COMEX: A Multi-task Benchmark for Knowledge-grounded COnversational Media EXploration

Zay Yar Tun
University of Glasgow
Glasgow, United Kingdom

Alessandro Speggiorin
University of Glasgow
Glasgow, United Kingdom
alessandro.speggiorin@glasgow.ac.uk

Jeffrey Dalton
University of Glasgow
Glasgow, United Kingdom
jeff.dalton@glasgow.ac.uk

Megan Stamper
BBC
Glasgow, United Kingdom
megan.stamper@bbc.co.uk

ABSTRACT

Open-domain conversational interaction with news, podcasts, and other types of heterogeneous content remains an open challenge. Interactive agents must support information access in a way that is fair, impartial, and true to the content and knowledge discussed. To facilitate this, systems building on interactive retrieval from knowledge-grounded media are a controllable and known base for experimentation. A conversational media agent should retrieve relevant content, understand key concepts in the content through grounding to a knowledge base, and enable exploration by offering to discuss a topic further or progress to describe related topics. In this work, we release a new multi-task benchmark on COnversational Media EXploration (COMEX) to measure knowledge-grounded conversational content exploration. It consists of a heterogeneous semantically annotated media corpus and topic-specific data for 1) entity Wikification and salience, 2) conversational passage ranking on heterogeneous media content, 3) background link ranking, and 4) background linking explanation. We develop COMEX with judgments and conversational interactions developed in partnership with professional editorial staff from the BBC. We study the behavior of state-of-the-art systems, with the results demonstrating significant headroom on all tasks. [1]

KEYWORDS

conversational search, news discovery, entity knowledge graphs

ACM Reference Format:
Zay Yar Tun, Alessandro Speggiorin, Jeffrey Dalton, and Megan Stamper. 2022. COMEX: A Multi-task Benchmark for Knowledge-grounded COnversational Media EXploration. In *4th Conference on Conversational User Interfaces (CUI 2022), July 26–28, 2022, Glasgow, United Kingdom.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3543829.3543830

[1] COMEX is available at https://github.com/grill-lab/COMEX with an MIT license.

CUI 2022, July 26–28, 2022, Glasgow, United Kingdom

ACM ISBN 978-1-4503-9739-1/22/07…$15.00
https://doi.org/10.1145/3543829.3543830

1 INTRODUCTION

The rise of voice-enabled smart speakers and other devices means that media consumption and interaction are shifting from the web and radio to ones mediated by personal assistants such as Alexa, Google Assistant, and others. However, the existing systems mainly focus on simple commands to play single pieces of content. In this work, we envision the future of conversational assistants that support consumption and exploration of rich multimedia content incorporating social discussion of the subject. In COMEX, we focus on building blocks of such a system that allows interactive discussion and interaction with content on topics of interest to facilitate learning and exploration.

Open-domain conversational interaction about information topics remains a fundamental challenge in conversational AI systems that build upon natural language understanding and large-scale information retrieval. Recent advances in conversational chatbots build on neural generative models [Adiwardana et al. 2020; Roller et al. 2021] leveraging large-scale pre-trained language models. However, while these models produce fluent and natural responses, they hallucinate, generating text that is nonsensical [Ji et al. 2022], and struggle to remain true to the source content [Dziri et al. 2021]. Recent efforts focus on grounding responses in longer text documents, often with knowledge based on Wikipedia and a small number of source documents [Dinan et al. 2019; Gopalakrishnan et al. 2019]. However, even in these cases generating responses remains challenging and error-prone. In contrast, in this benchmark, we build on conversational retrieval of content that is semantically grounded in entity knowledge bases.

COMEX brings together multiple separate threads of work that are usually evaluated separately in conversational systems and grounded language understanding into a unified multi-task benchmark, measuring key aspects of a conversational media exploratory system. It builds upon ideas and models from conversational retrieval in the TREC Conversational Assistance Track (CAsT) [Dalton et al. 2019]. To support a person exploring the news with a conversational agent it includes exploratory tasks (background linking) from the TREC News track [Soboroff et al. 2018] and knowledge-grounding of content (Wikification). To facilitate the voice assistant use case it includes diverse text and audio programs (podcasts, radio, etc.). Instead of navigation, it focuses on exploration of topics similar to previous work media exploration from the TREC Podcast track [Jones et al. 2020].

