TREC Complex Answer Retrieval Overview

Laura Dietz*, Manisha Verma, Filip Radlinski, Nick Craswell

Homepage: `http://trec-car.cs.unh.edu`
Google group mailinglist: `TREC-CAR`

1 Introduction

The SWIRL 2012 workshop on frontiers, challenges, and opportunities for information retrieval report [1] noted many important challenges. Among them, challenges such as conversational answer retrieval, sub-document retrieval, and answer aggregation share commonalities: We desire answers to complex needs, and wish to find them in a single and well-presented source. Advancing the state of the art in this area is the goal of this TREC track.

Consider a user investigating a new and unfamiliar topic. This user would often be best served with a single summary, rather than being required to synthesize his or her own summary from multiple sources. This is especially the case in mobile environments with restricted interaction capabilities. While these have led to extensive work on finding the best short answer, the target in this track is the retrieval of comprehensive answers that are composed of multiple text fragments from multiple sources. Retrieving high-quality longer answers is challenging as it is not sufficient to choose a lower rank-cutoff with the same techniques as for short answers. Instead, we need new approaches for finding relevant information in a complex answer space.

Many examples of manually created complex answers exist on the Web. Famous examples are articles from how-stuff-works.com, travel guides, or fanzines. These are collections of articles, that each constitutes a long answer to an information need represented by the title of the article.

The fundamental task of collecting references, facts, and opinions into a single coherent summary has traditionally been a manual process. We envision that automated information retrieval systems can relieve users from a large amount of manual work through sub-document retrieval, consolidation and organization. Ultimately, the goal is to retrieve synthesized *information* rather than documents.

2 Task Description

While the long-term goal of this track is to retrieve complex answers without any more information than the given complex topic, in the first year we focus on a simpler task, where both the topic and an appropriate outline is provided as a query. An example outline is given in Figure 1. We run two tasks: passage and entity.

Passage Task: Given an outline for complex topic outline Q, retrieve for each of its sections H_i, a ranking of relevant passages S.

Entity Task: Given outline for complex topic Q, retrieve for each of its sections H_i, a ranking of relevant entities E and with support passages S. These support passage should motivate the why the entity is relevant for the query.

The passage S is taken from the provided passage corpus. The entity E refers to an entry in the provided knowledge base. We define a passage or entity as relevant if the passage content or entity is appropriate for the knowledge article.

*dietz@cs.unh.edu

MUST be mentioned:
Sugar glass is made by dissolving sugar in water and heating it to at least the "hard crack" stage (approx. 150 C / 300 F) in the candy making process. Glucose or corn syrup is used to prevent the sugar from recrystallizing, by getting in the way of the sugar molecules forming crystals. Cream of tartar also helps by turning the sugar into glucose and fructose.

CAN be mentioned:
Sinuklob is a very sweet candy that is also used in making Bukayo. It is the cheaper version of caramel in the Philippines. Sinuklob is made from melted brown sugar hardened into a chewy consistency.

Roughly on TOPIC but non-relevant:
Most candies are made commercially. The industry relies significantly on trade secret protection, because candy recipes cannot be copyrighted or patented effectively, but are very difficult to duplicate exactly. Seemingly minor differences in the machinery, temperature, or timing of the candy-making process can cause noticeable differences in the final product.

Paragraph IDs:
left: 1b7d202a3dc37f69b4d2e3972eabb76920b6ca23
center: 5037e499734f75ecaccdce333021d793e59fd3f4
right: 6f503a6b2e3ee4fa33c2b0c4bee50aaba6c12dbf

Figure 2: Example passages and relevance for "Candy Making / Hard Candy" (Query ID "Candy%20making/Hard%20candy").

3 TREC CAR Data Set (v1.5)

The 2017 Complex Answer Retrieval track uses topics, outlines, and paragraphs that are extracted from English Wikipedia (XML dump from Dec 20th, 2016). Wikipedia articles are split into the outline of sections and the contained paragraphs. All paragraphs from all articles are gathered and deduplicated to form the paragraph corpus.

Each section outline is a description of a complex topic. By keeping the information which paragraph originates from which article and section, we have a means of providing a ground truth for the passage retrieval task. By preserving hyperlinks inside paragraphs that point to Wikipedia pages (also known as entities in the DBpedia knowledge graph), we have a means of providing a ground truth for the entity retrieval task.

```
Title: Candy Making
1. History
2. Safety
3. Hard candy
        3.1. Sugar stages
4. Soft candy
        4.1. Cotton Candy
        4.2. Marshmallows
5. Chocolatiering
6. Tools and machinery
```

Figure 1: Example outline for complex topic "Candy Making".

Through filtering and processing procedures described in Section 3.1, several datasets are derived. The size of the datasets is given in Table 1. The paragraph collection contains 29,678,367 unique paragraphs.

	benchmarkY1train	benchmarkY1test	train	test200
number of articles (complex topics)	117	133	285,924	198
hierarchical sections (queries)	1,816	2,125	2,180,868	1,860
total positive paragraphs assessments	4,530	5,820	5,276,624	4,706
total positive entity assessments	13,031	15,085	12,310,616	11,396

Table 1: Data set sizes in terms of articles, section, and automatic positive assessments.

