SemEval-2021 Task 12: Learning with Disagreements

Alexandra Uma[1] **Tommaso Fornaciari**[2] **Anca Dumitrache**[3] **Tristan Miller**[4]
Jon Chamberlain[5] **Barbara Plank**[6] **Edwin Simpson**[7] **Massimo Poesio**[1]

[1]Queen Mary University of London [2]Università Bocconi, Milano [3]Albert Heijn
[4]Austrian Research Institute for Artificial Intelligence [5]University of Essex
[6]IT University of Copenhagen [7]University of Bristol
m.poesio@qmul.ac.uk

Abstract

Disagreement between coders is ubiquitous in virtually all datasets annotated with human judgements in both natural language processing and computer vision. However, most supervised machine learning methods assume that a single preferred interpretation exists for each item, which is at best an idealization. The aim of the SemEval-2021 shared task on Learning with Disagreements (Le-wi-Di) was to provide a unified testing framework for methods for learning from data containing multiple and possibly contradictory annotations covering the best-known datasets containing information about disagreements for interpreting language and classifying images. In this paper we describe the shared task and its results.

1 Introduction

The assumption that natural language (NL) expressions have a single and clearly identifiable interpretation in a given context, or that images have a preferred labels, still underlies most work in NLP and computer vision. However, there is now plenty of evidence that this assumption is just a convenient idealization; virtually every project devoted to large-scale annotation has found that genuine disagreements are widespread.

In NLP, that annotator/coder disagreement can be genuine—i.e., resulting from debatable, difficult, or linguistic ambiguity—has long been known for anaphora and coreference (Poesio and Artstein, 2005; Versley, 2008; Recasens et al., 2011).[1] But in recent years, we have also seen evidence that disagreements among subjects/coders are common with virtually every aspect of language interpretation, from apparently simple aspects such as part-of-speech tagging (Plank et al., 2014b), to more complex ones like semantic role assignment (Dumitrache et al., 2019), to subjective tasks such as sentiment analysis (Kenyon-Dean et al., 2018), and to the inferences that can be drawn from sentences (Pavlick and Kwiatkowski, 2019).

In computer vision, as well, the assumption that gold labels may be specified for items has proven an idealization (Rodrigues and Pereira, 2018)—in fact, possibly even more than for NLP. In many widely used crowdsourced datasets for computer vision, different coders assign equally plausible labels to the same items. The problem of disagreement among coders, including experts, on the classification of noisy image data has arisen in many CV applications. This includes classification of astronomical images (Smyth et al., 1994), medical image classification (Raykar et al., 2010), and numerous others (Sharmanska et al., 2016; Rodrigues and Pereira, 2018; Firman et al., 2018).

Many AI researchers have concluded that rather than attempting to eliminate disagreements from annotated corpora, we should preserve them—indeed, some researchers have argued that corpora should aim to collect all distinct interpretations of an expression (Smyth et al., 1994; Poesio and Artstein, 2005; Aroyo and Welty, 2015; Sharmanska et al., 2016; Plank, 2016; Kenyon-Dean et al., 2018; Firman et al., 2018; Pavlick and Kwiatkowski, 2019). Poesio and Artstein (2005) and Recasens et al. (2012) suggest that the best way to create resources capturing disagreements is by preserving *implicit* ambiguity—i.e., having multiple annotators label the items, and then keeping all these annotations, not just an aggregated 'gold standard'. A number of corpora with these characteristics now exist (Passonneau and Carpenter, 2014; Plank et al., 2014a; Dumitrache et al., 2019; Poesio et al., 2019; Rodrigues and Pereira, 2018; Peterson et al., 2019)

Much recent research has explored the question of whether corpora of this type, besides being more

[1]See also the analysis of disagreements in OntoNotes and word senses in Pradhan et al. (2012), Passonneau et al. (2012), and Martínez Alonso et al. (2016).