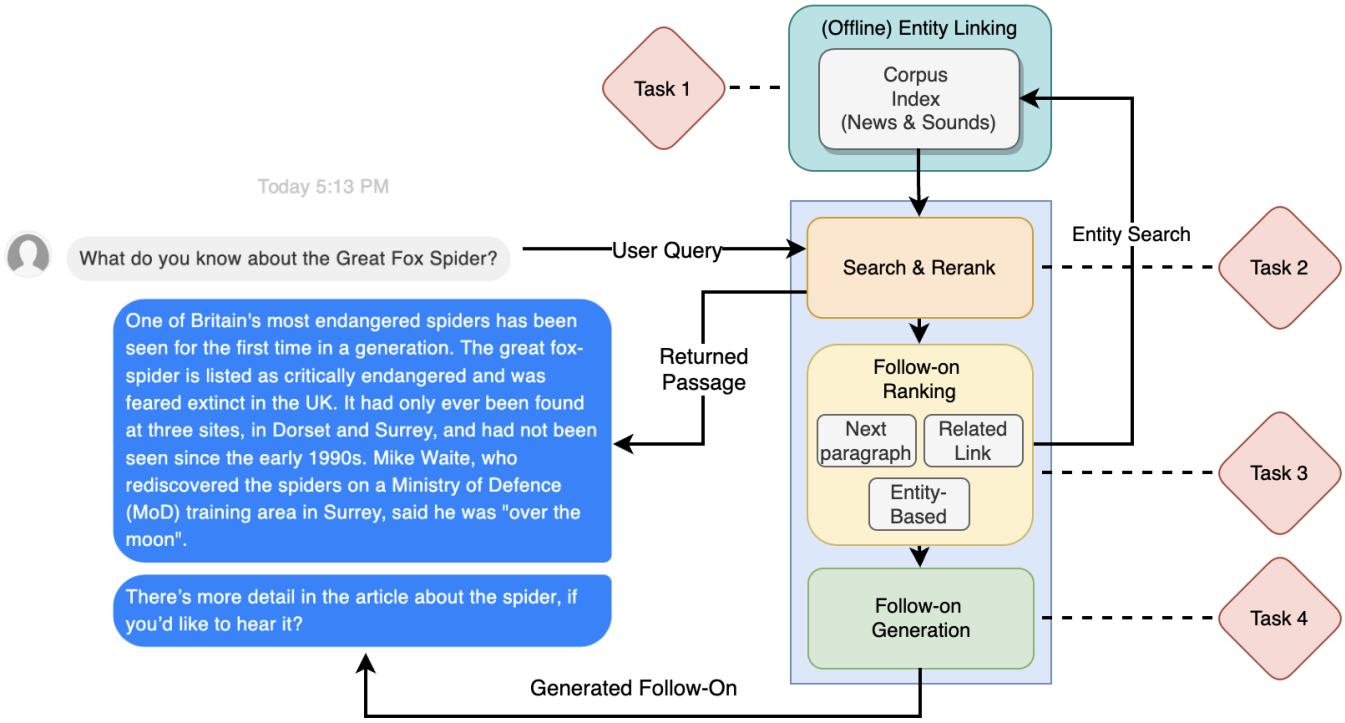

Figure 1: `UoG Bot` architecture and relationship to `COMEX` Tasks.

`COMEX` supports measurement social conversational news and audio content discovery developed in conjunction with the BBC and its internal AI team. The benchmark evaluates key components that correspond to distinct modules from the `University of Glasgow (UoG) Bot`, a research prototype conversational system integrated with the assistant developed at the BBC. These components include offline entity linking, conversational query-rewriting, and follow-up generation.

The goal of `COMEX` is to support research on conversational systems that facilitate social discussion of emerging topics in the news and media. Its objective is to support discovery and exploration of relevant related information across both breadth (related topics) and depth (sub-topic exploration). `COMEX` encompasses not just one type of content but heterogeneous content from the BBC News and Sounds to facilitate cross-media exploration.

`COMEX` includes fundamental tasks for a knowledge-grounded conversational agent with multiple novel elements that make it distinctive. First, unlike previous chat benchmarks [Dinan et al. 2019; Gopalakrishnan et al. 2019], it is designed as a large-scale open retrieval task for content across heterogeneous media. The corpus includes all the available News content and Sounds metadata (not full audio or transcripts) produced by the BBC from January 2017 to October 2020, with over 800,000 pieces of content. To support exploration, professional editorial staff develop topics and synthetic personas. Topics are popular or trending on news and social media platforms. The synthetic personas are artificially created by editorial staff inspired by the BBC aggregated audience data.

Figure 1 shows the high-level architecture of the `UoG Bot` prototype and how the tasks relate to key sub-components. The benchmark tasks correspond to critical components in the architecture spanning knowledge-grounding, passage retrieval, conversational suggestion, and suggestion explanation. The first task is the *entity Wikification and salience* task, with entity detection and disambiguation to the Wikidata knowledge base [Vrandečić and Krötzsch 2014]. This measures the system's ability to detect key topics in the content. The second task is *conversational passage response ranking* to identify relevant content in conversational search. The third task is an extension of the TREC *background linking* task [Soboroff et al. 2018] which identifies relevant related information to a piece of content. This is important because relevant related content forms the basis for system initiative to guide users to explore diverse aspects of topically related information.

The final `COMEX` task is *background linking explanation*, a new task that contextualizes the content found in background linking and provides an explanation for how two pieces of content are related to one another in the context of an exploratory conversation. The task is to create a natural language explanation that connects two pieces of content based upon previous conversational history. For example, consider a conversation where a turn includes the question: *"What do you know about the Great Fox Spider?"*. After finding a relevant response and related content to suggest (background linking), the system provides a natural language rationale for the recommendations. For example, *"If you're interested in spiders, there is a radio program called Nature that talks about the Fern*

Spider, or Gossip from the Garden Pond that talks about 'The Garden Spider and the Great Pond Snail'". This task is important because it facilitates conversational exploration.

Together, the tasks in COMEX measure key elements of a system's ability to ground content to knowledge, retrieve and rank relevant content, and find and explain related documents to carry a conversation forward. To our knowledge, no previous benchmark incorporates all of these components on a single heterogeneous real-world collection of media content. Moreover, in contrast to most work that uses crowdsourcing, this benchmark uses curated topics, annotations, and Wizard-of-Oz (WoZ) conversations with professional editorial staff. In this work, we use COMEX to study the behavior of state-of-the-art methods for each task and provide these as strong baselines for future comparison.

The contributions of this work include:

- We introduce a new heterogeneous media collection with editorially curated topics targeted for the task of conversational topic exploration.
- We introduce a multi-task benchmark encompassing diverse knowledge-grounded conversational tasks.
- We perform empirical evaluation of state-of-the-art neural systems for all components.
- We introduce a new *background linking explanation* task that generates natural language explanations to connect pieces of content.

We begin by introducing the heterogeneous BBC benchmark corpus used for all the tasks. We then describe the methodology for creating and annotating the data for each task. Next, we present results for state-of-the-art systems and analyze their behavior. We conclude by discussing future areas for extending the benchmark to encompass additional tasks and highlight open areas of research.

2 RELATED WORK

The related work spans multiple different areas of work in chatbots and dialogue systems, conversational search, and entity-based natural language processing (NLP) tasks.

Knowledge-grounded chat There are multiple recently released datasets focused on knowledge-grounded conversations with the aim of facilitating more factual and fluent conversations on a topic. A common approach is to crowdsource conversations in a Wizard-of-Oz setup with crowd workers. The approach by Zhou et al. [2018] focuses on the movie domain and grounds on Wikipedia and movie reviews. Later, Wizard-of-Wikipedia (WoW) [Dinan et al. 2019] focuses on chit-chat discussions of a single topic from a Wikipedia passage. Most turns focus on chat elements of a single topic, so it does not require deep knowledge-grounding. Later work on the Topical-Chat dataset [Gopalakrishnan et al. 2019] extends this paradigm beyond Wikipedia to include multiple sources on a limited number of fifty popular entity topics. Recently, the BEGIN benchmark [Dziri et al. 2021] examines the groundedness of systems trained on WoW and finds that state-of-the-art generative language models (GPT-2 [Radford et al. 2019] and T5 [Raffel et al. 2019]) leverage the document context to a very limited extent. Instead of focusing on generative approaches, COMEX focuses on measuring the ability to retrieve and connect content. More precisely, the BBC collection contains heterogeneous types of media content which are annotated and collected by professional editorial staff rather than crowd workers as in the previously described datasets.