3.1 Data Set Creation Pipeline

The TREC Complex Answer Retrieval benchmark (v1.5) is derived from Wikipedia so that complex topics are chosen from articles on open information needs, i.e., not people, not organizations, not events, etc. However, any paragraph or entity on Wikipedia is a legal paragraph/entity for the retrieval task even if a person entity or a paragraph from an article on an event. The data set creation process is as follows:

1. Mediawiki format of each article in the Wikipedia dump is parsed, preserving paragraph boundaries, intra-Wikipedia hyper links, and section hierarchy.
2. Templates, talk pages, portals, disambiguation, redirect, and category pages are discarded (redirect information is provided separately as **entity-redirects**).
3. Articles tagged with categories that indicate people, organizations, music, books, films, events, and lists are discarded.
4. Sections with headings that do not contain prose are discarded, for example external links, references, bibliography, notes, gallery etc.
5. Each article is separated into the outline of section headings on the one hand and paragraphs on the other hand.
6. The set of paragraphs across all of Wikipedia are collected, and uique paragraph IDs are derived through SHA256 hashes on the text content (ignoring links).
7. The paragraphs are further deduplicated with min hashing using word embedding vectors provided by GloVe. This is called the **paragraphcorpus**. Articles are rewritten to reference only the deduplicated set of paragraphs.
8. The set of articles is further filtered, to remove images, lead sections, sections with very long (>100 character), and very short headings (<3 letters). Articles with less than three remaining sections are discarded.
9. The set of articles is split into training and test data, training data is further split into five folds. To ensure uniform distribution and reproducibility, these decisions are made based on the SipHash of the article title.
10. The five folds of the training data, with separated outlines and paragraphs and extracted automatic qrels are made available as **train**.
11. For the benchmark, a total of 250 complex topics from English Wikipedia were manually selected, preferring topics that require a complex answer such as "Candy making", "Permaculture", "Soil Erosion", or "Air Ioniser". These were divided into train and test sets by training/test split defined previously in step 9.
12. Benchmark topics that are part of the training set are released with articles, outlines, and automatic ground truth as **benchmarkY1train**.
13. Benchmark topics that fall into the test set are only released as outlines as **benchmarkY1test.public**. Official contributed runs to be submitted on these topics. Articles and automatic ground truth will be released as benchmarkY1test after the TREC workshop.

Pages from step 2, that would be in the training set (based on the SipHash of the page ID) are released as **unprocessedtrain**.

A dataset based on an earlier selection of 200 complex topics from training fold 0 is released with articles, outlines, and automatic ground truth as **test200**.

3.2 Automatic Ground Truth

Two kinds of ground truth signals are collected: automatic and manual. For each, we release true paragraphs and true entities. While the manual ground truth is assessed after participants submit runs, the automatic ground truth is derived along with the dataset from the Wikipedia dump. The automatic ground truth is released for all training sets and derived as follows.

- If a paragraph is contained in the page/section it is defined as relevant, and non-relevant otherwise. The ground truth signal is released for three granularities: paragraph contained in section (`hierarchical`), paragraph contained in section hierarchy below top level section (`toplevel`), paragraph contained anywhere in the page (`article`).

- After resolving redirects, if a hyperlink to an entity is contained in the page/section it is defined as relevant, and non-relevant otherwise. As with paragraphs, three granularities hierarchical, top level, and article are collected.

These six automatic ground truth files are released for train and benchmarkY1train before the submission, and for benchmarkY1test after the evaluation. Paragraphs that are relevant according to the automatic ground truth will be added to the pool for manual assessments (cf. Section 5).

4 Submission

For the passage ranking task, participants were asked to submit a ranking of paragraph IDs per heading in the outlines of benchmarkY1test. To participate in the entity task, participants submitted a ranking of entity IDs for each heading. To provide provenance and simplify the manual assessment, participants where asked to complement each entityId with a paragraph that explains why the entity is relevant for the corresponding section heading.

Participants were allowed to consider all headings in the outline at once, use external resources such as knowledge graphs, entity linking tools, pre-trained word embeddings, and any of the provided TREC CAR data sets. The participants were not allowed to directly use a dump of Wikipedia, as this would allow them to look up the paragraphs on the page—the information used in the automatic ground truth.

Each participating team was allowed to submit up to three runs to the passage task and three runs to the entity task. Seven teams participated in this first year of the track.

