Disagreement is Natural

Crowdsourcing Labeled Data for Natural Language Processing

Anca Dumitrache

January 2019 – version 1.1

SIKS Dissertation Series No. XXXXXXXXXXXXXXXXXXXX

The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

VRIJE UNIVERSITEIT

Disagreement is Natural

Crowdsourcing Labeled Data for Natural Language Processing

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van rector magnificus
prof. dr. V. Subramaniam,
in het openbaar te verdedigen
ten overstaan van de promotiecomissie
van de Faculteit der Bètawetenschappen
op XXXXXXXXXXXXXXXXXXX
in de aula van de universiteit,
De Boelelaan 1105

door

Anca Dumitrache

geboren te Bistrița, Roemenië

promotor	prof. dr. L.M. Aroyo
promotor	prof. dr. C. Welty
copromotor	
promotiecomissie	

ACKNOWLEDGEMENTS

..

CONTENTS

1

INTRODUCTION

Language is the source of misunderstandings.
– Antoine de Saint-Exupéry, The Little Prince

In this chapter, we present the motivation of this thesis, as well as related work on crowdsourcing ground truth for natural language processing. We introduce the CrowdTruth methodology for crowdsourcing ground truth while preserving inter-annotator disagreement. CrowdTruth is based on the idea that disagreement is not noise, but an important signal that can be used to capture ambiguity in the annotation data. We define the main research goal of how to interpret disagreement in crowdsourcing ground truth for natural language processing, along with four research questions that address this goal. Finally, we outline the main contributions of this thesis.

This chapter is based on the paper titled *Crowdsourcing Disagreement for Collecting Semantic Annotation* in the European Semantic Web Conference [38].

1.1 MOTIVATION

As knowledge available on the Web expands, natural language processing methods have become invaluable for facilitating data navigation. Tasks such as knowledge base completion and disambiguation are solved with machine learning models for natural language processing that require a lot of data. Human-annotated gold standard, or ground truth, is used for training, testing, and evaluation of these machine learning components. The traditional approach to gathering this data is to employ domain experts to perform annotation tasks.

However, such an annotation process can be both expensive, and time consuming [4], due to the costs of working with domain experts. Furthermore, experts might prove difficult to find for broad, open domains (e.g. the annotation of news articles). This presents a challenge for extending natural language processing methods into new domains. Human annotation is needed to solve this problem, but the process of gathering this data is not scalable at the level of the large datasets currently available on the Web. Efficiently integrating human knowledge with automated methods is necessary for tackling this issue.

In recent years, crowdsourcing has become a viable alternative to using domain expert annotators, as it is both cheaper and more easily scalable [110]. This has been facilitated by platforms such as Amazon

Mechanical Turk[1] and Figure Eight[2] (formerly known as Crowdflower) that offer readily available crowds of workers. The main challenge posed by crowdsourcing is how to tune the annotation tasks (e.g. in terms of worker selection, task question and template) in order to get the best quality of data [37]. The quality of the annotated data can have a big impact on the performance of machine learning models that learn from it – so much so that Amazon has started offering a service[3] that optimizes the collection of human-labeled ground truth.

But what makes annotations high quality is still a matter of discussion. When collecting multiple annotations for the same task, it is likely that inter-worker disagreement will be present. In typical annotation setups it is assumed that one correct answer exists for every question, and that disagreement must be eliminated from the corpus. This traditional approach to gathering annotation, based on restrictive annotation guidelines, can often results in over-generalized observations, as well as a loss of ambiguity inherent to language [4], thus becoming unsuitable for use in training natural language processing systems.

The CrowdTruth[4] methodology [5, 8] has been proposed to perform crowdsourcing while preserving inter-annotator disagreement. CrowdTruth is based on the idea [8] that disagreement is not noise, but an important signal that can be used to capture ambiguity in the annotated data. It considers the crowdsourcing system as a triangle [7] with three components that are inter-connected: workers, input data, and annotations. CrowdTruth captures inter-annotator disagreement and uses it to calculate a set of quality metrics[5] [48, 64] for the three crowdsourcing components, by modeling the way that the components interact with each other – e.g. in an ambiguous sentence, we expect to have more disagreement between workers, therefore workers on those sentences should not be considered less trustworthy. Previous research in crowdsourcing medical relation extraction [5, 6] has shown that disagreement can be an informative, useful property, and its analysis can result in reduced time, lower cost, better scalability, and better quality human-annotated data.

