# AltEntities

**This doc:** GitHub Repo
**Dataset:** GitHub Repo
**Data Card Authors:** Javad Hosseini, Filip Radlinski

DATASET SUMMARY

AltEntities (Alternative-Entities) is a dataset of English-language alternative questions with their answers. In these questions, a user is presented with a choice of two entities. The user answer intends to select one entity from the pair by referring to it in an indirect way, i.e., without referring to its name or position. For example:

Question: “Do you mean 'Nemesis (Davis novel)' or 'Nemesis (Christie novel)'?”

Answer: “Not the more modern one, the one closer to BC than 2021.”

The entity names are sampled from Wikipedia. The alternative questions are generated using a simple template. The referring expressions are provided by human raters. The dataset contains 6,247 alternative questions in three domains (books, music, and recipes) and 42,529 referring expressions. The dataset is suitable for building machine learning systems to resolve indirect referring expressions for entity selection.

## Authorship

### Dataset Owners

| TEAM(S) | CONTACT DETAIL(S) | AUTHOR(S) |
|---|---|---|
| | **Dataset Owner(s):** Mohammad Javad Hosseini, Filip Radlinski, Silvia Pareti, Annie Louis<br><br>**Affiliation:** Google<br><br>**Contact:** javadh@google.com | • Javad Hosseini <javadh@google.com><br>• Filip Radlinski <filiprad@google.com><br>• Annie Louis <annielouis@google.com><br>• Silvia Pareti <spareti@google.com> |

## Dataset Overview

### DATA SUBJECT(S)

Sensitive Data about people

**Non-Sensitive Data about people**

Data about natural phenomena

**Data about places and objects**

Synthetically generated data

Data about systems or products and their behaviors

Unknown

Others (Please Specify)

### DATASET SNAPSHOT

| | |
|---|---|
| Size of dataset | 59 MB |
| Number of instances (referring expressions) | 42,529 |
| Number of questions | 6,247 |
| Number of Fields | 8 (domain, question, target, target_index, sampling_method, description_section, choices, expressions) |

### CONTENT DESCRIPTION

Each alternative question in the dataset consists of:

1. Information about the question: The question text ("Do you mean *A* or *B*", where *A* and *B* are entities); the target entity (selected at random between A and B) and its position; the question domain; the sampling method used to select the two entities (e.g., entities with a similar description in Wikipedia or entities with similar names); the Wikipedia section from which the entity description shown to raters was taken (only relevant for books and recipes domains).

2. Information about the two choices (entities): For each choice the dataset contains the entity name; its Wikipedia item identifiers; various textual descriptions (taken from Wikipedia verbatim, then truncated for length); Wikipedia image url for recipe entities.

3. Annotations (referring expressions): free-form text written by paid crowdworkers tasked with providing an indirect expression referring to just the target entity.

## Version And Maintenance

### MAINTENANCE STATUS

Regularly Updated

New versions of the dataset have been or will continue to be made available.

Actively Maintained

No new versions will be made available, but this dataset will be actively maintained, including but not limited to updates to the data.

**Limited Maintenance**

**The data will not be updated, but any technical issues will be addressed.**

Deprecated

This dataset is obsolete or is no longer being maintained.

### DATASET VERSION

**Current Version:** 1.0

**Last Updated:** 12/2022

**Release Date:** 01/2023

### MAINTENANCE PLAN

Any major issues reported will be fixed on a best effort basis.

| | NEXT PLANNED UPDATE(S) | EXPECTED CHANGE(S) |
|---|---|---|
| | N/A | N/A |

## Example Of Data Points

| PRIMARY DATA MODALITY | LINK(S) TO DATA POINT(S) | DATA FIELDS |
|---|---|---|
| Image Data<br>**Text Data**<br>Tabular Data<br>Audio Data<br>Video Data<br>Time Series<br>Graph Data<br>Geospatial Data<br>Multimodal<br>Please specify<br>Unknown<br>Others (Please Specify) | • [Full data]<br>• [Data in all domain] | |

| Field Name | Example Value | Description |
|---|---|---|
| domain | BOOKS | The domain of the example: “BOOKS”, “RECIPES”, or “SONGS” |
| question | Do you mean 'The Ghost and the Goth' or 'Shadowland'? | An alternative question providing two choices. |
| target | http://en.wikipedia.org/wiki/Shadowland_(Cabot_novel) | The URL of the target entity |
| target_index | 1 | The index (position) of the target entity in the question (0 or 1) |
| sampling_method | SIMILAR_DESCRIPTION | The sampling method used to sample the entities in the question (e.g., SAME_NAME, SIMILAR_NAME, or SIMILAR_DESCRIPTION) |
| description_section | plot summary | The title of the Wikipedia section shown to crowd workers. If empty, |

