Framework

Given query Q:

- **Retrieve documents** $\mathcal{D}_Q$.
- **Retrieve answers** $\mathcal{A}_Q$ from answer sources
- **Induce answer-reporting passages**
  1. A probabilistic framework for passage extraction
  2. Open-domain question answering
- **Re-rank documents** using passage quality.

Approach 1: A Probabilistic Framework

1. Use answers $\mathcal{A}_Q$ to improve relevance estimation of terms[6]:

$$p(t|Q) \propto \sum_{A \in \mathcal{A}_Q} \overbrace{p(t|A)}^{\text{term relevance}} \overbrace{p(Q|A)}^{\text{answer relevance}} \tag{1}$$

   - Estimating $p(t|A)$: **QL**, **BM25**, **Embedding LM (EMB)**
   - Estimating $p(Q|A)$: distributional assumptions e.g. **DCG** or **RBP**

2. Extract G that best approximates the answer-bearing content.
   - **Fixed-length passages (PSG)**[7].
   - **Integer-linear programming (ILP)**[8].

[6] Diaz and Metzler, 2006; Lavrenko and Croft, 2001.
[7] O'Connor, 1980; Callan, 1994.
[8] Takamura and Okumura, 2009; Gillick and Favre, 2009; Woodsend and Lapata, 2012.

Approach 2: Open-Domain Question Answering

Extract text fragments ("answers") directly from documents to address users' questions.

Document Reader (DR)[9]. The DR model takes query Q and document D as input and returns a best-matching passage $G^* = \langle g_1, g_2, \ldots, g_m \rangle$:

$$G^*_{\text{DR}} = \arg\max_{G \in D} \max_{1 \leq i \leq j \leq m} \log p_S(g_i|G, Q) + \log p_E(g_j|G, Q). \tag{2}$$

The score indicates the log-likelihood of G reporting an answer.

[9] Chen et al., 2017.

Passage Quality Based Ranking

All ranking signals combined by using a feature-based model:

$$\lambda_D f_{\text{SDM}}(q, D) + \sum_j \lambda_j f_j(q, G) \qquad \text{where } \lambda_D + \sum_j \lambda_j = 1. \tag{3}$$

Feature	Definition
PassageScore	Objective value to score the passage
PassageOverlap	Bigram overlap with respect to answers
NumSentences	Number of sentences
QueryOverlap	Number of query term occurrences
AvgWordWeight	Average passage term weight
AvgTermLen	Average passage term length
Entropy	Shannon entropy of the term distribution
FracStops	Fraction of passage terms that are stopwords
StopCover	Fraction of stopwords appear in the passage

Experiments

Setup

Test Collections

GOV2	TREC Topics 701–850	25,205,179 docs
ClueWeb09B	TREC Web Topics 1–200	50,066,642 docs

- Top 100 docs retrieved using SDM (Indri)
- Sentences extracted, stemmed and stopword removed.

External Resources

- Submit queries to Yahoo! Answers search engine
- Take the best answer for each of the top ten matching questions (they appear relevant but **do not address the queries**)
- Hold out 3 GOV2 topics and 5 CW09B topics.

Setup

Word Embeddings

- `EmbWiki`: 1M vectors, 300d (English Wikipedia, 16B tokens)
- `EmbYA`: 5M vectors, 100d (Y!A crawl 2013–2016, 5B tokens)

Baselines

- Sequential Dependence Model (`SDM`);
- Passage-Based Language Model (`MSP` and `SUM`);
- Quality-Biased Ranking (`QSDM`);
- External Expansion (`EE`).

Parameter Estimation

`MSP/SUM`: grid search + cross validation

`EE`: random search on randomly re-sampled 50%-50% split. In our experiments, $(n_T, \lambda_C, \lambda_Q)$ were set to $(60, 0.3, 0.2)$ on GOV2 and to $(50, 0.2, 0.2)$ on CW09B.

`PSG/ILP`: Passage size $K = 50$ (words) and $\lambda = 0.1$.