This thesis explores how the CrowdTruth methodology can be used to collect ground truth data for the training and evaluation of natural language processing models. We present work done across several tasks (relation extraction, semantic frame disambiguation) and domains (medical, open), showing the role of inter-annotator disagreement beyond simply identifying low quality workers. We argue that disagreement does not need to be eliminated from ground truth data in order to preserve data quality. Furthermore, we show that disagree-

1 https://www.mturk.com/
2 https://www.figure-eight.com/
3 https://aws.amazon.com/sagemaker/groundtruth/
4 http://crowdtruth.org
5 https://github.com/CrowdTruth/CrowdTruth-core

ment is a valuable quality to preserve in ground truth data, that can be effectively used in the training and evaluation of natural language processing models. This is because inter-annotator disagreement is a powerful signal for the ambiguity that is inherent in natural language. Our goal is to break the constraints of the typical methodology for collecting ground truth, and prove that disagreement is a necessary characteristic of annotated data that, when interpreted correctly, can improve the performance of natural language processing models, and make evaluations more attuned to the noise in real-world data.

1.2 RELATED WORK

Crowdsourcing is a widely used method to collect natural language processing ground truth [110]. In this section, we explore background work on four important crowdsourcing issues: (1) how to establish crowd data quality, (2) how to aggregate multiple crowd annotations, (3) how natural language processing models use crowd data in training and evaluation, (4) and what the relation is between inter-worker disagreement and natural language ambiguity. These issues make up the backbone and the main topics that will be discussed in this thesis. The related work on these issues is composed of papers published in a wide variety of venues, across three main fields of artificial intelligence:

- *human computation*, where the main venues are the Conference on Human Computation and Crowdsourcing (HCOMP), the Journal of Human Computation, the ACM Transactions on Interactive Intelligent Systems (TiiS) journal, and the Conference on Human Factors in Computing Systems (CHI);
- *natural language processing*, and *machine learning* more generally, where the main venues are the Annual Meeting of the Association for Computational Linguistics (ACL), the Conference on Empirical Methods in Natural Language Processing (EMNLP), the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), and the Conference on Neural Information Processing Systems (NeurIPS);
- *semantic web*, where the main venues are the International Semantic Web Conference (ISWC), the Extended Semantic Web Conference (ESWC), and the Semantic Web Journal.

1.2.1 *Crowd Data Quality*

Determining the quality of crowdsourced data collected from non-experts has been the subject of study since the work of Snow et al. [117], who have shown that the crowd can produce annotations with expert-level quality for a variety of natural language processing tasks:

affect recognition, word similarity, recognizing textual entailment, event temporal ordering, and word sense disambiguation. Despite these promising results, collecting high quality crowdsourced data is still a challenge, due primarily to the difficulty of identifying and preventing spam behavior of workers [34, 36]. This is especially the case when applying non-expert crowdsourcing to domains that are typically thought to require expertise on the part of the annotators [19].

The medical domain is particularly difficult, with even expert annotators sometimes producing low-quality data [87]. Nevertheless, there exists some research that successfully employed non-expert crowdsourcing to collect annotations in the medical domain. Mortensen, Musen, and Noy [95] use crowdsourcing to verify relation hierarchies in biomedical ontologies. On 14 relations from the SNOMED CT CORE Problem List Subset, the authors report the crowd's accuracy at 85% for identifying whether the relations were correct or not. Burger et al. [18] used crowdsourcing to extract the gene-mutation relations in Medical Literature Analysis and Retrieval System Online (MEDLINE) abstracts. Focusing on a very specific gene-mutation domain, the authors report a weighted accuracy of 82% over a corpus of 250 MEDLINE abstracts. Li, Good, and Su [83] performed a study exposing ambiguities in a gold standard for drug-disease relations with crowdsourcing. They found that, over a corpus of 60 sentences, levels of crowd agreement varied in a similar manner to the levels of agreement among the original expert annotators. Zhai et al. [131] describe a method for crowdsourcing a ground truth for medical named entity recognition and entity linking. In a dataset of over 1,000 clinical trials, the authors show no statistically significant difference between the crowd and expert-generated gold standard for the task of extracting medications and their attributes.