The ongoing Alexa Prize Socialbot competition [Gabriel et al. 2020; Ram et al. 2018] focuses on the goal of social chat for conversations up to twenty minutes. A common strategy is to discuss topics of interest to the user and use mixed-initiative to continuously ask questions to keep the participant engaged. Similarly, we add an element of mixed-initiative with the task of suggesting related content with natural language explanation of their connection. COMEX could be used to measure critical aspects of a typical Alexa system, including knowledge-grounding, retrieval, and related content identification widely used by leading teams [Chi et al. 2021; Gabriel et al. 2020].

Conversational search and question answering COMEX tasks relate strongly to work on conversational search and question answering. The comparison methods build upon the state-of-the-art methods from the TREC Conversational Assistance Track [Dalton et al. 2019]. The `UoG Bot` incorporates conversational generative query rewriting [Lin et al. 2021b] and neural ranking methods for response selection [Lin et al. 2021a]. COMEX adopts a single annotated response Wizard-of-Oz setup, similar to sparse judgments used in the MS MARCO benchmark [Nguyen et al. 2016]. The recent trend towards open retrieval for question answering in OR-QuAC and OR-CoQA [Qu et al. 2021, 2020] demonstrates the importance of full open retrieval. The ranking tasks in COMEX are closely related to the recently released Question Rewriting in Conversational Context (QReCC) dataset [Vakulenko et al. 2021]. However, the primary focus of COMEX is not question answering (QA) or query rewriting but instead enabling exploration and consumption of longer-form pieces of content by leveraging entity relationships.

Retrieval and NLP benchmarks Several of the tasks in COMEX are inspired by the TREC News Track [Soboroff et al. 2018], including the tasks of Wikification, entity salience, and background linking. We provide new data for these tasks on a new and more diverse collection of media content. We also introduce a novel task focusing on natural language explanations of the connections between pieces of content. COMEX is inspired by other multi-task benchmarks in the NLP community, such as KILT [Petroni et al. 2021] and SuperGLUE [Wang et al. 2019]. Similar to these, our benchmark is a collection of language tasks with a unified test collection. In contrast, it includes conversational response retrieval and exploration.

3 ANNOTATED BBC COLLECTION

In this section, we describe the annotated collection of content and topics that are the basis for the multi-task benchmark. This includes a description of the media collection, semantic annotations, and a description of the topic creation process.

3.1 Heterogeneous collection

Table 1 provides statistics on the BBC corpus. The BBC corpus contains a large quantity of News and Sounds (metadata) content. The News content consists of over 180,000 articles published from the beginning of 2017 to October 2020. The News content contains the title and content as well as metadata, including published date, updated date, author, and others. A News content item includes

editorially curated related links to other articles used in COMEX to create related article pairs across the corpus. Figure 2 shows an example document with key metadata fields highlighted.

The Sounds content includes all published digital on-demand audio content up to October 2020 and contains around 600,000 pieces of content. The Sounds content consists of metadata from the BBC on-demand audio, including live radio programs and podcasts. The content includes fields such as title, synopses of varying lengths, published date, and master brand (typically the first broadcast station).

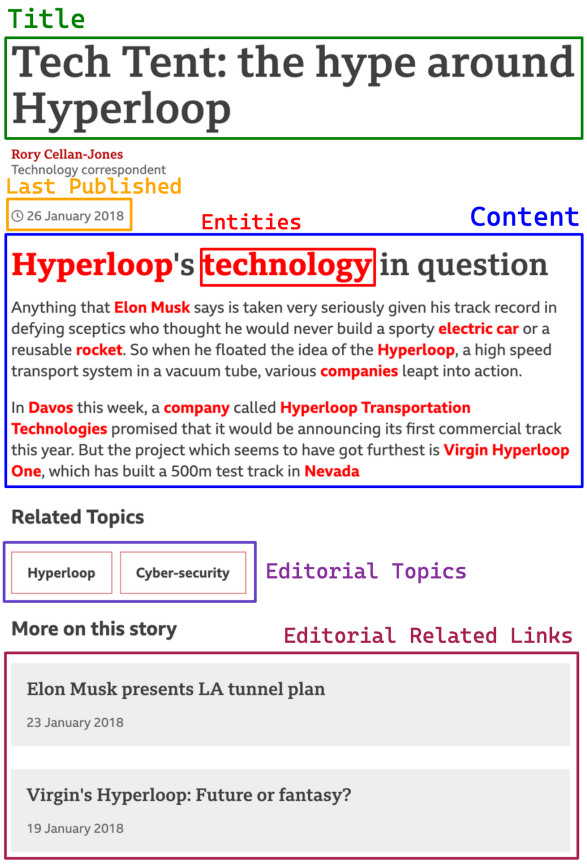

Figure 2: Example of a content item key fields.

3.1.1 *Semantic linking to Wikifipedia.*

COMEX includes automatic semantic annotations to ground the text content to the Wikipedia knowledge base (and subsequently Wikidata). The entity linking is performed using the state-of-the-art Generative ENtity REtrieval (GENRE) [Cao et al. 2021] system from Facebook. GENRE uses a BART-based generative model to auto-regressively generate text tagged with entities. GENRE uses the KILT Wikipedia dump to wikify the text, linking not only named entities, but also abstract concepts. This is vital for conversational exploration because many of the topics of conversation go beyond people, organization, and locations and includes general topical concepts.

Table 1: Corpus statistics

Total number of documents	807,384
News	182,798
Sounds	624,586
Average length in words	
News	437
Sounds	61
Average number of entity mentions	
News	80
Sounds	13

3.1.2 *Knowledge Base and Entity Retrieval.*

COMEX includes a knowledge base to ground the text as well as to use in conversational discovery. It is built on a combination of Wikipedia and Wikidata dumps from December 2020. Both dumps are ingested by the SLING pipeline [Ringgaard et al. 2017] to produce a unified frame representation. Included in the produced frame output is a mapping from Wikipedia to Wikidata.