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 338–347
Bangkok, Thailand (online), August 5–6, 2021.

accurate characterizations of the linguistic reality of language interpretation and image categorization, are also better resources for training NLP and computer vision models, and if so, what is the best way for exploiting disagreements in modeling. Beigman Klebanov and Beigman (2009) used information about disagreements to *exclude* items on which judgements are unclear ('hard' items). In the CrowdTruth project (Aroyo and Welty, 2015; Dumitrache et al., 2019) information about disagreement is used to *weigh* the items used for training. Plank et al. (2014a) proposed to use the information about disagreement to *supplement* the gold label during training. Finally, methods were proposed for training directly from the data with disagreements, without first obtaining an aggregated label (Sheng et al., 2008; Rodrigues and Pereira, 2018; Peterson et al., 2019; Uma et al., 2020). Only limited comparisons of these methods have been carried out (Jamison and Gurevych, 2015), and the sparse research landscape remains fragmented; in particular, methods applied in CV have not yet been tested in NLP, and vice versa.

The objective of SemEval-2021 Task 12, Learning with Disagreements (LE-WI-DI), was to provide a unified testing framework for learning from disagreements in NLP and CV using datasets containing information about disagreements for interpreting language and classifying images. The expectation being that unifying research on disagreement from different fields may lead to novel insights and impact AI widely.

2 Task organization

In order to provide a thorough benchmark for methods for learning from disagreements, we identified five well-known datasets for very different NLP and CV tasks, all characterized by providing a multiplicity of labels for each instance, by having a size sufficient to train state-of-the-art models, and by evincing different characteristics in terms of the crowd annotators and data collection procedure. We found or developed near–state-of-the-art models for the tasks represented by these datasets. Both 'hard' and 'soft' evaluation metrics were employed (Uma et al., n.d.).

The shared task was set up on the CodaLab Competitions platform,[2] which enables training and uniform evaluation on these datasets, such that the crowd learning adaptations of the base models proposed by participants to the task would be directly comparable.

In this section, we briefly introduce the five datasets included in the benchmark and our evaluation criteria. We also elaborate on the setup of the shared task.

2.1 Data

There are by now quite a few datasets preserving disagreements, and covering many levels of language interpretation; remarkably, none of these has ever been used for a shared task like the one we are proposing, and the majority of them have never been used for a shared task at all. Our shared task has aimed at leveraging this diversity. The datasets included are outlined in this section and their characteristics are summarized in Table 1. Figure 1 shows the observed agreement of each dataset.

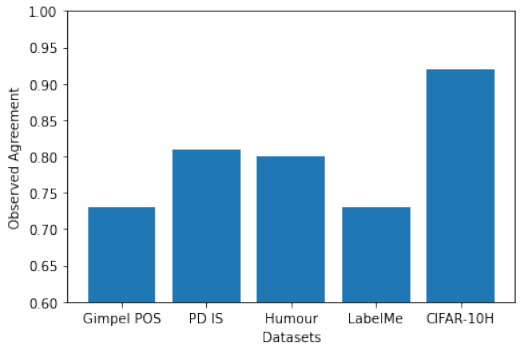

Figure 1: Observed Agreement for each dataset

2.1.1 The Gimpel et al. POS corpus

One widely used resource for developing disagreement-aware NLP models is the dataset of Twitter posts annotated with POS tags collected by Gimpel et al. (2011). Plank et al. (2014b) mapped the Gimpel tags to the universal POS tag set (Petrov et al., 2012) and collected at least five crowdsourced labels per token from 177 annotators. This dataset contains 14K training examples (English words/tokens) annotated by 177 annotators. Each item was annotated between five and 177 times, 16.38 times on average. For this shared task, we selected 8.3K, 3K, and 3.1K tokens as training, development and test sets respectively.