5 Assessment of the Manual Ground Truth

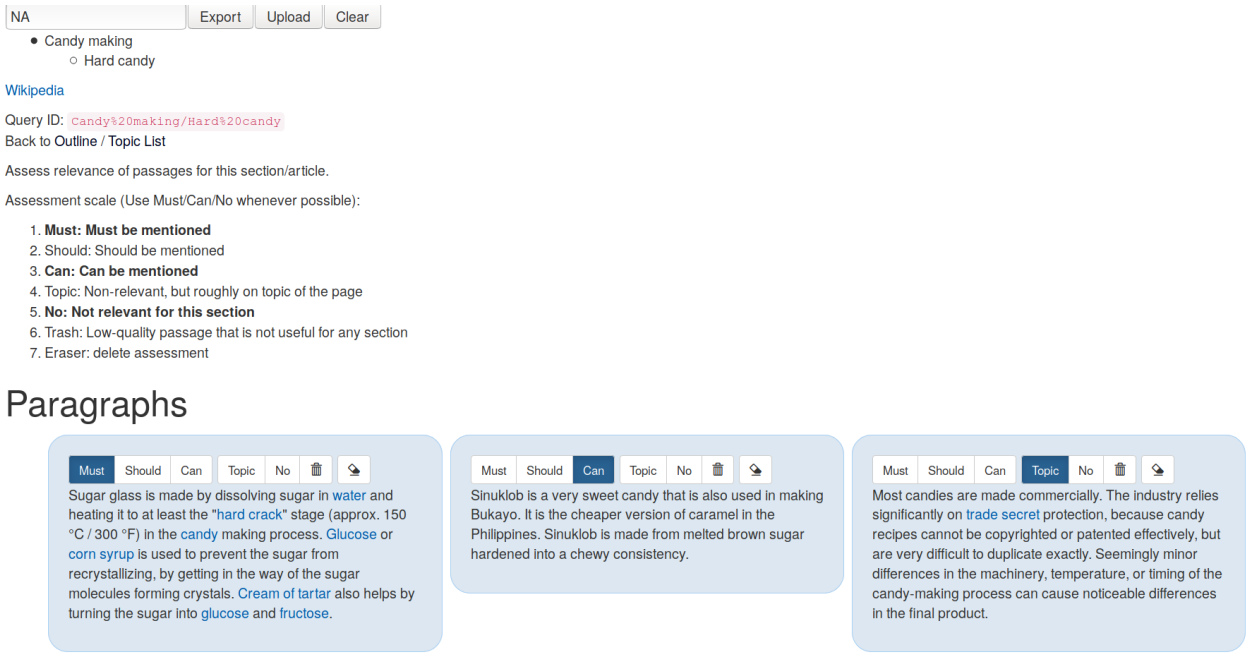
The top elements of participant contributed runs were merged to build a pool of 50 paragraphs/entities per topic section. Additionally paragraphs relevant according to the ground truth were added to the pool for verification. For all 113 complex topics in benchmarkY1test, at least three sections (of the assessors choice) were annotated. A small set of topic sections were annotated by all assessors in order to measure inter-annotator agreement across the six NIST assessors (cf. Section 5). A screenshot of the assessment interface is given in Figure 3.

The assessor is presented with the complex topic (page title) and the topic sections as a heading hierarchy, followed by a list of paragraphs when assessing the passage task. In the case of the entity task, the list displayed the canonical entity names together with the provenance paragraph if given. In cases where participants did not submit provenance, the first paragraph from the entity's Wikipedia pages was displayed. In both cases, assessors were asked to judge the relevance of a paragraph or entity solely based on information displayed, and not resort to their world-knowledge. If more context would be needed to for the paragraph to become relevant, the assessors were instructed to assess it as non-relevant.

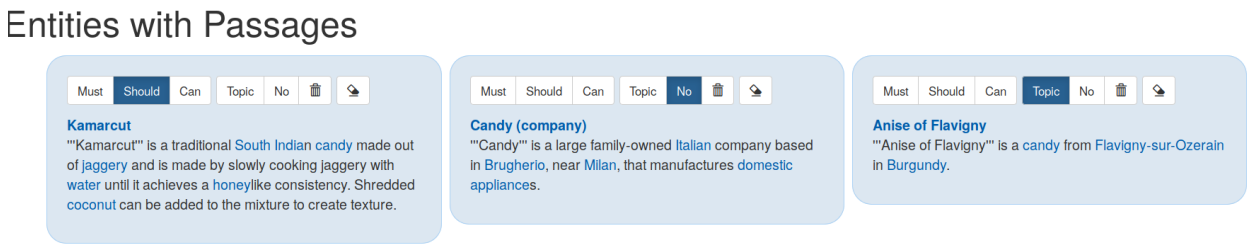
Assessors were asked to envision writing a Wikipedia article on the given complex topic. A graded assessment scale was used based on how much the paragraph/entity should be mentioned in this section of the article.

- MUST be mentioned
- SHOULD be mentioned
- CAN be mentioned
- Non-relevant, but roughly on TOPIC of the page
- NO, non-relevant
- Trash

Trash is assigned to paragraphs/entities that is of low quality therefore would not be relevant for any topic imaginable.



(a) Passage task.



(b) Entity task (excerpt).

Figure 3: Assessment interface.