The entities are extracted for ingest into a search system with the following field structure proven to be effective in fielded entity retrieval [Zhiltsov et al. 2015]:

- **ID:** Wikidata ID
- **URL:** Wikidata URL
- **Name:** Canonical name
- **Category:** Wikipedia categories
- **Related Entities:** Outgoing (object) entity names
- **Attributes:** Text and numeric object values

The structure also supports commonly used entity lookup requests in the UoG Bot, including lookup by name or Wikidata identifier to obtain properties.

3.2 Topics

One of the key aspects of COMEX is that it includes diverse and timely topics curated by professional editorial staff suitable for exploratory conversations.

Topics are editorially selected groupings of content items based upon popularity and key trends on leading social media platforms during 2020 that might appeal to diverse audiences. As shown in Figure 3, the selection criteria include that the topic is noteworthy and interesting as a topic of conversation or that an article on the topic would be appealing to a user who might discover it on social media.

Examples of topics in COMEX include: [Alternatives to plastic], [Love Theatre Day], [Vaccinations], [6 Music T-shirt day], [Google Stadia], [Bitcoin].

After identifying topics on popular social media, one or more related seed content items from the BBC corpus (in the 2020 time period) are manually selected for each topic. After a topic is defined, one or more synthetic user personas are constructed representing a user who might be interested in the topic. We note that the BBC

Corpus includes documents from the previous three years prior to the period of topic generation. This provides a rich selection of background information on the topics.

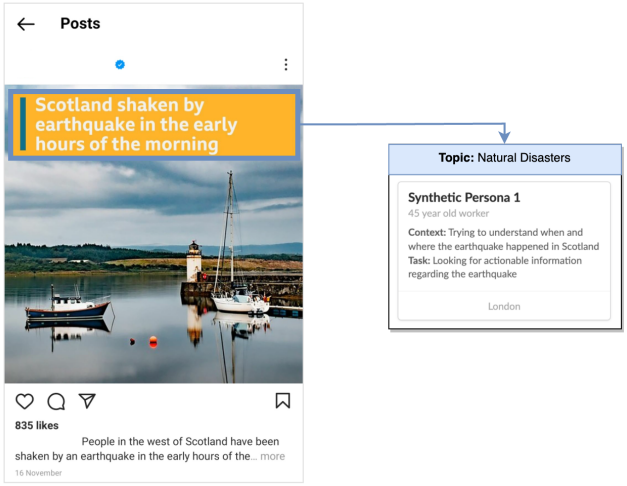

Figure 3: Example of a topic inspired by social media exploration and an associated user persona.

3.3 Synthetic personas

Defining user personas is an important step during the development and design of user-centered systems [Miaskiewicz and Kozar 2011]. More precisely, user personas can be defined as a fictitious representations of users specifically tailored to understand the target audience's behavior for a specific platform [Junior and Filgueiras 2005].

The BBC editorial team created several synthetic user personas as a result of analyzing the BBC target audience. More specifically, user personas are based on statistically representative users that interact with the BBC content and platforms on a daily basis in order to understand and model users' behavior as closely as possible to real-world settings.

Figure 4 shows examples of different personas. Each topic is associated with one or more editorially created personas that would be interested in learning about the topic. Personas are driven by high-level audience context and facts that simulate a user's interest in the topic. Persona properties include location, age, occupation, and interests. Personas also include a fictional narrative to explain the user's potential interest in the story and aspects of it. This allows personas to be used to personalize and target relevant information during data collection. For instance, user queries such as *[What about wildlife sightings near me?]* can be completed from the persona leading to localized and personalized results. As a result, COMEX is one of the few datasets that include a form of (synthetic) personalization. Although not included as a task in COMEX, this could be used to measure systems that perform elicitation to identify important user interests.

The topics and personas dataset includes 57 topics, each containing 1 to 3 different associated personas. The resulting topics and

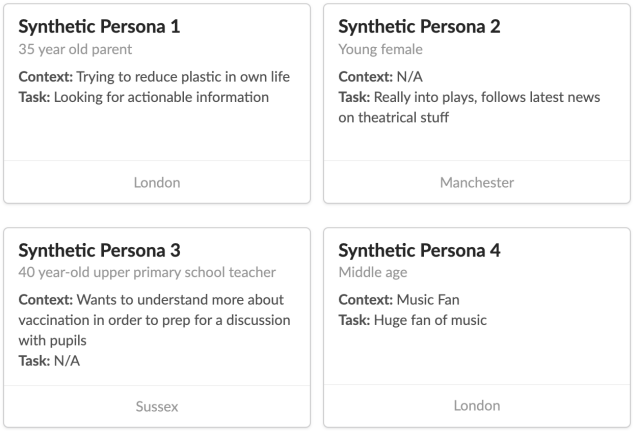

Figure 4: Examples of synthetic personas

personas are then used in the Wizard-of-Oz data collection for the target tasks.

4 BENCHMARK TASKS AND DATA

In this section, we describe the process for creating the benchmark task. We define the tasks formally, including the process for creating and annotating data for the tasks.

4.1 Topic-specific Entity Annotation

Supporting effective knowledge-grounded conversations requires high-quality topical linking of concepts in both the user utterances and the corpora. It requires not only Wikification, but also entity salience information that highlights the importance of the concept in the text. We develop a new topic-focused annotated collection to support evaluation of both entity tasks.

The new annotation collection is created from a subset of the News and Sounds corpus that are professionally annotated with entities linked to Wikidata, including their salience. The choice of content items resulted directly from the topics identified in the topics dataset mentioned in section 3.2. This means that on top of just an entity-linked corpus, the clean entity annotations can be used to identify salient entities for a topic and associated personas.

The annotated data is used to evaluate the effectiveness of state-of-the-art entity grounding systems for documents.

4.1.1 *Annotation Guidelines.*

Annotation is performed by BBC editorial staff using the Prodigy web interface. There are two main tasks within the overall annotation process. In the first step, for a given content item all mentions of entities are highlighted. The interface for selection is shown in Figure 5. To facilitate consistent annotation detailed guidelines on what to annotate and highlight are developed. An abbreviated version of this is outlined below:

What mentions to highlight

- *Proper nouns and named entities:* People, programs, places, organizations...
- *Abstract concepts:* Football, children's home...
- *Other concept:* Mental Health Day, hashtags...