2.1.2 The PDIS corpus

The *Phrase Detectives* corpus (Poesio et al., 2019) is a crowdsourced coreference corpus collected

[2] https://www.microsoft.com/en-us/research/project/codalab/

	POS	PDIS	HUMOUR	IC-LABELME	CIFAR-10H
Number of items	14,000	96,305	18,002	10,000	10,000
Number of crowd workers	177	1,741	272	59	2,457
Number of categories	12	2	2	8	10
Average annotations per item	16.37	11.87	5.00	2.50	51.10

Table 1: Summary of dataset characteristics

with the *Phrase Detectives* gamified online platform (Poesio et al., 2013).[3] We use PDIS, a simplified version of the corpus containing only binary information status labels: Discourse New (the entity referred to has never been mentioned before) and Discourse Old (it has been mentioned). PDIS consists of 542 documents, for a total of 408K tokens and over 96K markables. These documents were annotated by game players who produced an average of 11.87 annotations per markable.

Forty-five of the documents (5.2K markables), collectively called PD_{gold}, additionally contain expert-adjudicated gold labels. This subset of PDIS was designated as the test set. The training and development datasets consist of 473 documents (and 86.9K markables) and 24 documents (4.2K markables) respectively.

2.1.3 The Humour dataset

The comprehension and appreciation of humour is known to vary across individuals (Ruch, 2008), making disagreement over the perceived funniness of jokes an appealing subject of study. For our training data, we used the corpus of Simpson et al. (2019), which consists of 4,030 short texts (3398 jokes, mostly based on puns, and 632 non-jokes such as proverbs and aphorisms). 28,210 unique pairings of these texts were presented to five crowdsourcers each, who indicated which text in the pair (if either) they found to be funnier. The goal is to learn a model that can predict binary pairwise labels that can predict which of two short texts is funnier.

The 4,030 text instances were split into 60% (2,418 texts, 9,916 unique pairs) for the training set and 20% (806 texts, 1,086 unique pairs) for the development set. Since this dataset has already been published, we constructed a new test dataset along similar lines: 1,000 short texts (all punning jokes) were paired in 7,000 different ways, and each of these 7,000 pairs was then presented to five crowd workers for a preference judgement.[4]

2.1.4 The LabelMe corpus

Much research on learning from disagreements was motivated by computer vision datasets, so we intended to include some of these, too. Possibly the most widely used such corpus is the LabelMe dataset[5] (Russell et al., 2008). It classifies outdoor images according to 8 categories: *highway, inside city, tall building, street, forest, coast, mountain* or *open country*. Using Amazon Mechanical Turk, Rodrigues and Pereira (2018) collected an average of 2.5 annotations per image from 59 annotators for 10K images in this dataset.

We randomly selected 5K, 2.5K, and 2.5K images for training, development, and testing respectively, careful to keep the label proportions in each subset close to the proportions in the 10K dataset.

2.1.5 The CIFAR-10H corpus

Krizhevsky's (2009) CIFAR-10 dataset consists of 60K tiny images from the web, carefully labelled and expert-adjudicated to produce a single gold label for each image in one of 10 categories: *airplane, automobile, bird, cat, deer, dog, frog, horse, ship,* and *truck*. Peterson et al. (2019) collected crowd annotations for 10K images from this dataset (the designated test portion) using Amazon Mechanical Turk, creating the CIFAR-10H dataset[6] which we use for this shared task.

We randomly selected 7K, 1K, and 2K images for training, development and testing respectively. We kept as much data as we could for training without jeopardizing the evaluation process, as the base model was found to be sensitive to data size. As with the original dataset, each subset we created contains an equal number of images per category.

2.2 Evaluation metrics

While recent research questions the assumption that a single 'hard' label (a gold label) exists for every

[3] https://github.com/dali-ambiguity

[4] US-based workers from Amazon Mechanical Turk were employed, paid in line with the federal minimum wage.

[5] http://labelme.csail.mit.edu/Release3.0

[6] https://github.com/jcpeterson/cifar-10h

item in a dataset, the models proposed for learning from multiple interpretations are still largely evaluated under this assumption, using 'hard' measures like accuracy or class-weighted F_1 (Sheng et al., 2008; Plank et al., 2014a; Martínez Alonso et al., 2015; Sharmanska et al., 2016; Rodrigues and Pereira, 2018). For reference and comparison reasons, we also evaluate the models produced for this shared task using F_1.