Other difficult annotation tasks involve linguistics knowledge. For instance, frame disambiguation requires an understanding of the frame semantics theory [12], which can be difficult to explain to a crowd of non-experts. While Hong and Baker [60] showed a high accuracy when comparing the crowd to experts for the task of frame disambiguation by simply calculating the majority vote, Chang et al. [25] claim that a more complex multi-step annotation process is required in order to correct misunderstandings of the frame definition by the crowd.

In all of these experiments, disagreement between annotators is seen as undesirable and a sign of low quality data. In contrast, Jurgens [71] argues that ambiguity is an inherent feature of frame/word sense disambiguation, and that crowdsourcing can be used to capture it, by asking annotators to rate ambiguous examples on a Likert scale. Similarly, this thesis proposes that ambiguity is a useful property of natural language, but instead of asking workers directly to rate ambiguity, we study it through measuring inter-annotator disagreement. This presents an interesting challenge, as disagreement is usually removed from annotated datasets in order to improve their quality. Our goal in this work is to show that crowdsourced ground truth can still have quality comparable to that of domain experts, while still preserving the signals of worker disagreement.

1.2.2 *Crowdsourcing Aggregation Methods*

The most common way to aggregate crowd annotations is majority voting, where the label for an example is picked based on whether or not the majority of crowd workers agree that it exists. Inter-annotator agreement in crowdsourcing is usually employed as a method to determine the quality of the annotations. Typically, disagreement is considered an undesirable feature of the annotations – a byproduct either of low quality of the workers, or of an unclear annotation task. There are several metrics to capture inter-annotator agreement, most popular being Cohen's κ [33] and Krippendorff's α [76]. Artstein and Poesio [10] compared several of these metrics, finding that the choice of metric is not as important as it is to increase the number of annotators, in order to reduce the prevalence of personal bias.

In recent years, there is also a growing body of research on alternative crowdsourcing aggregation metrics. There is a particular focus on modeling the reliability of crowd workers, by identifying spam workers [16, 68, 75], and analyzing workers' performance for quality control and optimization of the crowdsourcing processes [115]. Whitehill et al. [129] and Welinder et al. [125] have used a latent variable model for task difficulty, as well as latent variables to measure the skill of each annotator, to optimize crowdsourcing for image labels. Werling et al. [128] use on-the-job learning with Bayesian decision theory to assign the most appropriate workers for each task, for both text and image annotation. Prelec, Seung, and McCoy [107] show that the surprisingly popular crowd choice (i.e. the answer that most workers thought would not be picked by other workers, even though it is correct) gave better results than the majority vote for a variety of tasks with unambiguous ground truths (state capitals, trivia questions and price of artworks). Finally, Paun et al. [101] compare majority vote with 6 different Bayesian methods that aggregate crowd results while also modeling worker reliability and task item difficulty. The evaluation over a variety of task settings (binary and multiple choice, different levels of quality for the workers) shows 5 out of 6 of the Bayesian methods consistently outperform majority vote.

Our research is part of this current trend of investigating the limitations of majority vote as a crowdsourcing aggregation method. The novel approach of CrowdTruth is the modeling of ambiguity as a latent variable of the crowdsourcing system, that is present in inter-worker disagreement. Therefore, instead of discarding it, the CrowdTruth approach preserves disagreement and uses it to identify ambiguous data points. In this thesis, we will show that the CrowdTruth method to aggregate crowdsourcing annotations is applicable to a variety of annotation tasks, where simply using majority vote would result in the loss of important information regarding the ambiguity in the data.

1.2.3 *Natural Language Processing with the Crowd*

Due to being both cheaper and more readily available than domain experts, crowdsourcing is used to collect training data for a variety of natural language processing tasks, across several domains: medical entity extraction [53, 122, 131], medical relation extraction [73, 122], open-domain relation extraction [78], clustering and disambiguation [81], ontology evaluation [98], web resource classification [22] and taxonomy creation [17]. Snow et al. [116] have shown that aggregating the answers of an increasing number of unskilled crowd workers with majority vote can lead to high quality natural language processing training data.