Table 2: Assessment scale for manual assessments. Horizontal line: Cutoff for positive/negative assessments.

	binary scale	manual scale	manual lenient scale
MUST be mentioned	1	3	5
SHOULD be mentioned	1	2	4
CAN be mentioned	1	1	3
Non-relevant, but roughly on TOPIC	0	0	2
NO, non-relevant	0	-1	0
Trash	0	-2	-2

Table 3: Grade histogram and distribution.

(a) Passage judgments.

	annotator1	annotator2	annotator3	annotator4	annotator5	annotator6	Total %
Trash	25	8	4	6	2	0	0%
No	2869	1875	2910	1971	1381	2254	43%
Topic	536	1000	1349	2853	1665	2491	32%
Can	213	703	1186	263	114	827	11%
Should	380	12	827	120	226	563	7%
Must	340	771	131	874	229	256	8%

(b) Entity judgments.

	annotator1	annotator2	annotator3	annotator4	annotator5	annotator6	Total %
Trash	2	1	10	17	10	1	0%
No	1363	1060	2142	1979	989	1901	85%
Topic	33	223	129	259	224	283	10%
Can	10	114	80	16	16	52	3%
Should	18	4	15	12	26	26	1%
Must	5	76	4	18	11	9	1%

Label distribution

Six assessors created 42,372 annotations on a total of 704 topic sections. For 702 topic sections passage assessments were created with 31,186 assessments in total. The grade histogram per annotator and the overall grade distribution is given in Table 3a. We notice that only a quarter of all passages are graded as relevant, while an additional third were annotated as being on topic. 43% of passages in the assessment pool were marked as absolutely non-relevant. This demonstrates the feasibility of the task, while indicating that it is much easier to identify relevant for the general topic than for a particular aspect.

For 640 topics sections, entity assessments were created with 11,095 assessments in total. Histogram and distribution is given in Table 3b. While the statistic seems to suggest that the entity retrieval problem is much harder, we believe that this is due to an issue with the submitted runs. One team indicated that a bug in the pipeline lead to random results. The remaining entity runs were derived from passage runs with some heuristics. An example of entities that are marked as relevant for the complex topic "Cocoa bean" is given in Figure 6.

Inter-annotator agreement

A small number of topic sections are selected for annotation by all assessors to measure inter-annotator agreement. Two passage and two entity topic sections were annotated at the beginning of the assessment cycle, additionally three passage and one entity topic sections were annotated towards the end of the assessment period.

We measure inter-annotator agreement using Cohen's κ for pairwise comparison and Fleiss' κ across all annotators. We analyze agreement on the derived binarized assessment (Table 4a) and the original graded assessment (Table 4b). We are aware of the subtle difference between neighboring grades very subtle. Therefore, we additionally consider the case of graded assessment we call "off by one" in Table 4c, where assessments that differ by no more than one grade step, e.g., grades "SHOULD" and "MUST", are also counted as agreements for both for p_0 and p_e.

Inspecting Cohen's κ, we find that on the whole the pair-wise agreement is relatively similar across all pairs of assessors. In other words, there is no "odd one out" which speaks to the quality of NIST's assessment procedures. As expected, the agreement for binarized judgments (Fleiss $\kappa = 0.574$) is higher than for graded judgments (Fleiss $\kappa = 0.273$). This may sound small, yet it is comparable to previous work [2]. However, once neighboring grades are counted as agreement ("off by one"), the inter-annotator agreement is even a bit higher agreement on binarized assessments.

Table 4: Inter annotator agreement according to Cohen's κ and Fleiss' κ.

(a) Binary (TOPIC counting as negative). Fleiss $\kappa = 0.574$

	annotator1	**annotator2**	**annotator3**	**annotator4**	**annotator5**	**annotator6**
annotator1		0.569	0.602	0.742	0.574	0.544
annotator2	0.569		0.574	0.658	0.458	0.417
annotator3	0.602	0.574		0.667	0.615	0.553
annotator4	0.742	0.658	0.667		0.638	0.575
annotator5	0.574	0.458	0.615	0.638		0.733
annotator6	0.544	0.417	0.553	0.575	0.733	

(b) Graded. Fleiss $\kappa = 0.273$

	annotator1	**annotator2**	**annotator3**	**annotator4**	**annotator5**	**annotator6**
annotator1		0.255	0.425	0.326	0.281	0.152
annotator2	0.255		0.214	0.379	0.181	0.336
annotator3	0.425	0.214		0.314	0.322	0.238
annotator4	0.326	0.379	0.314		0.334	0.380
annotator5	0.281	0.181	0.322	0.334		0.377
annotator6	0.152	0.336	0.238	0.380	0.377	

(c) Graded, counting grades that are "off by one" as agreement.

	annotator1	**annotator2**	**annotator3**	**annotator4**	**annotator5**	**annotator6**
annotator1		0.603	0.686	0.758	0.635	0.658
annotator2	0.603		0.581	0.739	0.485	0.520
annotator3	0.686	0.581		0.692	0.685	0.693
annotator4	0.758	0.739	0.692		0.623	0.625
annotator5	0.635	0.485	0.685	0.623		0.812
annotator6	0.658	0.520	0.693	0.625	0.812	

We conclude that, aside from subtle nuances in the grading scale, assessors agree on the whether the passage or entity should be included in the article on the complex topic.