`QL/BM25`: $\mu = 100$; $b_1 = 1.2$ and $k_1 = 0.75$

`EmbWiki/EmbYA`: $\kappa = 10$ and $x_0 = 0$ (cross validation)

`DR`: query/passage vectors encoded using 128 hidden units in 3-layer LSTM network, model trained on SQuAD using AdaMax, dropout rate set to 0.5.[10]

All trained on 10-fold CV, using Coordinate Ascent (NDCG@20)

[10] Chen et al., 2017; Rajpurkar et al., 2016; Kingma and Ba, 2014.

Comparisons

	GOV2 (NDCG@20)	CW09B (NDCG@20)
SDM$^{(\dagger)}$	0.4751	0.2462
QSDM$^{(\diamond)}$	0.5022‡	0.2639†
SDM+EE$^{(*)}$	0.5057‡	0.2736‡
SDM+MSP	0.4745	0.2469
SDM+SUM	0.4749	0.2409
SDM+PSG (EmbWiki)	0.4975‡	0.2588†
SDM+PSG (EmbYA)	0.4957†	0.2644†
SDM+PSG (QL)	0.5068‡	0.2569
SDM+PSG (BM25)	0.5116‡	0.2687‡
SDM+ILP (EmbWiki)	0.4967‡	0.2652†
SDM+ILP (EmbYA)	0.4951‡	0.2665†
SDM+ILP (QL)	0.5052‡	0.2901$^{\ddagger\diamond\diamond}$
SDM+ILP (BM25)	**0.5171**‡	**0.2955**$^{\ddagger\diamond\diamond *}$
SDM+DR (Title)	0.4786	0.2505
SDM+DR (Desc)	0.4894†	0.2681‡

Comparisons

	GOV2 (NDCG@20)	CW09B (NDCG@20)
SDM$^{(\dagger)}$	0.4751	0.2462
QSDM$^{(\diamond)}$	0.5022‡	0.2639†
SDM+EE$^{(*)}$	0.5057‡	0.2736‡
SDM+MSP	0.4745	0.2469
SDM+SUM	0.4749	0.2409
SDM+PSG (EmbWiki)	0.4975‡	0.2588†
SDM+PSG (EmbYA)	0.4957†	0.2644†
SDM+PSG (QL)	0.5068‡	0.2569
SDM+PSG (BM25)	0.5116‡	0.2687‡
SDM+ILP (EmbWiki)	0.4967‡	0.2652†
SDM+ILP (EmbYA)	0.4951‡	0.2665†
SDM+ILP (QL)	0.5052‡	0.2901$^{\ddagger\diamond\diamond}$
SDM+ILP (BM25)	**0.5171**‡	**0.2955**$^{\ddagger\diamond\diamond *}$
SDM+DR (Title)	0.4786	0.2505
SDM+DR (Desc)	0.4894†	0.2681‡

Comparisons

	GOV2 (NDCG@20)	CW09B (NDCG@20)
SDM$^{(\dagger)}$	0.4751	0.2462
QSDM$^{(\diamond)}$	0.5022‡	0.2639†
SDM+EE$^{(*)}$	0.5057‡	0.2736‡
SDM+MSP	0.4745	0.2469
SDM+SUM	0.4749	0.2409
SDM+PSG (EmbWiki)	0.4975‡	0.2588†
SDM+PSG (EmbYA)	0.4957†	0.2644†
SDM+PSG (QL)	0.5068‡	0.2569
SDM+PSG (BM25)	0.5116‡	0.2687‡
SDM+ILP (EmbWiki)	0.4967‡	0.2652†
SDM+ILP (EmbYA)	0.4951‡	0.2665†
SDM+ILP (QL)	0.5052‡	0.2901$^{\ddagger\diamond\diamond}$
SDM+ILP (BM25)	**0.5171**‡	**0.2955**$^{\ddagger\diamond\diamond *}$
SDM+DR (Title)	0.4786	0.2505
SDM+DR (Desc)	0.4894†	0.2681‡