In this thesis, we focus on the training of one natural language processing task – relation extraction from sentences. This task usually requires large amounts of training data, meaning that completely crowdsourcing the ground truth is cost-prohibitive. However, active and semi-supervised methods can be used to scale-up the signal in labeled data to unlabeled examples. Angeli et al. [2] used an active learning approach to identify candidate sentences for crowd labeling that will most impact the performance of their relation extraction model. Levy et al. [82] have shown that a small crowdsourced dataset of questions about relations can be exploited to perform zero-shot learning. Pershina et al. [104] used a small dataset of hand-labeled data to generate relation-specific guidelines that are used as additional features in the relation extraction.

The approach in these works is to restrict disagreement between annotators by using either of the following methods: restricting annotator guidelines, picking one answer that reflects some consensus usually through majority voting, or using a small number of annotators. In this thesis, we explore the question of whether crowdsourced data that preserves disagreement can be used as ground truth for the task of relation classification in sentences. We investigate whether inter-annotator disagreement in particular is a useful signal that the relation classification model can learn from, and whether our crowdsourcing method can be scaled-up through a semi-supervised learning approach.

1.2.4 *Capturing Ambiguity*

Our work is part of a continuous effort in exploring the link between inter-annotator disagreement and ambiguity of the input data, as applied to a variety of tasks and domains.

In an experiment for crowdsourcing anaphora resolution, Poesio and Artstein [106] found that inter-annotator disagreement is linked to ambiguity in the text, and that directly asking the annotators to identify the ambiguous annotations is not enough to identify all the implicitly ambiguous cases. In assessing the OAEI benchmark, Cheatham and Hitzler [27] found that disagreement between annotators (both crowd and expert) is an indicator for inherent uncertainty in the domain knowledge, and that current benchmarks in ontology alignment and evaluation are not designed to model this uncertainty. Plank, Hovy, and Søgaard [105] found similar results for the task of crowdsourced part-of-speech tagging – most inter-annotator disagreement was indicative of debatable cases in linguistic theory, rather than faulty annotation. Bayerl and Paul [14] also investigate the role of inter-annotator disagreement as a possible indicator of ambiguity inherent in natural language. Chang, Amershi, and Kamar [23] found that ambiguous cases cannot simply be resolved by better annotation guidelines or through worker quality control. Across a series of textual annotation tasks, Chang, Lee-Goldman, and Tseng [24] found that the vast majority of annotators that disagree with the gold standard were correct in their assessment, either because the gold standard was faulty, or the task allowed for multiple correct answers.

Beyond text, studies in the annotation of music similarity [55], time series [113, 114], and medical images [31] have shown that disagreement between annotators can be an indicator for interesting properties of the data, such as ambiguity and uncertainty.

In most of these works, ambiguity is treated like a curious outlier, a property of the data that is unclear how it should be handled. We claim that ambiguity is an inherent part of natural language, and should be treated as such, by clearly defining it in ground truth corpora, and using it for training and evaluation of natural language processing models.

1.3 RESEARCH QUESTIONS & CONTRIBUTIONS

Based on the issues we identified in Section 1.2, the overall goal of this thesis is to *investigate the role of inter-annotator disagreement in crowdsourcing ground truth for natural language processing*, as collected using CrowdTruth methodology and metrics. The main research goal is addressed by answering the following research questions:

- **RQ1:** *Does allowing disagreement in crowdsourcing ground truth yield the same quality as asking domain experts?*

Chapter 2 explores this question for the task of medical relation extraction. In the medical domain it is typically assumed that expert annotators are required to get the best quality ground truth. This work shows that, by capturing the inter-annotator disagreement with the CrowdTruth method, medical relation classifiers trained on crowd annotations perform the same as those trained on expert annotations. Furthermore, classifiers trained on crowd annotations perform better than those trained with automatically-labeled data. Using the crowd also reduces the cost (monetary and in time required to find annotators) for collecting the data. This chapter is based on the following publication:

  – Dumitrache, Anca, Lora Aroyo, and Chris Welty. "Crowdsourcing ground truth for medical relation extraction." *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8.2 (2018): 12. [45]

- **RQ2:** *How does allowing disagreement in diverse crowdsourcing tasks influence the quality of the data?*