| | | |
|---|---|---|
| | | indicates the first section. |
| choices | [choice1, choice2] | Information about the two choices. See below fields for more information |
| choice/name | The Ghost and the Goth | The name of the choice |
| choice/wikipedia_url | http://en.wikipedia.org/wiki/The_Ghost_and_the_Goth | The Wikipedia URL of the choice |
| choice/description | Alona Dare was the most popular girl in school... | The choice description shown to crowd workers |
| choice/infobox | author: Stacey Kade. country: United States... | Textual representation of the choice's Wikipedia infobox |
| choice/unshown_background | ...The Ghost and the Goth is a 2010 paranormal romance... | "choice/infobox" above concatenated with Wikipedia sections that were *not* shown to crowd workers |
| expressions | ["Not the one with Will.", ...] | Referring expressions written by crowd workers to select the target entity |

EXAMPLE: TYPICAL DATA POINT

```
``` {
        "domain": "BOOKS",
        "question": "Do you mean 'The Ghost and the
Goth' or 'Shadowland'?",
        "target":
"http://en.wikipedia.org/wiki/Shadowland_(Cabot_nove
l)",
        "target_index": 1,
        "sampling_method": "SIMILAR_DESCRIPTION",
        "description_section": "plot summary",
```

EXAMPLE: ATYPICAL DATA POINT

In the below example, two of the expressions (in bold) are correct for both entities, but they have to be correct only for the target entity.

```
```{
        "domain": "BOOKS",
        "question": "Do you mean 'Scarlet' or
'Cress'?",
        "target":
"http://en.wikipedia.org/wiki/Scarlet_(novel)",
        "target_index": 0,
        "sampling_method": "SIMILAR_DESCRIPTION",
```

```
        "choices": [
            {
                "name": "The Ghost and the Goth",
                "wikipedia_url":
"http://en.wikipedia.org/wiki/The_Ghost_and_the_Goth
",
                "description": "Alona Dare was the
most popular girl in school…",
                "infobox": ": Infobox book. author:
Stacey Kade. country: United States. followed_by:
Queen of The Dead…",
                "unshown_background": ": Infobox
book. author: Stacey Kade. country: United States.
followed_by: Queen of The Dead… The Ghost and the
Goth is a 2010 paranormal romance young adult novel
written by Stacey Kade and published by Hyperion
Books…"
            },
            {
                "name": "Shadowland",
                "wikipedia_url":
"http://en.wikipedia.org/wiki/Shadowland_(Cabot_nove
l)",
                "description": "Sixteen-year-old
Susannah 'Suze' Simon is a mediator, which means she
can see and talk to ghosts…",
                "infobox": ": Infobox book. author:
Meg Cabot. caption: First edition. country: United
States…",
                "unshown_background": ": Infobox
book. author: Meg Cabot. caption: First edition.
country: United States… Shadowland is a young adult
novel written by author Meg Cabot and published by
Avon Books in 2000…"
            }
        ],
        "expressions": [
            "Not the one with Will.",
            "The one with Suze.",
            "The one where Suze helps ghosts cross
over.",
            "about 16 year old",
            "talks to ghosts",
            "helps them to afterlife"
        ]
    } ```
```

```
        "description_section": "",
        "choices": [
            {
                "name": "Scarlet",
                "wikipedia_url":
"http://en.wikipedia.org/wiki/Scarlet_(novel)",
                "description": "Scarlet is a 2013
young adult science fiction novel written by
American author Marissa Meyer and published by
Macmillan Publishers through their subsidiary Feiwel
& Friends. It is the second novel in The Lunar
Chronicles…",
                "infobox": ": Infobox book. author:
Marissa Meyer. caption: Book cover of Scarlet.
congress: PZ7.M571737 Sc 2013. country: United
States. cover_artist: Michael O. followed_by: Cress.
genre: Children, Romance, Science Fiction,
dystopian. …",
                "unshown_background": ": Infobox
book. author: Marissa Meyer. caption: Book cover of
Scarlet. congress: PZ7.M571737 Sc 2013. country:
United States. cover_artist: Michael O. followed_by:
Cress. genre: Children, Romance, Science Fiction,
dystopian… Scarlet Benoit is a young girl living on
her grandmother's farm in Rieux, France…
            },
            {
                "name": "Cress",
                "wikipedia_url":
"http://en.wikipedia.org/wiki/Cress_(novel)",
                "description": "Cress is a 2014
young adult science fiction novel written by
American author Marissa Meyer and published by
Macmillan Publishers through their subsidiary Feiwel
& Friends…",
                "infobox": ": Infobox book. author:
Marissa Meyer. caption: Book cover of Cress.
congress: PZ7.M571737 Cre 2014b. country: United
States. cover_artist: Michael O. followed_by:
Winter. genre: Young adult, Romance, Science
fiction, Dystopian…",
                "unshown_background": ": Infobox
book. author: Marissa Meyer. caption: Book cover of
Cress. congress: PZ7.M571737 Cre 2014b. country:
United States. cover_artist: Michael O. followed_by:
Winter. genre: Young adult, Romance, Science
fiction, Dystopian… The Lunar Chronicles. The novel
begins with an introduction to Crescent \"Cress\"
Moon Darnel, a sixteen-year-old girl…"
            }
        ],
        "expressions": [
            "its apart of the lunar chronicles ",
            "its a young adult sci fi novel",
            "its the second book in the series i
believe ",
            "The oldest one",
```

```
            "The second of its series",
            "Not the one based on Rapunzel"
        ]
    }```
```