However, a way of evaluating models as to their ability to capture disagreement is needed, especially for datasets with substantial extent of disagreement. The simplest 'soft' metric of this type is to evaluate ambiguity-aware models by treating the probability distribution of labels they produce as a **soft label**, and comparing that to the full distribution produced by annotators, using, for example, cross-entropy. This approach was adopted in, *inter alia*, (Peterson et al., 2019; Uma et al., 2020). Peterson et al. (2019) tested this approach on image classification tasks, generating the soft label by transforming the item annotation distribution using standard normalization. In this shared task we also use standard normalization to produce soft labels for the humour dataset. Uma et al. (2020) show that the choice of soft label encoding function depends on the characteristics of the dataset. For POS and IC-LABELME, they show that a softmax function over the annotator distribution is preferable over standard normalization. On the other end, for PDIS, training a soft-loss model using the posterior probability produced by Hovy et al.'s (2013) MACE probabilistic aggregation model as a soft label produces predictions that a most accurate with respect to the gold.

Therefore, in this shared task we used different soft label encoders to generate soft labels from annotator distributions for the test data.

2.3 Task setup

CodaLab was the designated site for hosting SemEval-2021 competitions.[7] LE-WI-DI was run in two main phases:

Practice phase. In the practice phase, the goal was to train models for each task to learn from crowd annotations, given (1) the training data (consisting of raw and preprocessed input data and crowd annotations), (2) the development data with no labels, and (3) the base models (discussed in Section 3). While participants were encouraged to start with the base models and extend them, we did not make this mandatory. Participants could test the performance of their models on the development set by making predictions on the given development input data and then uploading their submissions to CodaLab for preliminary testing. We permitted up to 999 submissions in this phase. The 'leader board' was made public to allow participants not only to see how their models performed, but also to compare the performance of their model to those submitted by other participants.

Evaluation phase. The evaluation phase was the official testing phase of the competition. In this phase, we released test data (without labels) but we also released the gold labels and crowd annotations for the development set to facilitate quick offline testing and refining of models and model selection. The number of submissions for this phase was limited to ten submissions per participant to prevent the participants from fine-tuning their models on the test data.[8] The allowed number of submissions was later increased to 999 to more encourage submission attempts. The leader board was also kept public in this phase. Each participant could see the best model of each of the tasks using each of the evaluation metrics.

Post-campaign evaluation. As our aim was to make this benchmark available beyond the competition to researchers developing disagreement-aware models, we included a third, post-evaluation phase to allow lifetime access to the data. Researchers participating in this phase will be able to access the same data as in the evaluation phase and test their models on the test data for the various tasks.

3 Base models and baselines

In order to encourage the participants to focus on the development of methods for learning from disagreement, as opposed to achieving higher performance by developing better models, we provided 'base' models for each of the tasks represented by the aforementioned corpora. In this section, we briefly discuss the baseline models for each task that we provided. In Section 5, we report the results using these base models and two crowd learning approaches: majority voting and the soft loss method (Peterson et al., 2019; Uma et al., 2020).

[7] Our competition can be found at https://competitions.codalab.org/competitions/25748.

[8] This proved unnecessary as the inherent difficulty of the shared task was enough of a deterrent.

The POS tagging model. The POS tagger is a bi-LSTM (Plank et al., 2016) with additional use of attention over the input word and character embeddings, as used in Uma et al. (2020).

The PDIS classification model. The model for this task was developed by comparing architectures from two models: a state-of-the-art coreference model and a state-of-the-art IS classification model. We combined the mention representation component of Lee et al.'s (2018) coreference resolution system with the mention sorting and non-syntactic feature extraction components of the IS classification model proposed by Hou (2016)[9] to create a novel IS classification model that outperforms Hou (2016) on the PDIS corpus. The training parameters were set following Lee et al. (2018).