Annotation Time

Over the course of two weeks, six assessors were hired for 40 hours each, yielding a total of 240 hours. Excluding breaks and training, the average annotation time per passage or entity judgments is 22 seconds.

6 Participant Submitted Runs

In total seven teams contributed runs. The majority of submissions were passage runs. Three teams submitted neural network ranking methods. Most methods are based on Lucene's BM25 ranking model as a candidate method. Some methods used entity linking, pre-trained word vectors, and other forms of query expansion. Table 5 gives an overview over the kinds of methods contributed by participating teams. Below, detailed descriptions of submitted runs:

- The MPIID5 submission includes three runs for passage retrieval, all of which are extensions of the PACRR [5] deep neural relevance ranking architecture: *nn6pos*, *nn4poshperc*, and *nn6postprob*. Based on a set of initial results provided by BM25, PACRR reranks the results based on query-document interactions captured by convolutional filters. We extend the PACRR architecture by providing new contextual signals along with the existing query term match signals. All three runs include a contextual

	CUIS	ECNU	ICTNET	MPIID5	NYUDL	TREMA-UNH	UTDHLTRI
Uses entity linking						X	X
Uses a knowledge graph						X	X
Uses pre-trained word embeddings				X	X		X
Used neural network				X	X		X
Uses learning to rank		X			X		X
unsupervised			X			X	
Uses BM25	X	X	X	X	X	X	X
Uses SDM	X						
Language model							X
Passage Task	X	X	X	X	X	X	X
Entity Task	X					X	

Table 5: Participant contributed runs.

heading position vector (*pos*), indicating if a given query term was from the title, intermediate heading, or bottommost heading. The *hprec* run includes a vector that models the prevalence of each heading in the training set. The *tprob* run includes a vector that estimates the likelihood that each query term is present in relevant paragraphs.

- The UTDHLTRI team developed the Complex Answer PAragraph Retrieval (CAPAR) system to perform complex answer retrieval consisting of the following five modules: (1) The Paragraph Indexing Module creates a searchable index of paragraphs from Wikipedia articles; (2) The Query Processing Module processes a Wikipedia article outline into a set of queries - one for each section of the outline; (3) The Paragraph Search Module searches each query against the paragraph index, resulting in a list of relevant paragraphs for each section in the article outline; (4) The Feature Extraction Module is used to extract features from each paragraph; (5) The Paragraph Ranking Module produces a separate ranking of the retrieved paragraphs for each section. We use one of two Learning to Rank (L2R) systems to calculate relevance scores for each paragraph: (1) the Siamese Attention Network (SANet) for Pairwise Ranking and (2) AdaRank. A ranking is produced for each section of the article outline using the relevance scores from one of the two L2R systems. Their best run uses SANet which combines traditional IR features with a learned, attention-based semantic matching function.

- Team TREMA-UNH provided two passage runs and three entity runs. The passage runs are based on Lucene/Solr's BM25 method ($k1 = 1.2$, $b = 0.75$) using the page title, section heading and headings of parent sections as a query. The first method is just BM25 the second BM25 with uses query expansion. Expansion terms are selected through two sources: 1) Using the TagMe API [4] with the "include_abstract" option and only considering terms with the highest IDF score; 2) Based on ideas of Banerjee and Mitra [3], in the provided collection of Wikipedia articles in training set ("unprocessed_train"), content of sections with the same headings from other pages are extracted and identifying terms with the highest IDF terms. Up to 25 query expansion terms were selected, giving preference to terms selected through both sources. The entity runs are based on graph walks on the knowledge graph generated by co-mentions of entities.

- Team CUIS provided one passage and one entity run. The passage run uses Lucene's BM25 implementation to create a candidate set. This set is re-ranked using the sequential dependence model [6]. From this passage run, an entity run is derived by replacing the paragraph id with the containing article.

- Team NYUDL provides three passage runs using their neural retrieval network[1] [8]. The system consists

[1] `https://github.com/nyu-dl/QueryReformulator`

of two stages. The first stage is query reformulation based on neural networks that rewrites a query to maximize the number of relevant documents returned. We train this neural network with reinforcement learning. The actions correspond to selecting terms to build a reformulated query, and the reward is the document recall. The second stage is a binary classifier based on a neural network that selects relevant documents. This classifier is trained with supervised learning. For all runs, BM25 is used to create a candidate set. The run "ds" is a simple document classifier using avg word embeddings in the documents as document vector and last hidden state of an LSTM as query vector. A 2-layer feed forward neural net is used to select which documents are relevant given a query. The run "qr" uses BM25 with query reformulation with deep reinforcement learning. The run "qrds" uses query reformulation using reinforcement learning + Lucene + Neural Net Classifier to select documents. All of these methods are trained with data provided in train-v1.5.