## Provenance

### Data Collection & Sources

| METHOD(S) USED | METHODOLOGY DETAIL(S) | SOURCE DESCRIPTION(S) |
|---|---|---|
| API<br>Artificially Generated<br>**Crowdsourced - Paid**<br>Crowdsourced - Volunteer<br>Vendor Collection Efforts<br>**Scraped or Crawled**<br>Survey, forms or polls<br>Taken from other existing datasets<br>Unknown<br>To be determined<br>Others (Please Specify) | **Source :** Contributed by paid crowd workers.<br>**Platform:** Proprietary crowdworking platform<br>**Is this source considered sensitive or high-risk?** No<br>**Dates of Collection:** August 2021 - February 2022<br>**Primary modality of collected data:**<br>• Image Data<br>• **Text Data**<br>• Tabular Data<br>• Audio Data<br>• Video Data<br>• Time Series<br>• Graph Data<br>• Geospatial Data<br>• Unknown<br>• Multimodal (Please specify)<br>• Others (Please specify)<br>**Update Frequency for collected data:**<br>• Yearly / Quarterly / Monthly / Weekly / Daily / Hourly<br>• **Static**<br>• Others (Please specify) | • **[Source]:** English Wikipedia to obtain entities and their descriptions.<br>• **[Source]:** Crowd annotators to obtain referring expressions. |

| COLLECTION CADENCE | DATA INTEGRATION | DATA PROCESSING |
|---|---|---|
| **Static**<br>**Data was collected once from single or multiple sources.**<br><br>Dynamic | None | N/A |

| | | |
|---|---|---|
| Data is updated regularly from single or multiple sources.<br><br>Streamed<br>Data is streamed from single or multiple sources.<br><br>Others<br>(please specify) | | |

## Collection Criteria

| DATA SELECTION | DATA INCLUSION | DATA EXCLUSION |
|---|---|---|
| • Questions were created from three domains: Books, Recipes, and Music.<br>• A set of candidate entities were extracted from English Wikipedia for each domain and background section.<br>• The entities `A' and `B' were filled by sampling pairs of entities based on different criteria. | • English Wikipedia articles were collected by checking the presence of domain-specific Wikipedia templates (Wikipedia infoboxes).<br>• For the Recipes domain, articles that have a section with title “ingredients” were additionally included.<br>• Questions were created by sampling entity pairs based on different criteria, e.g., entities with the same or similar name, similar description, etc. See the research paper for details. | The following articles were filtered:<br><br>• Very short articles (with fewer than 250 characters in the main section in the Songs domain, or the section shown to the raters in the other domains).<br>• Entities that could be assigned to more than one domain.<br>• Books or Music entities without a genre.<br>• In the Music domain, article length (number of sections/subsections) is used as a proxy for popularity. Songs other than the first ~1000 songs with longest articles were filtered out.<br>• To remove any sensitive or offensive content, articles whose content matched a list of sensitive words on the date of sampling were filtered out. The filters were intended to avoid potentially biased or insensitive language on the part of crowd workers, but note that the filters may have also excluded some books and songs for which Wikipedia articles describe adult themes, even if they are not the topic of insensitive or biased language. |

## Data Sampling

| METHOD(S) USED | CHARACTERISTIC(S) | SAMPLING CRITERIA |
|---|---|---|

| | | |
|---|---|---|
| Cluster Sampling<br>Haphazard Sampling<br>Multi-stage Sampling<br>**Random Sampling**<br>Retrospective Sampling<br>Stratified Sampling<br>Systematic Sampling<br>Weighted Sampling<br>Unknown<br>Unsampled<br>Others: **Sampling entities with the same name, similar title, similar descriptions and similar attributes.** | See the research paper for details. | See the research paper for the details. |