The humour preference learning model. We use as base model for this task Gaussian process preference learning (GPPL) with stochastic variational inference, as described and implemented by Simpson and Gurevych (2020). As an input vector to GPPL, we first take the mean word embedding of a text, using 300-dimensional word2vec embeddings trained on the Google News corpus (Mikolov et al., 2013). Then, we compute the frequency of each unigram in the text in a 2017 Wikipedia dump, and each bigram in the text in a Google Books Ngram dataset. Finally, we concatenate the mean unigram and bigram frequencies with the mean word embedding vector to obtain the input vector representation for each short text. The GPPL model is trained on pairwise labels from the training set to obtain a ranking function that can be used to score test instances or output pairwise label probabilities. As a Bayesian model, it takes into account sparsity and noise in the crowdsourced training labels, and moderates its confidence accordingly. Hence, it is a strong baseline for accounting for disagreement among annotators. This same GPPL approach set the previous state of the art on the humour dataset (Simpson et al., 2019).

The LabelMe image classification model. For this task, we replicated the model from Rodrigues and Pereira (2018). The images were encoded using pretrained CNN layers of the VGG-16 deep neural network (Simonyan et al., 2013). This encoding is passed into a feed-forward neural network layer with a RELU activated hidden layer with 128 units. A 0.2 dropout is applied to this learned representation which is then passed through a final layer with softmax activation to produce the model's predictions.

The CIFAR-10 image classification model. The trained model provided for this task is the ResNet-34A model (He et al., 2016), a deep residual framework which is one of the best performing systems for the CIFAR-10 image classification. We made available to participants the publicly available Pytorch implementation of this ResNet model.[10]

4 Participating systems

Unfortunately, we observed a dramatic difference in the number of participants that signed up to the competition (over 100 groups), the number of groups that participated in the trial phase, and the number of groups that submitted a run for official evaluation.[11] Only one group, UOR, submitted in the evaluation phase (Osei-Brefo et al., 2021). However, they did submit models for each of the tasks, and did adopt a learning from disagreements approach.

POS tagging. For POS tagging, UOR developed a novel POS tagging model by fine-tuning the BERT language model (Devlin et al., 2019). The (tweet, token) pairs were encoded in the form

[CLS] Tweeted text [SEP] Token [SEP]

where the '[CLS]' token was added for classification and the '[SEP]' token separated the tweet from the token under consideration. To learn the class for the token, the learned classification token was passed through a single feed-forward neural network layer with softmax activation. The output of this layer represented the probabilities of the token belonging to each of the 12 classes.

To extend this model for crowd learning, UOR added an adaptation of the crowd layer from Rodrigues and Pereira (2018). Rather than compute a single loss from the crowd layer as Rodrigues and Pereira (2018) do, UOR compute a joint loss from both the crowd layer and the base model (without the crowd layer bottleneck).

[9] This model was developed for fine-grained information status classification on the ISNOTES corpus (Markert et al., 2012; Hou et al., 2013).

[10] https://github.com/KellerJordan/ResNet-PyTorch-CIFAR10

[11] Two participating groups cited an inability to come up with a novel crowd learning paradigm as the reason they did not submit for official evaluation.

PDIS classification. For this task, UOR also used a fine-tuned BERT together with Rodrigues and Pereira's (2018) crowd layer. Each (document, markable) pair was encoded as follows:

[CLS] + Document + [SEP] + Markable + [SEP]

where the '[CLS]' and '[SEP]' tokens are used in the same manner as in POS tagging.

Humour preference learning. For humour preference learning, the participant submitted predictions using the base model without modifications.

LabelMe image classification (IC-LABELME). For this task, UOR adapted the Rodrigues and Pereira (2018) crowd layer to the base model.

CIFAR-10H image classification (IC-CIFAR10H). For IC-CIFAR10H, the crowd labels were aggregated into hard labels using majority voting. However, UOR combined Zagoruyko and Komodakis's (2016) WideResNet model, which has been shown to outperform He et al.'s (2016) ResNet with the novel Sharpness-Aware Minimization (SAM) optimization technique, proposed by Foret et al. (2020), that has been shown to efficiently improve model generalization, especially on noisy, singly labelled data.