- Team ICTNET provides one passage run based on BM25.
- Team ECNU provides on passage run as follows: First, Lucene's BM25 is used to select candidate paragraphs, and then we use the BM25 score and word matching is used as features in a learning-to-rank framework.

7 Results

We evaluate participant-contributed runs on automatic and manual assessments with respect to four standard TREC evaluation measures, R-Precision (RPrec), Mean-average Precision (MAP), Reciprocal Rank (MRR), and Normalize Discounted Cumulative Gain (NDCG). Of these measures only NDCG includes a graded scale, for all other methods will use the positive/negative cutoff indicated in Table 2.

Results are presented in Figure 4 on the automatic binary scale, the manual graded scale, and a lenient variant of the manual graded scale. Standard error bars are given for reference. All analyses across all measures as well as automatic and manual assessments are painting the same picture. Acknowledging consistent patterns in the results, here the ranking of methods by across-the-board performance:

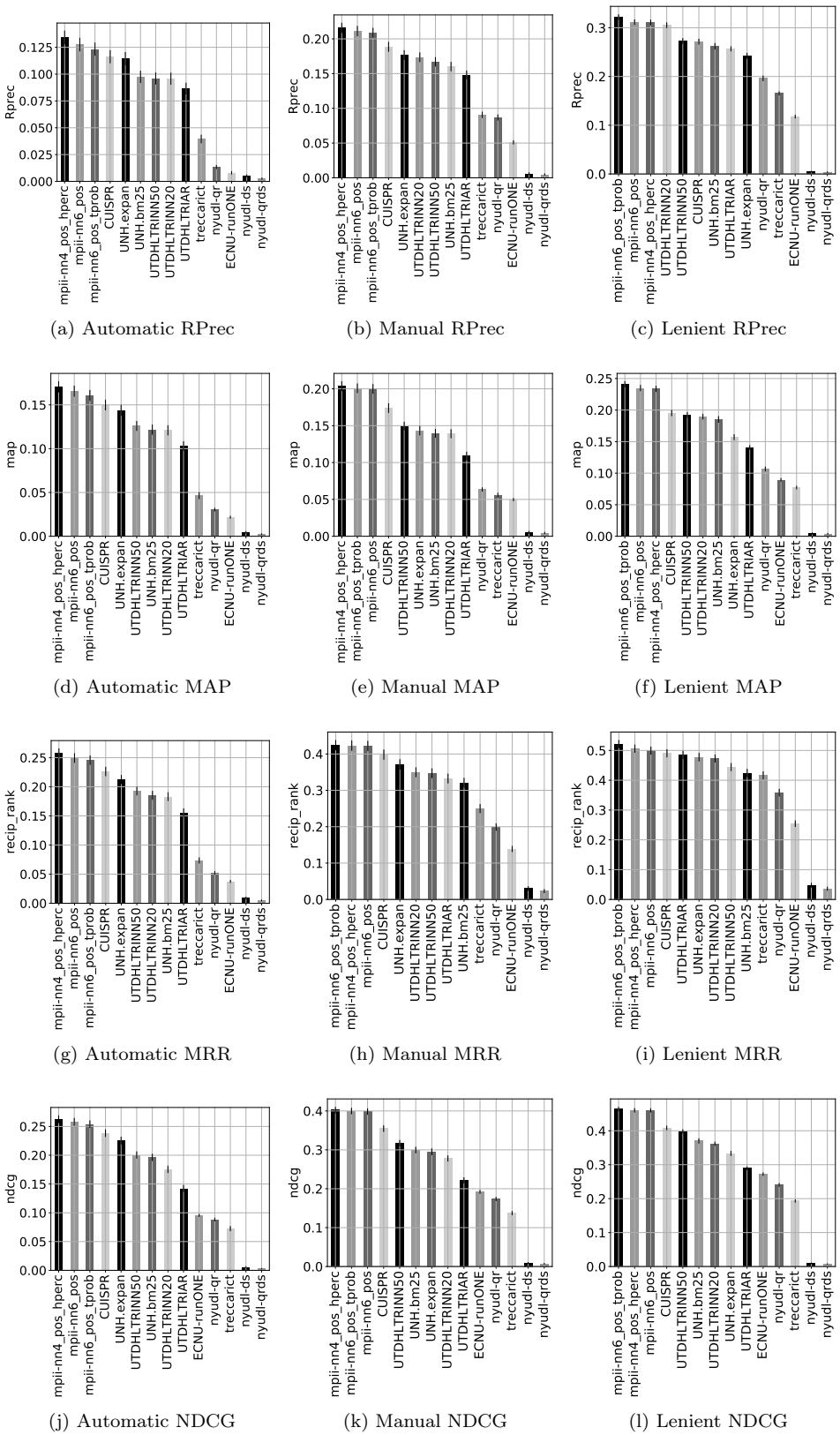
1. All variations of the neural network PACRR (MPIID5), which re-ranks a BM25 candidate set.
2. CUISPR (CUIS), which uses the sequential dependence model to rerank a BM25 candidate set.
3. BM25, BM25 with DBpedia expansion (UNH-TREMA), and the neural Siamese attention network ranking variants (UTDHLTRI), which are also DBpedia expansion and pre-trained embeddings. In comparison to plain BM25, DBpedia expansion is advantageous most of the time, while neural network-based ranking seems to retrieve more paragraphs on that are roughly on topic.
4. DBpedia expansion features combined with AdaRank-based learning-to-rank (UTDHLTRIAR).
5. ICT's method, "qr" method (NYUDL), and "runOne" (ECNU),
6. Neural network methods "ds" and "qrds" (NYUDL)