  Chapter 3 compares the quality of crowd data aggregated with CrowdTruth metrics and majority vote, a consensus - enforcing metric, over a diverse set of crowdsourcing tasks. We show that, by applying the CrowdTruth methodology, we collect richer data that allows us to reason about ambiguity of content. Furthermore, an increased number of crowd workers leads to growth and stabilization in the quality of annotations, going against the usual practice of employing a small number of annotators. This chapter is based on the following publication:

  – Dumitrache, Anca, et al. "Empirical methodology for crowdsourcing ground truth." *Semantic Web Journal (in publication)*. 2018. [49]

- **RQ3:** *Can we improve the performance of natural language processing models by using disagreement-aware ground truth data?*

  In Chapter 4 we discuss how CrowdTruth data can be used to better models for relation classification for sentences. We build on work from Chapter 2, where we have shown that training models on on crowd annotations gives better results than training with data automatically-labeled with distant supervision [93]. However, crowd data is expensive to collect. Chapter 4 describes how to correct a large corpus of training data for relation classification by using only a relatively small crowdsourced corpus, with two different methods: (1) by manually propagating the false positive and cross-relation signals identified with the help of the crowd, and (2) by adapting the semantic label propagation method [119] to work with CrowdTruth data. This chapter is based on the following publications:

  - Dumitrache, Anca, Lora Aroyo, and Chris Welty. "False positive and cross-relation signals in distant supervision data." *Proceedings of the Sixth Workshop on Automated Knowledge Base Construction (AKBC) at NIPS*. 2017. [42]
  - Dumitrache, Anca, Lora Aroyo, and Chris Welty. "Crowdsourcing semantic label propagation in relation classification." *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER) at EMNLP*. 2018. [46]

- **RQ4:** *Is inter-annotator disagreement an accurate indicator for ambiguity in natural language?*

  In Chapter 5, we explore this question as applied to the task of disambiguating semantic frames (i.e. high-level concepts that represent the meanings of words). Similarly to Chapter 2, we show that the crowd achieves comparative quality with domain experts. A qualitative evaluation of cases when crowd and expert disagree shows that inter-annotator disagreement is an indicator of ambiguity in both frames and sentences. We demonstrate that the cases in which the crowd workers could not agree exhibit ambiguity, either in the sentence, frame, or the task itself, arguing that collapsing such cases to a single, discrete truth value (i.e. correct or incorrect) is inappropriate, creating arbitrary targets for machine learning. This chapter is based on the following publication:

  - Dumitrache, Anca, Lora Aroyo, and Chris Welty. "Capturing ambiguity in crowdsourcing frame disambiguation." *Proceedings of the Sixth AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*. 2018. [43]
  - Dumitrache, Anca, Lora Aroyo, and Chris Welty. "A crowdsourced frame disambiguation corpus with ambiguity." In submission.

In addition to addressing these research question, another contribution of this thesis is a collection of ground truth datasets for the tasks of medical relation extraction [41], open domain relation extraction [44], and semantic frame disambiguation [47]. These datasets have been collected with crowdsourcing and processed with the CrowdTruth methodology. The disagreement-aware metrics have allowed us to label the data with continuous truth labels for sentences, relations, semantic frames and workers, allowing us to capture the ambiguity inherent in these tasks.

2

CROWDS VS. THE MEDICAL EXPERT

One is always wrong; but with two, truth begins. One cannot prove his case, but two are already irrefutable.

– Friedrich Nietzsche, The Gay Science

Natural language processing models require human labeled data for evaluation, and often for training. The standard practice used in gathering this data minimizes disagreement between annotators. This chapter investigates whether allowing disagreement in crowdsourcing ground truth can still yield quality of data comparable to that of experts, while accounting for the ambiguity inherent in language.

We have proposed the CrowdTruth method for collecting ground truth through crowdsourcing, that reconsiders the role of people in machine learning based on the observation that disagreement between annotators provides a useful signal for phenomena such as ambiguity in the text. We report on using this method to build an annotated data set for medical relation extraction for the *cause* and *treat* relations, and how this data performed in a supervised training experiment. We demonstrate that by modeling ambiguity, labeled data gathered from crowd workers can (1) reach the level of quality of domain experts for this task while reducing the cost, and (2) provide better training data at scale than distant supervision. We further propose and validate new weighted measures for precision, recall, and F-measure, that account for ambiguity in both human and machine performance on this task.