5 Results and discussion

Table 2 contains the results of various models discussed in Sections 3 and 4 on this shared task when evaluated based on the hard metric (i.e., the class-weighted F_1 with respect to the gold labels) and the soft metric (the cross-entropy between the soft labels for each task—see Section 2.2—and the model prediction for that task). The best results for each task are highlighted in bold.

UOR concentrated their effort on the IC-CIFAR10H dataset, on which they did achieve good results and outperformed the baseline (see below). In the other datasets, their official results at the end of the evaluation phase were less competitive.

With the POS and PDIS datasets, the model proposed by UOR, adding a crowd layer on top of BERT, achieved substantially worse results than training from a label aggregated using majority voting or training using a soft-loss function, both according to the hard evaluation metric (F_1) and the soft metric (CE). The ranking between soft-loss method, aggregation, and crowd layer with POS is consistent with that obtained by Uma et al. (n.d.), but the results obtained by UOR are much worse for reasons that will require further investigation. (With PDIS, Uma et al. (n.d.) obtain comparable results with soft-loss functions and with the crowd layer.) More generally, the results show that although the hard label (the majority voting aggregate of the annotator distribution) and the soft label (a probability distribution encoding of the annotator distribution) were drawn from the same annotator distribution with this dataset, given the same base model, training by targeting the soft label (base model + soft loss) outperforms training using majority voting aggregates (base model + majority voting) regardless of which evaluation metric is used to compare the models.

For the humour preference learning task, again, the base model outperforms UOR's submission on both metrics, but in this case the difference in performance between GPPL and UOR is much less substantial with the hard metric, although it remains large according to the soft metric. This large difference may be due to a technical issue that requires further investigation, since UOR's submission was also supposed to have been produced by the same base system. A possible reason for poor cross-entropy error is the use of discrete labels, which are heavily penalized for overconfidence by cross-entropy error. On this soft metric, the Bayesian probabilistic approach of GPPL may have advantages over approaches with poorer calibration, which remains to be explored in future work. The GPPL approach therefore remains the state of the art with this dataset.

For IC-LABELME, again, soft-loss training achieved better hard and soft scores than both aggregation training with majority voting labels and the UOR extension of the base model using a crowd layer adapted from Rodrigues and Pereira (2018). The finding that the UOR group's adaption of the Rodrigues and Pereira (2018) crowd layer yielded lower F_1 than training using majority voting is unexpected, given that in Rodrigues and Pereira (2018); Uma et al. (2020) and Uma et al. (n.d.), the crowd layer, particularly the DL-MW variant, was shown to be a competitive approach to learning from crowds and always outperforms majority voting. However, UOR's crowd layer does achieve better soft evaluation (cross-entropy) scores than majority voting.

There is one dataset, however, on which UOR outperformed the two baselines: IC-CIFAR10H. For this dataset, UOR used Zagoruyko and Komodakis's (2016) WideResNet image classifier trained using majority voting aggregated labels and

Task	Model	Hard score (F_1)	Soft score (cross-entropy)
POS	base model + majority voting	0.753	2.263
POS	base model + soft loss	**0.767**	**1.084**
POS	UOR (BERT + Crowd Layer)	0.125	2.331
PDIS	base model + majority voting	0.906	0.397
PDIS	base model + soft loss	**0.928**	**0.273**
PDIS	UOR (BERT + Crowd Layer)	0.474	0.830
HUMOUR	base model (GPPL)	**0.557**	**0.728**
HUMOUR	UOR	0.513	3.697
IC-LABELME	base model + majority voting	0.806	2.833
IC-LABELME	base model + soft loss	**0.833**	**1.691**
IC-LABELME	UOR (base model + Crowd Layer)	0.784	1.769
IC-CIFAR10H	base model + majority voting	0.646	2.610
IC-CIFAR10H	base model + soft loss	0.698	1.052
IC-CIFAR10H	UOR (WideResNet + SAM)	**0.769**	**0.827**

Table 2: Results on the benchmarks and participant submissions on all the tasks using F_1 (higher is better) and cross-entropy (lower is better)