Comparisons

	GOV2 (NDCG@20)	CW09B (NDCG@20)
SDM$^{(\dagger)}$	0.4751	0.2462
QSDM$^{(\diamond)}$	0.5022‡	0.2639†
SDM+EE$^{(*)}$	0.5057‡	0.2736‡
SDM+MSP	0.4745	0.2469
SDM+SUM	0.4749	0.2409
SDM+PSG (EmbWiki)	0.4975‡	0.2588†
SDM+PSG (EmbYA)	0.4957†	0.2644†
SDM+PSG (QL)	0.5068‡	0.2569
SDM+PSG (BM25)	0.5116‡	0.2687‡
SDM+ILP (EmbWiki)	0.4967‡	0.2652†
SDM+ILP (EmbYA)	0.4951‡	0.2665†
SDM+ILP (QL)	0.5052‡	0.2901$^{\ddagger\diamond\diamond}$
SDM+ILP (BM25)	**0.5171**‡	**0.2955**$^{\ddagger\diamond\diamond *}$
SDM+DR (Title)	0.4786	0.2505
SDM+DR (Desc)	0.4894†	0.2681‡

Overall Effectiveness

	GOV2 (NDCG@20)	CW09B (NDCG@20)
SDM$^{(\dagger)}$	0.4751	0.2462
SDM+ILP (EmbYA)	0.4951$^{\ddagger **}$	0.2665$^{\dagger *}$
SDM+ILP (BM25)	0.5171‡	0.2955$^{\ddagger \diamond\diamond}$
QSDM$^{(\diamond)}$	0.5022‡	0.2639†
QSDM+ILP (EmbYA)	0.5126‡	0.2691†
QSDM+ILP (BM25)	0.5245$^{\ddagger \diamond\diamond}$	0.2959$^{\ddagger \diamond\diamond}$
QSDM+EE$^{(*)}$	0.5213$^{\ddagger \diamond\diamond}$	0.2819$^{\ddagger +}$
QSDM+EE+ILP (EmbYA)	0.5208$^{\ddagger \diamond}$	0.2864$^{\ddagger \diamond\diamond}$
QSDM+EE+ILP (BM25)	**0.5311**$^{\ddagger \diamond\diamond}$	**0.3015**$^{\ddagger \diamond\diamond **}$

Overall Effectiveness

	GOV2 (NDCG@20)	CW09B (NDCG@20)
SDM$^{(\dagger)}$	0.4751	0.2462
SDM+ILP (EmbYA)	$0.4951^{\ddagger **}$	$0.2665^{\dagger *}$
SDM+ILP (BM25)	$0.5171^{\ddagger}$	$0.2955^{\ddagger \diamond\diamond}$
QSDM$^{(\diamond)}$	$0.5022^{\ddagger}$	$0.2639^{\dagger}$
QSDM+ILP (EmbYA)	$0.5126^{\ddagger}$	$0.2691^{\dagger}$
QSDM+ILP (BM25)	$0.5245^{\ddagger \diamond\diamond}$	$0.2959^{\ddagger \diamond\diamond}$
QSDM+EE$^{(*)}$	$0.5213^{\ddagger \diamond\diamond}$	$0.2819^{\ddagger +}$
QSDM+EE+ILP (EmbYA)	$0.5208^{\ddagger \diamond}$	$0.2864^{\ddagger \diamond\diamond}$
QSDM+EE+ILP (BM25)	$\mathbf{0.5311}^{\ddagger \diamond\diamond}$	$\mathbf{0.3015}^{\ddagger \diamond\diamond **}$

Other Results

Combining `ILP` and `DR` significantly improves QSDM

	GOV2 (NDCG@20)	CW09B (NDCG@20)
`QSDM`$^{(\diamond)}$	0.5022	0.2639
`QSDM+Combined`	0.5280$^{\diamond\diamond}$	0.2896$^{\diamond\diamond}$

Falling back to using offline CQA data (Yahoo! L6) still shows improvements.