This chapter was published as *Crowdsourcing Ground Truth for Medical Relation Extraction* in the ACM Transactions on Interactive Intelligent Systems, and was co-authored by Lora Aroyo and Chris Welty [45].

2.1 INTRODUCTION

Many methods for Natural Language Processing (NLP) rely on *gold standard* annotations, or *ground truth*, for the purpose of training, testing and evaluation. In clinical NLP and other difficult domains, researchers assume that expert knowledge of the field is required from annotators. This means that, aside from the monetary costs of hiring humans to label data, simply finding suitable annotators bears a big time cost. The lack of annotated datasets for training and benchmarking is considered one of the big challenges of clinical NLP [26]. Understanding the role of people in machine learning is crucial in this context, as human annotation is considered the most reliable method for collecting ground truth. Because of this, in this chapter we tackle the question of whether *allowing disagreement in crowdsourcing ground truth yields the same quality as asking domain experts* (**RQ1**).

Disagreement in annotated data is typically considered to lower data quality, and is therefore removed from the ground truth. Data labeling is performed by humans, by reading text and following a set of guidelines to ensure a uniform understanding of the annotation task. It is assumed that the gold standard represents a universal and reliable model for language. However, Schaekermann et al. [113] and Bayerl and Paul [14] criticize this approach by investigating the role of inter-annotator disagreement as a possible indicator of ambiguity inherent in text. Previous experiments we performed in medical relation extraction [5] support this view by identifying two issues with the standard data labeling practice:

1. disagreement between annotators is usually eliminated through overly prescriptive annotation guidelines, thus creating artificial data that is neither general nor reflects the ambiguity inherent in natural language,

2. the process of acquiring ground truth by working exclusively with domain experts is costly and non-scalable, both in terms of time and money.

Ambiguity in text also impacts automated processes for extracting ground truth. Specifically, in the case of relation extraction from text, distant supervision [93, 126] is a well-established semi-supervised method that uses pairs of entities known to be related (e.g. from a knowledge base) to select sentences from a corpus that are used as positive training examples for the relations that relate the pairs. However, this approach is also prone to generating low quality training data, as not every mention of an entity pair in a sentence means a relation is also present. The problems are further compounded when dealing with ambiguous entities, or incompleteness in the knowledge base.

The goal of this chapter is to demonstrate that *preserving inter-annotator disagreement results in high quality ground truth data*, that is comparable to that of domain experts, and can be used as training data for NLP models. To capture inter-worker disagreement, we have proposed the *CrowdTruth* method for crowdsourcing training data for machine learning. We present an alternative approach for guiding supervised machine learning systems beyond the standard data labeling practice of a universal ground truth, by instead harnessing disagreement in crowd annotations to model the ambiguity inherent in text. We claim that, even for complex annotation tasks such as relation extraction, lack of domain expertise of the crowd is compensated by collecting a large enough set of annotations.

Previously, we studied medical relation extraction in a relatively small set of 90 sentences [6], comparing the results from the crowd with that of two expert medical annotators. We found that disagreement within the crowd is consistent with expert inter-annotator disagreement. Furthermore, sentences that registered high disagreement tended to be vague or ambiguous when manually evaluated. In this chapter, we build on these results by training a classifier for medical relation extraction with CrowdTruth data, and evaluating its performance. Our contributions are the following:

1. a comparison between using annotations from crowd and from medical experts to train a relation extraction classifier, showing that, with the processing of disagreement, *classifiers trained on crowd annotations perform the same as to those trained on expert annotations*;

2. a similar comparison between crowd annotations and distant supervision, showing that *classifiers trained on crowd annotations perform better than those trained on distant supervision*;

3. a *dataset of 3,984 English sentences for medical relation extraction*, centering on the *cause* and *treat* relations, that have been processed with disagreement analysis to capture ambiguity [41].