What mentions not to highlight

- *Duplicate mentions of concept previously highlighted with identical text:* If **Mr. White** is tagged, you should not tag Mr. Dan White if they refer to the same person as they are coreferent with minor word variations.
- *Concepts that are overly "general" and unimportant to the article:* "Food", "People", "Animal", etc.

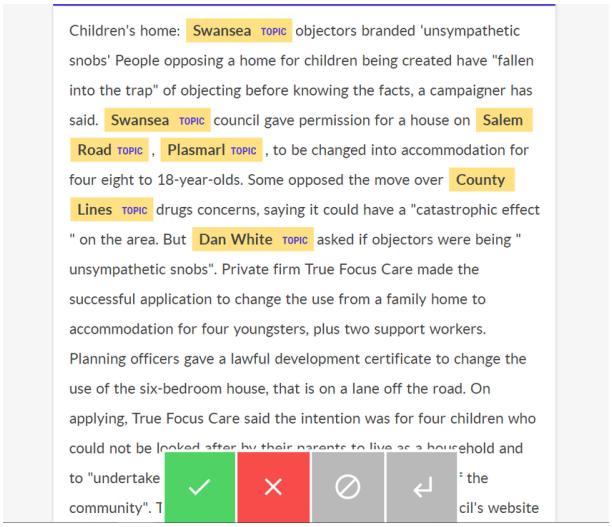

Figure 5: Entity mention tagging UI

The rationale for not tagging duplicate mentions is important because it results in significant entity linking annotation time savings for annotators. These are identified and linked via coreference resolution. Any ambiguous mentions that refer to different entities are tagged separately.

After entity detection, the concepts are Wikified and linked to the Wikidata knowledge base. Using existing entity linking tools, a pool of candidate Wikidata entries is presented in a multiple-choice format for the annotator. There is also an option to add an identifier, not in the candidate pool. Entities can also be tagged as out of KB if they do not have an entry in Wikidata. Out of the 2580 entities tagged, 5% were not in the knowledge base. The UI for linking is shown in Figure 6.

After linking, the annotation process also collected entity salience, the importance of the entity to the document. This is important to be able to discuss the important key concepts in the documents and to identify important entities for a topic. Salience for each entity is annotated on a five-point scale from 0 to 4. The guidelines with examples are included in the Appendix in Table 8. Table 2 shows a breakdown of the resulting entity annotation statistics. This includes information on the number of topics, documents, their type, and a breakdown of links by document type. Table 3 presents the distribution of entity salience.

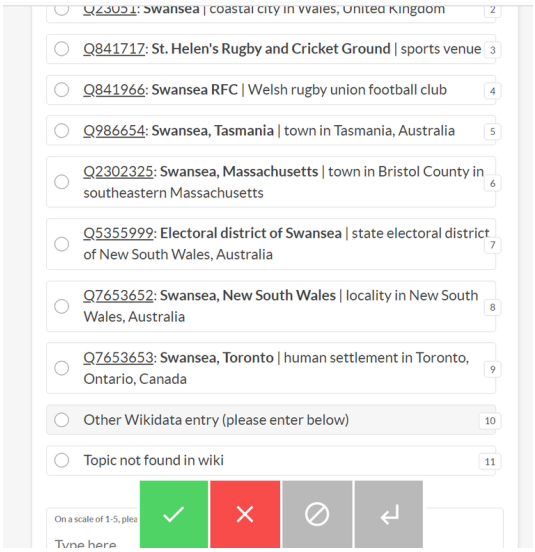

Figure 6: Mention linking to Wikidata UI

Table 2: Entity annotation statistics

	Count
Topics	56
Documents	141
Sounds documents	22
News documents	119
Entity KG links	4403
Sounds KG links	337 (15 avg/doc)
News KG links	4055 (22 avg/doc)

Table 3: Distribution of entity salience

Salience	Count	Percentage
0	558	23.3%
1	401	16.7%
2	618	25.8%
3	428	17.8%
4	394	16.4%

4.2 Wizard Guidelines

Wizard-of-Oz (WoZ) experiments are experiments in which a user interacts with a system that is totally or partially operated by a human ('Wizard') in a controlled manner. The main goal is to explore and collect data regarding the interaction between a user and the system under simulated experimental conditions. We conduct a WoZ experiment internally with BBC editorial staff with the goal of generating, collecting, and evaluating conversations with a conversational media exploration agent. Figure 7 shows an example of such interaction.

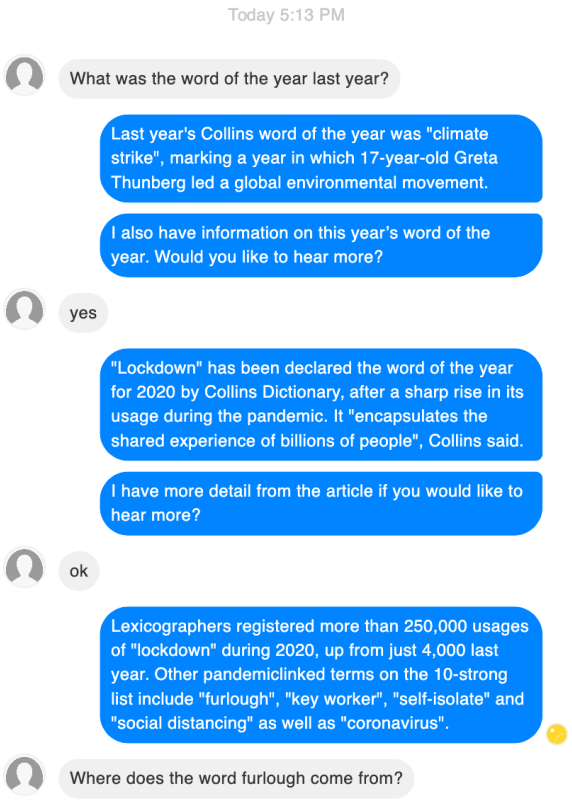

Figure 7: Example of a conversation between a user and a wizard.