	GOV2 (NDCG@20)	CW09B (NDCG@20)
`QSDM`$^{(\diamond)}$	0.5022	0.2639
`QSDM+ILP (BM25)`	0.5083	0.2804$^{\diamond\diamond}$

Feature Importance

GOV2		*CW09B*	
Feature	Diff.	Feature	Diff. [11]
SDM	0.0306	SDM	0.0223
FracStop	0.0101	StopCover	0.0092
AvgWordWeight	0.0076	PassageScore	0.0086
UrlDepth	0.0063	FracVisText	0.0077
QueryOverlap	0.0052	EE	0.0076
FracAnchorText	0.0049	StopCover[P]	0.0046
FracStop[P]	0.0048	AvgTermLen	0.0038
FracVisText	0.0045	NumTitleTerm	0.0035

[11] Quality features with a [P] are the passage version.

Feature Importance

GOV2		*CW09B*	
Feature	Diff.	Feature	Diff. [11]
SDM	0.0306	SDM	0.0223
FracStop	0.0101	StopCover	0.0092
AvgWordWeight	0.0076	PassageScore	0.0086
UrlDepth	0.0063	FracVisText	0.0077
QueryOverlap	0.0052	EE	0.0076
FracAnchorText	0.0049	StopCover[P]	0.0046
FracStop[P]	0.0048	AvgTermLen	0.0038
FracVisText	0.0045	NumTitleTerm	0.0035

[11] Quality features with a [P] are the passage version.

Examples

Answer passages extracted for TREC Web Topic 65, “ *Find information and resources on the Korean language*.” (query: korean language)

`ILP (EmbYA)`	For example, different endings are used based on whether the subjects and listeners are friends, parents, or honoured persons. in a similar way European languages borrow from Latin and Greek. Its use limited some cases and the aristocracy prefers Classical Chinese for its writing. “Mortal enemy” and “head of state” are homophones in the South. Learn to read, write and pronounce Korean
`ILP (BM25)`	Yanbian (People’s Republic of China) Given this, it is sometimes hard to tell which actual phonemes are present in a certain word. Unlike most of the European languages, Korean does not conjugate verbs using agreement with the subject, and nouns have no gender. The Korean language used in the North and the South exhibits differences in pronunciation, spelling, grammar and vocabulary.
`DR (Desc)`	Korean is similar to Altaic languages in that they both lack certain grammatical elements, including number, gender, articles, fusional morphology, voice, and relative pronouns (Kim Namkil). Korean especially bears some morphological resemblance to some languages of the Northern Turkic group, namely Sakha (Yakut).

Discussions

Conjectures

- High-quality answer sources $\Longrightarrow$ rich query model
- ILP representation $\Longrightarrow$ compressed, highly diverse content

As a result, **non-relevant doc generates poor quality passages**

Examples

dinosaur plates, dinosaur cups, dinosaur mural, dinosaur balloons, dinosaur candy, dinosaur napkins, dinosaur tableware. Dinosaur Party Pack, Dinosaur Party Tableware, 12 Guests Dinosaur Times, Dinosaur Table Cover Dinosaur Mural, Dinosaur Wall Mural Banner Dinosaur Masks, Dinosaur Party Masks, Triceratops Masks, 4 pcs Dinosaur Pinatas, T-rex Pinata, Prehistoric Birthday Pinata

Conclusions

1. Propose a **quality-biased ranking approach** that incorporates signals from answer-reporting passages.
2. Achieve **strong empirical results** to support our hypothesis, which expands on the theory of passage-level evidence.
3. Simpler methodologies win on overall efficacy; open-domain question answering is not shown useful yet. (Task mismatch?)

Future Work:

- Understanding human perceptions of passage quality
- Further exploration into combining other ranking signals
- End-to-end architecture (anyone?)

Thanks for Your Attention!

Questions?

https://github.com/rmit-ir/AnswerPassageQuality