2.2 RELATED WORK

2.2.1 *Medical Crowdsourcing*

There exists some research using crowdsourcing to collect semantic data for the medical domain. Mortensen, Musen, and Noy [95] use crowdsourcing to verify relation hierarchies in biomedical ontologies. On 14 relations from the SNOMED CT CORE Problem List Subset, the authors report the crowd's accuracy at 85% for identifying whether the relations were correct or not. In the field of Biomedical NLP, Burger et al. [18] used crowdsourcing to extract the gene-mutation relations in Medical Literature Analysis and Retrieval System Online (MEDLINE) abstracts. Focusing on a very specific gene-mutation domain, the authors report a weighted accuracy of 82% over a corpus of 250 MEDLINE abstracts. Finally, Li, Good, and Su [83] performed a study exposing ambiguities in a gold standard for drug-disease relations with crowdsourcing. They found that, over a corpus of 60 sentences, levels of crowd agreement varied in a similar manner to the levels of agreement among the original expert annotators. All of these approaches present preliminary results from experiments performed with small datasets.

To our knowledge, the most extensive study of medical crowdsourcing was performed by Zhai et al. [131], who describe a method for crowdsourcing a ground truth for medical named entity recognition and entity linking. In a dataset of over 1,000 clinical trials, the authors show no statistically significant difference between the crowd and expert-generated gold standard for the task of extracting medications and their attributes. We extend these results by applying crowdsourcing to the more complex task of medical relation extraction, that *prima facie* seems to require more domain expertise than named entity recognition. Furthermore, we test the viability of the crowdsourced ground truth by training a classifier for relation extraction.

2.2.2 *Crowdsourcing Ground Truth*

Crowdsourcing ground truth has shown promising results in a variety of other domains. Snow et al. [116] have shown that aggregating the answers of an increasing number of unskilled crowd workers with majority vote can lead to high quality NLP training data. Hovy, Plank, and Søgaard [61] compared the crowd versus experts for the task of part-of-speech tagging. The authors also show that models trained based on crowdsourced annotation can perform just as well as expert-trained models. Kondreddi, Triantafillou, and Weikum [78] studied crowdsourcing for relation extraction in the general domain, comparing its efficiency to that of fully automated information extraction approaches. Their results showed the crowd was especially suited to identifying subtle formulations of relations that do not appear frequently enough to be picked up by statistical methods.

Other research for crowdsourcing ground truth includes: entity clustering and disambiguation [81], Twitter entity extraction [53], multilingual entity extraction and paraphrasing [28], and taxonomy creation [32]. However, all of these approaches rely on the assumption that one black-and-white gold standard must exist for every task. Disagreement between annotators is discarded by picking one answer that reflects some consensus, usually through using majority vote. The number of annotators per task is also kept low, between two and five workers, in the interest of reducing cost and eliminating disagreement. Whitehill et al. [129] and Welinder et al. [125] have used a latent variable model for task difficulty, as well as latent variables to measure the skill of each annotator, to optimize crowdsourcing for image labels. The novelty in our approach is to consider language ambiguity, and consequently inter-annotator disagreement, as an inherent feature of the language. Language ambiguity can be related to, but is not necessarily a direct cause of task difficulty. The metrics we employ for determining the quality of crowd answers are specifically tailored to measure ambiguity by quantifying disagreement between annotators.

2.2.3 *Disagreement and Ambiguity in Crowdsourcing*

In addition to our own work [5], the role of ambiguity when building a gold standard has previously been discussed by Lau, Clark, and Lappin [79]. The authors propose a method for crowdsourcing ambiguity in the grammatical correctness of text by giving workers the possibility to pick various degrees of correctness. However, inter-annotator disagreement is not discussed as a factor in measuring this ambiguity. After empirically studying part-of-speech datasets, Plank, Hovy, and Søgaard [105] found that inter-annotator disagreement is consistent across domains, even across languages. Furthermore, most disagreement is indicative of debatable cases in linguistic theory, rather than faulty annotation. It is not unreasonable to assume that these findings manifest even more strongly for NLP tasks involving semantic ambiguity, such as relation extraction.

In assessing the Ontology Alignment Evaluation Initiative (OAEI) benchmark, Cheatham and Hitzler [27] found that disagreement between annotators (both crowd and expert) is an indicator for inherent ambiguity of alignments, and that current benchmarks in ontology alignment and evaluation are not designed to model this ambiguity. Schaekermann et al. [113] propose a framework for dealing with uncertainty in ground truth that acknowledges the notion of ambiguity, and uses disagreement in crowdsourcing for modeling this ambiguity. To our knowledge, our work presents the first experimental results of using disagreement-aware crowdsourcing for training a machine learning system.

