Appendix : Weakly Supervised Identification of Cross-lingual Semantic Relations via Knowledge Distillation

Crowd-sourcing Protocol

Collecting Samples for Annotation To build our dataset, we start from the human annotated examples collected by Pavlick et al. (2015). [1] We select lexical pairs from this dataset corresponding to the five relations we are interested in. The right-hand-side term in each pair is then passed through Google Translate to obtain a cross-lingual pair. This pool of cross-lingual word pairs along with the label assigned to the original word pair forms the pool that crowd-workers on FigureEight then annotate.

Task setup We frame the crowdsourcing task as a binary classification task. Given a cross-lingual word pair and the relation assigned to the original monolingual pair, five annotators are asked if the relations holds for the cross-lingual pair. A natural language description of the relation and the cross-lingual pair are put together into a full sentences, and annotators have to answer whether the statement is true or false. Each page shown to annotators contains 10 tasks like the one shown in Figure 1.

Before annotators get access to the task, they are explained the task and the meaning of each relation. Figure 2 shows the explanations of each relation provided to the workers. We also show annotators examples of each relation and explain why the relation holds / does not hold.

Quality Control We control the quality of our annotations in several ways. First, before doing the task, annotators are required to attempt a quiz containing 10 examples and score at least 70% on the quiz, in order to gain access to the task. Next, one of the 10 examples on each page that annotators are provided for annotation is a gold example.

[1] https://cs.brown.edu/people/epavlick/resources/natlog-labeled-rte-pairs.gz

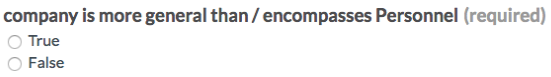

Figure 1: Example task

Annotators are required to maintain the 70% accuracy on these hidden examples as well. If an annotator's performance drops below 70%, all their judgments are marked as "tainted", and we do not use such tainted judgments.

To minimize annotations by speakers who are not fluent in the language we are annotating for, we also use FigureEight's language settings, which only allow annotators with a demonstrated proficiency in particular language to work on the task. We combine this by translating some of the instructions in the task descriptions into the target language.

Aggregating annotations using MACE We use MACE (Hovy et al., 2013), a Bayesian model that estimates the trustworthiness of annotators and accordingly assigns a label to each instance. We run MACE using default parameters in the semi-supervised setting i.e. we use the annotations obtained for quiz questions, along with their true answers as controls, which can help improve the predictions of the model. All samples that are assigned "True" as the final label by MACE form our final test sets.

References

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning Whom to Trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130. Association for Computational Linguistics.

Ellie Pavlick, Johan Bos, Malvina Nissim, Charley

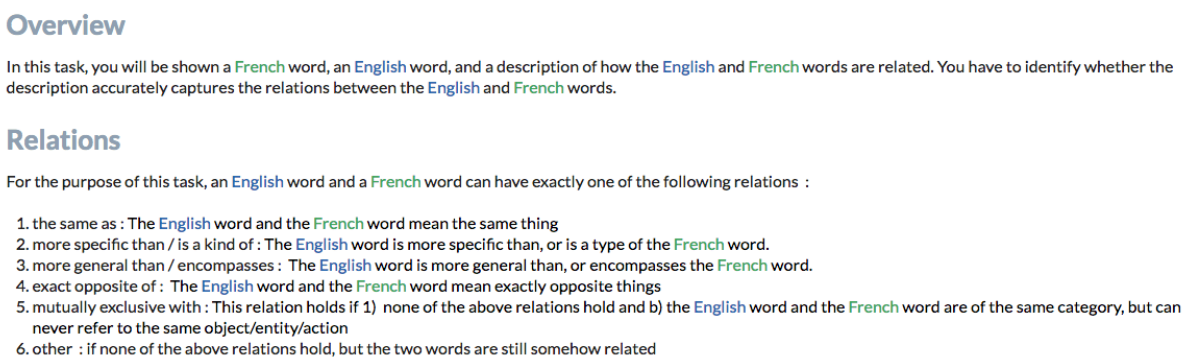

Figure 2: Explanation of the task and the relations

Beller, Benjamin Van Durme, and Chris Callison-Burch. 2015. Adding Semantics to Data-Driven Paraphrasing. pages 1512–1522. Association for Computational Linguistics.