To attempt an explanation, NYUDL's submissions only included a binary set of relevant (versus non-relevant) paragraphs. In the vast majority of cases these included only a single paragraph per topic section.

To study whether the differences are due to better performance on easy queries, difficult queries, or overall, we include divide the set of all annotated topic sections into percentiles ranging from easy to difficult according to the BM25 baseline. The results are presented in Figure 5 and show, for instance, that all methods of MPIID5 are consistently the best on all but the easiest quartile of section topics.

8 Conclusion

In this first year of the TREC Complex Answer Retrieval track we learned a lot about the structure of the problem.



(a) Automatic RPrec (b) Manual RPrec (c) Lenient RPrec

(d) Automatic MAP (e) Manual MAP (f) Lenient MAP

(g) Automatic MRR (h) Manual MRR (i) Lenient MRR

(j) Automatic NDCG (k) Manual NDCG (l) Lenient NDCG

Figure 4: Results of contributed passage runs under automatic and manual ground truth. Lenient is based on the manual graded scale, but counting TOPIC as relevant.

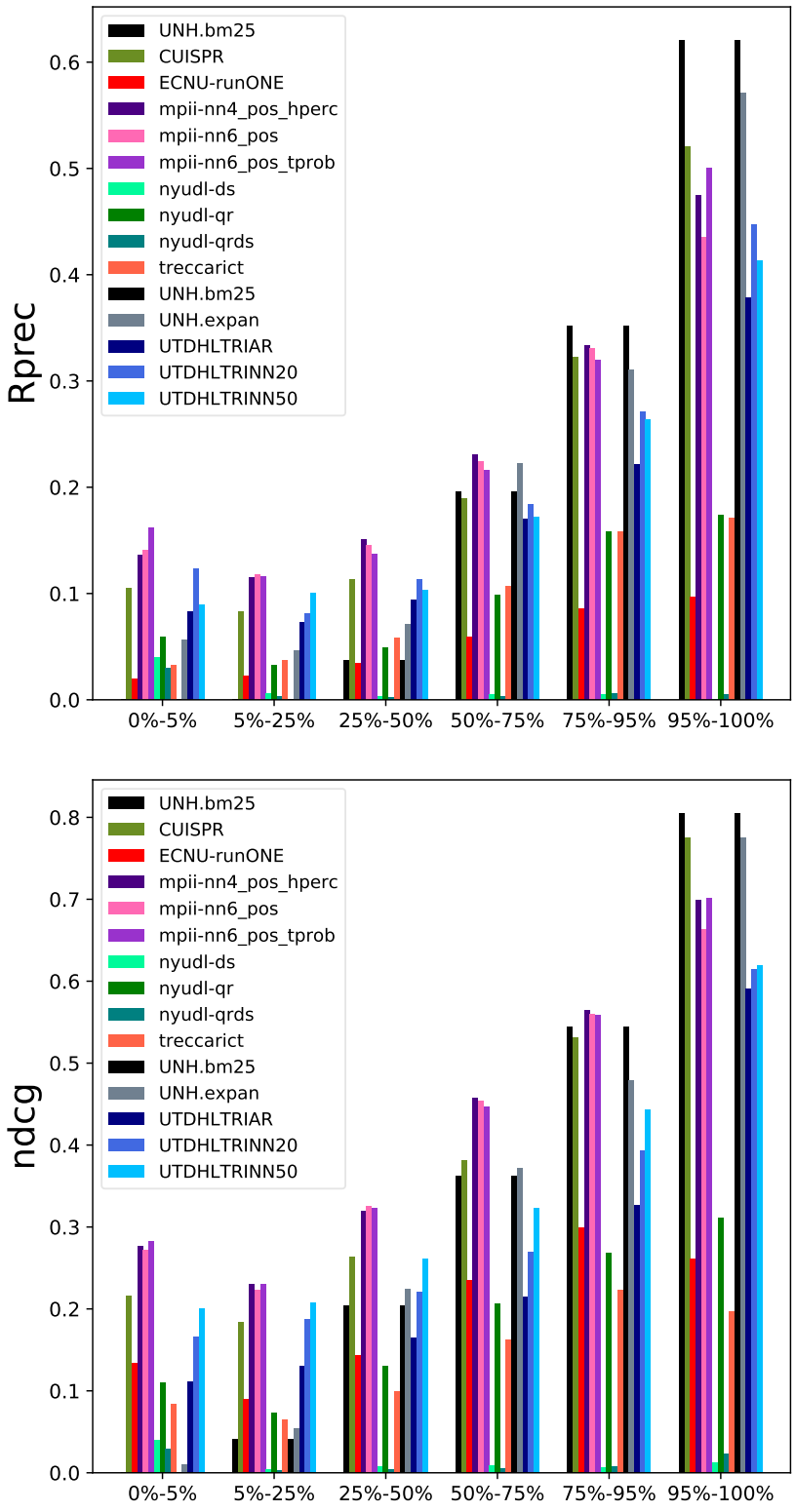


Figure 5: Performance according to manual truth on difficulty percentiles according to BM25.

- Human assessors can agree which passages and entities are relevant. We found that there are two predominant reasons why passages are not relevant: they are either completely missing the topic (NO) or they are roughly on topic, but non-relevant for the given section (TOPIC).
- Automatic and Manual ground truth agree which of the contributed systems works better than others. However, many retrieved passages that are marked as incorrect by the automatic ground truth, are actually correct when manually inspected.
- With much care, neural network methods can work significantly better than unsupervised methods. However, the distance to unsupervised methods such as, the sequential dependence model and BM25 is small (which corroborates earlier findings [7]). Using DBpedia as a source for expansion provides an advantage.

After the success of the first year, we are looking forward to exploring exciting directions in year two.

Acknowledgement

We express our gratitude for many suggestions of several experts in the field, who helped to make this track successful. Special thanks to Fernardo Diaz on whose ideas this evaluation is based on. We thank the University of New Hampshire for providing computational resources and web servers. We are grateful for Ben Gamari's invaluable support in developing the benchmark creation and assessment interface software. We are deeply thankful for Ellen Voorhees' experience, patience, and persistence in running the assessment process. Finally we thank all our participants.