Foret et al.'s (2020) SAM optimization technique. The results show that WideResNet outperforms ResNet with this task both according to the hard metric and the soft metric. Interestingly, this is the one dataset in which the Deep Learning from Crowds approach of Rodrigues and Pereira (2018) works best according to Uma et al. (n.d.), outperforming both soft-loss training and majority voting training. It would thus be interesting to understand if the performance of UOR's model could be further increased by adopting one of these methods.[12]

6 Conclusion

This shared task presented the first unified testing framework for learning with disagreements. The datasets include sequence labelling, three classification tasks, and preference learning, hence provide a testbed for a wide range of challenges when learning from multiple annotators. We proposed to evaluate not just the 'hard' performance against a gold standard, but also the ability to predict the distribution of different interpretations of the data—that is, the alternative labellings provided by different annotators. The results show the benefit of soft loss functions that account for the distribution of labels in the training data. However, modelling alternative interpretations of data remains an under-researched topic in NLP and computer vision. To encourage future work on learning with disagreements, the shared task and datasets will remain available for evaluating new methods.

[12]As a postscript, we should note that after the end of the official competition we did carry out an investigation of the reasons for the poor performance of UOR's models on the tasks other than IC-CIFAR10H. Some points emerging from the discussion are presented in the participants' paper for the shared task.

Acknowledgments

Alexandra Uma, Jon Chamberlain, and Massimo Poesio were partially supported by the DALI project, ERC Advanced Grant 695662. Tristan Miller was supported by the Austrian Science Fund (FWF) under project M 2625-N31. Barbara Plank is supported in part by the Independent Research Fund Denmark (DFF) grant 9131-00019B and 9063-00077B.

References

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Beata Beigman Klebanov and Eyal Beigman. 2009. From annotator agreement to noise models. *Computational Linguistics*, 35(4):495–503.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, volume 1, pages 4171–4186. Association for Computational Linguistics.

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2019. A crowdsourced frame disambiguation corpus with

ambiguity. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, volume 1, pages 2164–2170. Association for Computational Linguistics.

Michael Firman, Neill D. F. Campbell, Lourdes Agapito, and Gabriel J. Brostow. 2018. DiverseNet: When one right answer is not enough. *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5598–5607.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. 2020. Sharpness-aware minimization for efficiently improving generalization. *CoRR*, abs/2010.01412.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 42–47. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Yufang Hou. 2016. Incremental fine-grained information status classification using attention-based LSTMs. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1880–1890. The COLING 2016 Organizing Committee.

Yufang Hou, Katja Markert, and Michael Strube. 2013. Global inference for bridging anaphora resolution. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 907–917. Association for Computational Linguistics.

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*, pages 1120–1130. Association for Computational Linguistics.

Emily Jamison and Iryna Gurevych. 2015. Noise or additional information? Leveraging crowdsource annotation item agreement for natural language tasks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 291–297. Association for Computational Linguistics.

Kian Kenyon-Dean, Eisha Ahmed, Scott Fujimoto, Jeremy Georges-Filteau, Christopher Glasz, Barleen Kaur, Auguste Lalande, Shruti Bhanderi, Robert Belfer, Nirmal Kanagasabai, Roman Sarrazingendron, Rohit Verma, and Derek Ruths. 2018. Sentiment analysis: It's complicated! In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, volume 1, pages 1886–1895. Association for Computational Linguistics.

Alex Krizhevsky. 2009. Learning multiple layers of features from tiny images. https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, volume 2, pages 687–692. Association for Computational Linguistics.

Katja Markert, Yufang Hou, and Michael Strube. 2012. Collective classification for fine-grained information status. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 795–804. Association for Computational Linguistics.

Héctor Martínez Alonso, Anders Johannsen, and Barbara Plank. 2016. Supersense tagging with inter-annotator disagreement. In *Proceedings of the 10th Linguistic Annotation Workshop*, pages 43–48. Association for Computational Linguistics.

Héctor Martínez Alonso, Barbara Plank, Arne Skjærholt, and Anders Søgaard. 2015. Learning to parse with IAA-weighted loss. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1357–1361. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, volume 2, pages 3111–3119. Curran Associates Inc.