4.2.1 *Conversational interface.*

To conduct WoZ experiments, a custom interface based upon the TREC CAsT Web UI [2] is used by staff to perform conversational retrieval over the collection. The Wizard types a manually rewritten version of the user utterance into the search interface to find relevant content. Sample results are shown in the Appendix in Figure 8. The interface returns ranked document results and the associated metadata. The Wizard manually selects the most relevant document and passage to return to the user.

4.2.2 *Experimental details.*

Participants for the WoZ experiment include various members of the BBC staff. Wizards are nominated from the BBC editorial and UI staff. Experiments are conducted asynchronously on the Slack chat platform by pairing users with Wizards on private Slack channels. Users believed Wizards were providing responses exactly as generated by the BBC conversational agent rather than manually selected by a Wizard. The reason for doing so is that the primary goal of the experiment is to observe users participating in a natural conversation that reflects real-world interactions with a fully automatic agent.

Experiments are conducted by providing participants with synthetic personas and associated topics to converse about, prompting them to ask related and relevant questions. The Wizard is given access to a custom interface, based on the CAsT Web UI, with direct access to a search index built on the BBC corpus. The Wizard has access to the BBC annotated corpus in order to look up and retrieve passages upon participant queries. After choosing a relevant passage and returning this to the participant, the Wizard writes a follow-on sentence, which is an utterance recommending another content item. The aim is to use a form of initiative to steer the conversation towards additional relevant content. Throughout the interaction, Wizards record and log conversational metadata, such as the retrieved article ID or URL.

Participant guidelines: Participants are provided with a synthetic user persona (as defined in section 3.3) and a topic (as defined in section 3.2). Based on the persona's interests, they initiate conversation. Throughout the interaction, participants are asked to consider the following questions:

- What sort of things might your persona ask the agent about this topic?
- What sort of conversation might they wish to have with the BBC agent?

Wizard guidelines: Wizards are instructed to extract content from the search interface and keep track of the generated conversational metadata. An example of the conversational metadata stored by the Wizard for a single turn in a conversation is shown in Table 4. A summary of the statistics for the WoZ experiments is presented in Table 5.

Table 5: Wizard-of-Oz experiment statistics

	Count
Total conversations	48
Total turns	261
Average turns per conversation	5.4
Turns with results in corpus	143
Turns without results in corpus	77
Other turns (chit chat etc.)	41

5 EXPERIMENTAL RESULTS

In this section, we provide baselines and initial resources for the tasks in the benchmark.

5.1 Task 1: Topic-focused document entity linking

This task aims to study the effectiveness of document entity linking for knowledge-grounded conversations. We use the annotated dataset described in Section 4.1. We consider two state-of-the-art entity linkers:

- **GENRE:** This is a Wikifier that uses a Transformer-based neural model to tag and link every instance of an entity that has a page in Wikipedia [Cao et al. 2021].
- **Radboud Entity Linker (REL):** REL is a modular open-source toolkit for entity detection and linking [van Hulst et al. 2020]. It uses statistical algorithms to link every occurrence of traditional named entities.

[2]TREC CAsT Web UI available at https://github.com/grill-lab/CAsTSearcher.

Table 4: Wizard-of-Oz single turn conversation example

Conversation Medatata	Value
Topic Title	Wildlife Sightings in Urban Areas
Turn Number	1
User Query	Tell me about those dolphins in Istanbul
Bot Response/Relevant Extract	The Bosphorus in Istanbul, Turkey is normally one of the world's busiest marine routes. Huge tankers, cargo ships and passenger boats criss-cross the straits that cut the city in half 24 hours a day.Now, with a lull in traffic and fishermen staying at home during the city's lockdown, dolphins are swimming and jumping in the waters.
Source	https://www..../news/world-52459487
Type	News article
Comments	
Follow-on Question	Speaking of wildlife and lockdown, do you want to hear about the goats in Llandudno?

To evaluate the entity linkers, we apply them to the corpus and compare their results to the ground-truth labels using standard linking metrics: Macro-precision (Macro-P), Macro-recall (Macro-R), and Macro-F1 (Macro-F1).

The results presented in Table 6 show that the two entity linkers have different strengths. GENRE has a high recall but suffers from low precision as it produces more linked entities per content item. By contrast, REL has lower recall but higher precision and F1 score. We include in the annotated corpus the entity links from GENRE because it includes general concepts as well as named entities. A high recall system is important for identifying conversational topics.

Table 6: Entity linking results

	Macro-P	Macro-R	Macro-F1
GENRE	0.259	**0.688**	0.330
REL	**0.525**	0.459	**0.445**

5.2 Task 2: Conversational Passage Ranking

In order to engage the user and facilitate learning and exploration, the core of the agent focuses on the ability to retrieve relevant content. This requires a conversational ranking task that handles multi-turn information seeking and is able to track references to entities and topics in a natural way. The focus of this task is long answer responses. These may be shortened or summarized, which is a task we leave for future work.

The baseline setup for this task uses a proven system and pipeline used as a baseline for TREC CAsT [Dalton et al. 2019]. The entity annotated corpus is indexed with Pyserini, a Python-based information retrieval toolkit based on Anserini [Yang et al. 2018]. The multi-turn query understanding is handled by a conversational query rewriter based on a T5 sequence to sequence model that uses previous conversation context from the user and system turns [Lin et al. 2021b].

The baseline system is a multi-stage retrieval that uses the rewritten query for retrieval. First, we perform retrieval over the *documents* using BM25 [Robertson and Walker 1994]. After document retrieval, we produce a ranking of the top passages from each. We also experiment with the use of a MonoT5 passage re-ranker [Nogueira et al. 2020], which is a widely used baseline. The MonoT5 model used is fine-tuned on the standard MS MARCO web data collection. For the baselines, we use the default retrieval parameters provided by the toolkit. Passages are split into non-overlapping fixed-size passages.

5.2.1 *Results.*

To focus on the conversational retrieval task, the evaluation uses retrieval measures that focus heavily on precision in top ranks as well as standard retrieval measures. These include: Precision at 1 (P@1), Recall at 5 (R@5), Mean Reciprocal Rank (MRR), and Mean Average Precision (MAP).