References

[1] James Allan, W. Bruce Croft, Alistair Moffat, and Mark Sanderson. Frontiers, challenges, and opportunities for information retrieval: Report from SWIRL 2012 the second strategic workshop on information retrieval in Lorne. *SIGIR Forum*, 46(1):2–32, 2012. doi: 10.1145/2215676.2215678. URL `http://doi.acm.org/10.1145/2215676.2215678`.

[2] Omar Alonso and Stefano Mizzaro. Using crowdsourcing for trec relevance assessment. *Information Processing & Management*, 48(6):1053–1066, 2012.

[3] Siddhartha Banerjee and Prasenjit Mitra. Wikikreator: Improving wikipedia stubs automatically. In *ACL (1)*, pages 867–877, 2015.

[4] Paolo Ferragina and Ugo Scaiella. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1625–1628. ACM, 2010.

[5] Kai Hui, Andrew Yates, Klaus Berberich, and Gerard de Melo. Pacrr: A position-aware neural ir model for relevance matching. In *EMNLP*, 2017.

[6] Donald Metzler and W Bruce Croft. A markov random field model for term dependencies. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 472–479. ACM, 2005.

[7] Federico Nanni, Bhaskar Mitra, Matt Magnusson, and Laura Dietz. Benchmark for complex answer retrieval. In *ICTIR '17 Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, pages 293–296. https://doi.org/10.1145/3121050.3121099, 2017.

[8] Rodrigo Nogueira and Kyunghyun Cho. Task-oriented query reformulation with reinforcement learning. *arXiv preprint arXiv:1704.04572*, 2017.

Title: Cocoa bean

1. Etymology
- Could: Criollo (cocoa bean)

2. History
- Should: Cadbury

3. Production
3.1. Cocoa pod
- Should: Phytophthora

3.2. Varieties
- Must: Criollo (cocoa bean)

3.3. Harvesting
3.4. Harvest processing
- Must: Theobroma cacao
- Could: Chocolate liquor

3.5. World production
- Must: Cote d'Ivoire
- Could: Criollo (cocoa bean)

3.6. Child slavery
- Must: Cote d'Ivoire
- Must: Children in cocoa production
- Must: Trafficking of children
- Could: Sub-Saharan Africa

4. Cocoa trading
4.1. Fair trade
- Must: Fair Trade
- Must: Global Exchange
- Must: Green America
- Must: Office of Fair Trading
- Could: Ghana Cocoa Board
- Could: Organic chocolate

4.2. Consumption
- Could: Cameroon # 5. Chocolate production
- Must: Baking chocolate
- Must: Chocolate liquor
- Must: Milk chocolate
- Must: Sweet chocolate
- Must: White chocolate
- Could: International Cocoa Organization

6. Health benefits

7. Environmental impact
7.1. Agroforestry
- Must: CGIAR

Figure 6: Positively annotated entities for the outline of the complex topic "Cocoa bean".