Emmanuel Osei-Brefo, Thanet Markchom, and Huizhi Liang. 2021. UOR at SemEval-2021 Task 12: On crowd annotations; learning with disagreements to optimise crowd truth. In *Proceedings of the 15th International Workshop on Semantic Evaluation*. To appear.

Rebecca J. Passonneau, Vikas Bhardwaj, Ansaf Salleb-Aouissi, and Nancy Ide. 2012. Multiplicity and word sense: evaluating and learning from multiply labeled word sense annotations. *Language Resources and Evaluation*, 46(2):219–252.

Rebecca J. Passonneau and Bob Carpenter. 2014. The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics*, 2:311–326.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, and Olga Russakovsky. 2019. Human uncertainty makes classification more robust. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision*, pages 9616–9625.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 2089–2096. European Language Resources Association.

Barbara Plank. 2016. What to do about non-standard (or non-canonical) language in NLP. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014a. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751. Association for Computational Linguistics.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014b. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 507–511. Association for Computational Linguistics.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 412–418. Association for Computational Linguistics.

Massimo Poesio and Ron Artstein. 2005. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 76–83. Association for Computational Linguistics.

Massimo Poesio, Jon Chamberlain, Udo Kruschwitz, Livio Robaldo, and Luca Ducceschi. 2013. Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Transactions on Intelligent Interactive Systems*, 3(1).

Massimo Poesio, Jon Chamberlain, Silviu Paun, Juntao Yu, Alexandra Uma, and Udo Kruschwitz. 2019. A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1778–1789. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Proceedings of the Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes, Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1–40. Association for Computational Linguistics.

Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. 2010. Learning from crowds. *Journal of Machine Learning Research*, 11(43):1297–1322.

Marta Recasens, Ed Hovy, and M. Antònia Martí. 2011. Identity, non-identity, and near-identity: Addressing the complexity of coreference. *Lingua*, 121(6):1138–1152.

Marta Recasens, M. Antònia Martí, and Constantin Orasan. 2012. Annotating near-identity from coreference disagreements. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 165–172. European Language Resources Association.

Filipe Rodrigues and Francisco C. Pereira. 2018. Deep learning from crowds. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pages 1611–1618.

Willibald Ruch. 2008. Psychology of humor. In Victor Raskin, editor, *The Primer of Humor Research*, number 8 in Humor Research, pages 17–100. Mouton de Gruyter, Berlin.

Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman. 2008. LabelMe: A database and Web-based tool for image annotation. *International Journal of Computer Vision*, 77:157–173.

Viktoriia Sharmanska, Daniel Hernández-Lobato, José Miguel Hernández-Lobato, and Novi Quadrianto. 2016. Ambiguity helps: Classification with disagreements in crowdsourced annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2194–2202.

Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. 2008. Get another label? Improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 614–622.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034.

Edwin Simpson, Erik-Lân Do Dinh, Tristan Miller, and Iryna Gurevych. 2019. Predicting humorousness and metaphor novelty with Gaussian process preference learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*,

pages 5716–5728. Association for Computational Linguistics.

Edwin Simpson and Iryna Gurevych. 2020. Scalable Bayesian preference learning for crowds. *Machine Learning*, 109(4):689–718.

Padhraic Smyth, Usama Fayyad, Michael Burl, Pietro Perona, and Pierre Baldi. 1994. Inferring ground truth from subjective labelling of venus images. In *Proceedings of the 7th International Conference on Neural Information Processing Systems*, page 1085–1092. MIT Press.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2020. A case for soft-loss functions. In *Proceedings of the 8th AAAI Conference on Human Computation and Crowdsourcing*, pages 173–177.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. n.d. Learning from disagreements. Submitted.

Yannick Versley. 2008. Vagueness and referential ambiguity in a large-scale annotated corpus. *Research on Language and Computation*, 6(3):333–353.

Sergey Zagoruyko and Nikos Komodakis. 2016. Wide residual networks. *CoRR*, abs/1605.07146.