Table 7: Passage ranking results

	P@1	R@5	MRR	MAP
BM25	0.271	0.507	0.360	0.360
BM25 + T5	**0.300**	**0.521**	**0.389**	**0.389**

Table 7 shows the results of the baseline ranking. It shows that using the second stage reranker improves over the BM25 baseline. The results show that the golden passages are in the top five approximately half the time and ranked first approximately a third of the time. We note that retrieval effectiveness is based on the sparse judgments available from the WoZ experiments. Additional assessments and pooling would be needed to robustly evaluate alternative retrieval methods.

5.3 Task 3: Follow-on Content Ranking

The third task in the benchmark focuses on ranking content for suggesting related information to the user during the course of the conversation.This is a form of system initiative where related information is posed to the user to recommend the next step in the conversational trajectory. The suggested related information should be both topically coherent in the conversational trajectory as well as engaging. In the next section we discussed multiple methods for generating follow-on candidates.

5.3.1 *Models*.

In this section, we describe several baseline ranking methods for ranking related content. We study methods that use both explicit structure as well as structure from shared entity relationships.

Intra-article - Given the short system responses, this ranking model uses the discourse structure and recommends additional content from the same article.

Editorial Related Links - The article contains editorially created links to other pieces of content. Since these are selected manually, they form a natural follow-on. However, although they may be relevant to the article overall, they may not follow the topic of the conversation.

Entity-based - We propose a new model based upon using topically entities. All of the content in the BBC corpus is tagged automatically with entities. This method ranks entities to find ones that are the most relevant.

We consider the following features:

- Frequency of the entity.
- Position of the entity and its existence in the title.
- Cosine similarity between the SBERT [Reimers and Gurevych 2019] embedded description of the entity, inferred from its Wikipedia description, and
  - The embedded form of the previous response passage.
  - The embedded form of the previous user utterance.
- Type of the feature, boosting the score of entities of type Person, Organization, Location.

The features are used in a custom learning-to-rank (LTR) function to rank the entities based on relevance. To find related content, the top-ranked entities are used as a query, and documents containing links to them are retrieved from the collection.

We do not explicitly provide an evaluation of these methods in this work. For future evaluation, the WoZ data provides wizard-generated follow-on descriptions that include content links for comparison.

5.4 Task 4: Background Linking Explanation

As shown in Figure 1, the main objective of the fourth task is to produce a natural language connection from the current conversational context to the suggested candidate content from generated in Task 3.

5.4.1 *Template-based approach*.

A simple approach is to use a template-based method given the source and target articles. We say that X is an entity/topic in the article ,and Y is the title of the related article. The following are basic templates:

- "If you are interested in this, would you be interested in Y?"
- "Would you like to hear more about this?"
- "Speaking of X, would you like to hear more about Y?"

However, these simplistic connections only connect the pieces of content in a shallow and simplistic way.

5.4.2 *Supervised Generative Model*.

To create more meaningful semantic connection between pieces of content, we create a new dataset explicitly for this task. The input to the task is a pair of of content items from the WoZ data. The output is a summary of the connection between them. BBC editorial staff write and label the generated output following a variety of different formats, such as "Speaking of {insert entity/phrase}, {insert how they are related}. Would you like to hear more?". We produce a supervised dataset containing 377 pairs of connecting articles.

Given that existing models demonstrate transfer learning capability, we experiment with a model-based upon automatic text generation. As a baseline, we use a T5-base model. Due to the token limit constraints we first generate a summary of the body of the articles using the HuggingFace summarizer pipeline [Wolf et al. 2019]. We introduce special tokens encoding the data for T5. We also pre-process the input data to generate clean text by removing punctuation, including underscores, dashes, and semicolons. The input encoding to T5 has the following structure "follow: <firsttitle> article A title <firstbody> summarised article A <secondtitle> article B title <secondbody> summarised article B".

Although we do not provide an evaluation of the model in this work, the labeled data in the resource can be used to evaluate the task using standard text generation evaluation measures.

6 CONCLUSION AND FUTURE WORK

We present COMEX, a new multi-task benchmark for conversational exploration of heterogeneous media content. We introduce a new semantically annotated collection of BBC News and Sounds data. We develop these datasets in conjunction with professional BBC editorial staff. The tasks we propose encompass key components of a conversational exploration agent including: knowledge-grounding of documents and interactions, conversation response ranking, conversational background linking, and background linking explanation. We perform an evaluation of some of the key task components using state-of-the-art approaches and neural models. Although limited in scale, the entity link and Wizard-of-Oz data is labeled by expert professional staff. The resource provides researchers an end-to-end dataset that spans the spectrum of tasks for a knowledge-grounded social conversational agent.

ACKNOWLEDGMENTS

This work is supported by a grant from The Data Lab in partnership with the BBC. Additionally, this work is supported by a Turing AI Acceleration Fellowship from the Engineering and Physical Sciences Research Council, grant number EP/V025708/1.

The authors would like to thank the BBC technical and editorial staff who contributed to BBC UoGBot and creation of this resource. This includes: Hazel Morton, James Fletcher, Fiona Linton Forrest, Claire Dimeo, Sandy Cormie, Matthew Gamble, James Barndard, Matthew Clarke, Leo Currie, Andy Pick, Ren Padron, and Lee McWhinnie. We would also like to thank Daniel Whaley, Diane McDonald, and Chris Dix for their guidance and support.

REFERENCES

Daniel De Freitas Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a Human-like Open-Domain Chatbot. *ArXiv* abs/2001.09977 (2020).

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive Entity Retrieval. arXiv:2010.00904 [cs.CL]

Ethan A. Chi, Chetanya Rastogi, Alexander Iyabor, Hari Sowrirajan, Avanika Narayan, and Ashwin Paranjape. 2021. Neural, Neural Everywhere: Controlled Generation Meets Scaffolded, Structured Dialogue.

Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2019. TREC CAsT 2019: The Conversational Assistance Track Overview. (2019).

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-Powered Conversational agents. *ArXiv* abs/1811.01241 (2019).

Nouha Dziri, Hannah Rashkin, Tal Linzen, and D. Reitter. 2021. Evaluating Groundedness in Dialogue Systems: The BEGIN Benchmark. *ArXiv* abs/2105.00071 (2021).

