Datasheet for the REFreSD dataset

- ✓ **Dataset name:** **R**ationalized-**E**nglish **Fre**nch **S**emantic **D**ivergences
- ✓ **Citation:** Detecting Fine-Grained Cross-Lingual Semantic Divergences without Supervision by Learning to Rank, *In EMNLP 2020*
- ✓ **Datasheet developers:** Eleftheria Briakou & Marine Carpuat
- ✓ **Datasheets author:** Eleftheria Briakou
- ✓ **Link to dataset:** https://github.com/Elbria/xling-SemDiv/tree/master/REFreSD

A. Motivation

1. For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created to evaluate the performance of computational approaches that aim to capture fine-grained meaning mismatches between English and French sentence-pairs drawn from parallel corpora.

2. Who created the dataset (e.g., which team, research group) & on behalf of which entity (e.g., company, institution, organization)?

The dataset was created by members of the Computational Linguistics and Infomation Processing (CLIP) lab at the University of Maryland, College Park.

3. Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

The creation of the dataset is funded from a National Science Foundation grant under Award Number No. 1750695. The Principal Invistigator of the grant is Marine Carpuat.

B. Composition

1. What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances of the dataset consist of parallel text in English and French.

2. How many instances are there in total (of each type, if appropriate)?

There are 1,039 annotated sentence-pairs.

3. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

*The annotated instances are sampled from the English-French part of the **WikiMatrix** corpus. We choose this resource because (1) it is likely to contain diverse, interesting divergence types since it consists of mined*

parallel sentences of diverse topics which are not necessarily generated by (human) translations, and (2) Wikipedia and WikiMatrix are widely used resources to train semantic representations and perform cross-lingual transfer in NLP. We exclude obviously noisy samples by filtering out sentence-pairs that a) are too short or too long, b) consist mostly of numbers, c) have a small token-level edit difference.

4. What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance consists of an English-French sentence-pair in the form of unprocessed text.

5. Is there a label or target associated with each instance? If so, please provide a description.

Each instance is associated with a single label at a sentence-level describing the semantic relation of the individual sentences and rationales in the form of highlighted spans.

6. Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

We do not provide information on the language variety/varieties of the annotated instances or the origin of the specific Wikipedia articles the sentences are sampled from, as this information is not available in WikiMatrix.

7. Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links) If so, please describe how these relationships are made explicit.

N/A

8. Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

*The dataset was primarily collected to serve as an **evaluation dataset**. Therefore, there are no splits associated with it.*

9. Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

None that we are aware of.

10. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained.

11. Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

The dataset builds upon public data and therefore, does not contain information that might be considered confidential.

12. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

None that we are aware of. The dataset covers Wikipedia topics.

13. Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Not as a whole. Individual sentence-pairs are likely to refer to people as discussed in Wikipedia pages.

14. Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

Not as a whole. Individual sentence-pairs are likely to refer to peope as discussed in Wikipedia pages.

15. Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

Individual sentence-pairs are likely to refer to people as discussed in Wikipedia pages, yet no sensitive information is included in them.

16. Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?

We do not suspect REFreSD contains text that reveals sensitive data.

(P.S.: As the dataset contains sentence-pairs drawn from Wikipedia, it is possible that it contains topics that might be considered sensitive in certain ways. For example, we have not looked at whether/how Wikipedia gender and racial bias is introduced to REFreSD through the sampling process.)

C. Collection process

1. How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Participants were given raw instances.

2. What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The raw sentence-pairs are sampled from the WikiMatrix corpus that consists of bitext mined from Wikipedia pages. Participants are then asked to annotate the raw instances through a web-based server that was configured using the BRAT annotation toolkit for the task at hand.

3. If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

Instances are sampled randomly.

4. Who was involved in the data collection process (e.g., students, crowdworkers, contractors) & how were they compensated (e.g., how much were crowdworkers paid)?

The population targeted in this project is UMD students who are proficient in English and French. Annotators were compensated with Amazon gift cards at a rate of $2 per 10 instances, with a bonus of $5 for completing the first session of 120, and $10 for completing additional sessions after the first. The maximum amount of money a participant can get for completing all 8 sessions is $267.

5. Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the time- frame in which the data associated with the instances was created.

The dataset was collected on April 2020, over the span of three weeks.

6. Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

The collection of the dataset passed through an Expedited Review that was conducted by the Institutional Review Board at Univesity of Maryland, College Park. The process was approved at March 27, 2020.

7. Does the dataset relate to people? If not, you may skip the remainder of the questions in this section.

Yes.

8. Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

The data (here annotations of instances) are collected from the individuals directly.

9. Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

Participants are notified avout the collection of data through advertising emails; the exact wording used can be found here.

10. Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

*All participants were presented with **informed consent** and asked to sign a consent form **electronically** before the start of the annotation process. Consent forms were written in English. All participants received a copy of the consent form for their records. Since the session took place remotely, participants signed the consent form from their homes and returned it to the investigators via email in a pdf format. The exact wording used in the consent forms can be found here.*

11. If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

Participants were not provided with a mechanism to revoke their consent in future.

12. Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

No. There are no known risks associated to participants of this study.

D.Preprocessing/cleaning/labeling

1. Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of miss- ing values)?If so, please provide a description. If not, you may skip the remainder of the questions in this section.

*Yes. Examples are drawn from the English-French section of the **WikiMatrix***

corpus. We choose this resource because (1) it is likely to contain diverse, interesting divergence types since it consists of mined parallel sentences of diverse topics which are not necessarily generated by (human) translations, and (2) Wikipedia and WikiMatrix are widely used resources to train semantic representations and perform cross-lingual transfer in NLP. We exclude obviously noisy samples by filtering out sentence-pairs that a) are too short or too long, b) consist mostly of numbers, c) have a small token-level edit difference.

2. Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?If so, please provide a link or other access point to the "raw" data.

Yes, the raw text is publicly available here.

3. Is the software used to preprocess, clean, label the instances available? If so, please provide a link or other access point.

The cleaning/preprocessing software can be found here. The labeling software can be found here.

E. Uses

1. Has the dataset been used for any tasks already? If so, please provide a description.

Yes, the dataset has been used to evaluate computational approaches in unsupervised detection semantic divergences spanning two languages at a sentence and token-level.

2. Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Yes, the dataset is stored a public GitHub repository that can be found here.

3. What (other) tasks could the dataset be used for?

The dataset could be used to evaluate the performance of computational approaches in detecting meaning differences at token, span and sentence-level.

4. Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

Not that we are aware of.

5. Are there tasks for which the dataset should not be used? If so, please provide a description.

Not that we are aware of.

F. Distribution

1. Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

Yes, the dataset is publicly available.

2. How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset is stored at the following GitHub repository: https://github.com/Elbria/xling-SemDiv/tree/master/REFreSD.

3. When will the dataset be distributed? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant

licensing terms or ToU, as well as any fees associated with these restrictions. If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset was released on 10/04/2020.

4. Will the dataset be distributed under a copyright or other intel- lectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

There is no license associated with the dataset, but there is a request to cite the corresponding paper if the dataset if used: Eleftheria Briakou and Marine Carpuat, Detecting Fine-Grained Cross-Lingual Semantic Divergences without Supervision by Learning to Rank, In EMNLP 2020.

5. Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

There are no fee restrictions.

6. Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

N/A

G. Maintenance

1. Who is supporting, hosting, maintaining the dataset?

This dataset is hosted at the CLIP lab at the University of Maryland.

2. How can the owner, curator, manager of the dataset be contacted?

All questions and comments can be sent to Eleftheria Briakou: ebriakou AT umd DOT edu DOT.

3. Is there an erratum? If so, please provide a link or other access point.

No.

4. Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

All changed to the dataset will be announce at the corresponding GitHub repository.

5. If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

No.

6. Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

All versions of the dataset will be supported unless otherwise communicated on the GitHub repository.

7. If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for com- municating/distributing these contributions to other users? If so, please provide a description.

All information regarding the dataset replicability is publicly available.