Raefer Gabriel, Yang Liu, Anna Gottardi, Mihail Eric, Anju Khatri, Anjali Chadha, Qinlang Chen, Behnam Hedayatnia, Pankaj Rajan, Ali Binici, Shui Hu, Karthik Gopalakrishnan, Seokhwan Kim, Lauren Stubel, Kate Bland, Arindam Mandal, and Dilek Z. Hakkani-Tür. 2020. Further Advances in Open Domain Dialog Systems in the Third Alexa Prize Socialbot Grand Challenge.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Z. Hakkani-Tür. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *INTERSPEECH*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of Hallucination in Natural Language Generation. *CoRR* abs/2202.03629 (2022). arXiv:2202.03629 https://arxiv.org/abs/2202.03629

R. Jones, Ben Carteree, Ann Clion, Maria Eskevich, G. Jones, Jussi Karlgren, Aasish Pappu, Sravana Reddy, and Yongze Yu. 2020. TREC 2020 Podcasts Track Overview. *ArXiv* abs/2103.15953 (2020).

Plinio Thomaz Aquino Junior and Lucia Vilela Leite Filgueiras. 2005. User Modeling with Personas. In *Proceedings of the 2005 Latin American Conference on Human-Computer Interaction* (Cuernavaca, Mexico) *(CLIHC '05)*. Association for Computing Machinery, New York, NY, USA, 277–282. https://doi.org/10.1145/1111360.1111388

Jimmy J. Lin, Rodrigo Nogueira, and Andrew Yates. 2021a. Pretrained Transformers for Text Ranking: BERT and Beyond. *Proceedings of the 14th ACM International Conference on Web Search and Data Mining* (2021).

Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and Jimmy Lin. 2021b. Multi-stage conversational passage retrieval: An approach to fusing term importance estimation and neural query rewriting. *ACM Transactions on Information Systems (TOIS)* 39, 4 (2021), 1–29.

Tomasz Miaskiewicz and Kenneth A. Kozar. 2011. Personas and user-centered design: How can personas benefit product design processes? *Design Studies* 32, 5 (2011), 417–430. https://doi.org/10.1016/j.destud.2011.03.003

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *CoRR* abs/1611.09268 (2016). arXiv:1611.09268 http://arxiv.org/abs/1611.09268

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document Ranking with a Pretrained Sequence-to-Sequence Model. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. 708–718.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vassilis Plachouras, Tim Rocktaschel, and Sebastian Riedel. 2021. KILT: a Benchmark for Knowledge Intensive Language Tasks. *ArXiv* abs/2009.02252 (2021).

Chen Qu, Liu Yang, Cen Chen, William Bruce Croft, Kalpesh Krishna, and Mohit Iyyer. 2021. Weakly-Supervised Open-Retrieval Conversational Question Answering. *ECIR European Conference on Information Retrieval*.

Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, William Bruce Croft, and Mohit Iyyer. 2020. Open-Retrieval Conversational Question Answering. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2020).

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *CoRR* abs/1910.10683 (2019). arXiv:1910.10683 http://arxiv.org/abs/1910.10683

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, Eric King, Kate Bland, Amanda Wartick, Yi Pan, Han Song, Sk Jayadevan, Gene Hwang, and Art Pettigrue. 2018. Conversational AI: The Science Behind the Alexa Prize. *ArXiv* abs/1801.03604 (2018).

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *CoRR* abs/1908.10084 (2019). arXiv:1908.10084 http://arxiv.org/abs/1908.10084

Michael Ringgaard, Rahul Gupta, and Fernando C Pereira. 2017. SLING: A framework for frame semantic parsing. *ArXiv* abs/1710.07032 (2017).

Stephen E Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. *SIGIR'94 Conference on Research and Development in Information Retrieval*, 232–241.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric Michael Smith, Y.-Lan Boureau, and Jason Weston. 2021. Recipes for Building an Open-Domain Chatbot. *EACL Conference of the European Chapter of the Association for Computational Linguistics*.

Ian Soboroff, Shudong Huang, and Donna K. Harman. 2018. TREC 2018 News Track Overview. *TREC Text Retrieval Conference*.

Svitlana Vakulenko, S. Longpre, Zhucheng Tu, and R. Anantha. 2021. Question Rewriting for Conversational Question Answering. *Proceedings of the 14th ACM International Conference on Web Search and Data Mining* (2021).

Johannes M. van Hulst, Faegheh Hasibi, Koen Dercksen, Krisztian Balog, and Arjen P. de Vries. 2020. REL: An Entity Linker Standing on the Shoulders of Giants. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2020).

Denny Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM* 57 (2014), 78–85.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In *NeurIPS*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *CoRR* abs/1910.03771 (2019). arXiv:1910.03771 http://arxiv.org/abs/1910.03771

Peilin Yang, Hui Fang, and Jimmy Lin. 2018. Anserini: Reproducible Ranking Baselines Using Lucene. *J. Data and Information Quality* 10, 4, Article 16 (oct 2018), 20 pages. https://doi.org/10.1145/3239571

Nikita Zhiltsov, Alexander Kotov, and Fedor Nikolaev. 2015. Fielded Sequential Dependence Model for Ad-Hoc Entity Retrieval in the Web of Data. *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2015).

Kangyan Zhou, Shrimai Prabhumoye, and Alan W. Black. 2018. A Dataset for Document Grounded Conversations. *EMNLP Conference on Empirical Methods in Natural Language Processing*.

Figure 8: Wizard-of-Oz search interface

Table 8: Guidelines for entity salience annotation

Rating	Description	Example
4	The target topic is the main topic of the document and topic and has the right level of specificity.	For an article about what's happening in the West End of Glasgow, the entities West End and Glasgow would have this rating.
3	The highlighted content is important to the document and topic but is not the main topic of the article and is not specific enough.	The same article from above, mentions such as Scotland, Byres Road (if it's mentioned and is one of the main highlights of the article) or the name of the local council overseeing would fall under this rating.
2	The highlighted content is somewhat relevant to the article, but is rather general.	The same article from above, activities going on in the West End such as parade, festivals would fall under this rating.
1	The highlighted content is borderline irrelevant and is highly general.	Using the same article, if there is a comparison being made to other UK cities, such cities would be listed under here. If there is some organization being mentioned but is barely relevant to the context, it would also be listed here.
0	The highlighted content is completely irrelevant.	If by any chance there were a few sentences containing mentions of other countries or randomly mentioning other entities such as "Mercedes" without much context, those would fall under this rating.