2.3 EXPERIMENTAL SETUP

The goal of our experiments is to assess the quality of our disagreement-aware crowdsourced data in training a medical relation extraction model. We use a binary classifier [124] that takes as input a set of sentences and two terms from the sentence, and returns a score reflecting the confidence of the model that a specific relation is expressed in the sentence between the terms. This manifold learning classifier was one of the first to accept weighted scores for each training instance, although it still requires a discrete positive or negative label. This property seemed to make it suitable for our experiments, as we expected the ambiguity of a sentence to impact its suitability as a training instance (in other words, we decreased the weight of training instances that exhibited ambiguity). We investigate the performance of the classifier over two medical relations: *cause* (between symptoms and disorders) and *treat* (between drugs and disorders).

The quality of the crowd data in training the classifier is evaluated in two parts: first by comparing it to the performance of an expert-trained classifier, and second with a classifier trained on distant supervision data. The training is done separately for each relation, over the same set of sentences, with different relation existence labels for crowd, expert and baseline.

2.3.1 *Data Selection*

The dataset used in our experiments contains 3,984 medical sentences extracted from PubMed article abstracts. The sentences were sampled from the set collected by [124] for training the relation extraction model that we are re-using. Wang & Fan collected the sentences with *distant supervision* [93, 126], a method that picks positive sentences from a corpus based on whether known arguments of the seed relation appear together in the sentence (e.g. the *treat* relation occurs between terms *antibiotics* and *typhus*, so find all sentences containing both and repeat this for all pairs of arguments that hold). The MetaMap parser [3] was used to recognize medical terms in the corpus, and the UMLS vocabulary [15] was used for mapping terms to categories, and relations to term types. The intuition of distant supervision is that since we know the terms are related, and they are in the same sentence, it is more likely that the sentence expresses a relation between them (than just any random sentence).

We started with a set of 12 relations important for clinical decision making, used also by Wang & Fan. Each of these relations corresponds to a set of UMLS relations (Tab.1), as UMLS relations are sometimes overlapping in meaning (e.g. *cause of* and *has causative agent* both map to *cause*). The UMLS relations were used as a seed in distant supervision. We focused our efforts on the relations *cause* and *treat*. These two relations were used as a seed for distant supervision in two thirds of the sentences of our dataset (1,043 sentences for *treat*, 1,828 for *cause*). The final third of the sentences were collected using the other 10 relations as seeds, in order to make the data more heterogeneous.

To perform a comparison with expert-annotated data, we randomly sampled a set of 975 sentences from the distant supervision dataset. This set restriction was done not just due to the cost of the experts, but primarily because of their limited time and availability. To collect this data, we employed medical students, in their third year at American universities, that had just taken United States Medical Licensing Examination (USMLE) and were waiting for their results. Each sentence was annotated by exactly one person. The annotation task consisted of deciding whether or not the UMLS seed relation discovered by distant supervision is present in the sentence for the two selected terms. The expert annotation costs are about $2.00 per sentence.

The crowdsourced annotation setup is based on our previous medical relation extraction work [7]. For every sentence, the crowd was asked to decide which relations (from Tab.1) hold between the two extracted terms. The task was multiple choice, workers being able

Relation	Corresponding UMLS Relation(s)	Definition	Example
treat	may treat	therapeutic use of a drug	penicillin treats infection
cause	cause of; has causative agent	the underlying reason for a symptom or a disease	fever induces dizziness
prevent	may prevent	preventative use of a drug	vitamin C prevents influenza
diagnoses	may diagnose	diagnostic use of an ingredient, test or a drug	RINNE test is used to diagnose hearing loss
location	disease has primary anatomic site; has finding site	body part in which disease or disorder is observed	leukemia is found in the circulatory system
symptom	disease has finding; disease may have finding	deviation from normal function indicating the presence of disease or abnormality	pain is a symptom of a broken arm
manifestation	has manifestation	links disorders to the observations that are closely associated with them	abdominal distention is a manifestation of liver failure
contraindicate	contraindicated drug	a condition for which a drug or treatment should not be used	patients with obesity should avoid using danazol
side effect	side effect	a secondary condition or symptom that results from a drug	use of antidepressants causes dryness in the eyes
associated with	associated with	signs, symptoms or findings that often appear together	patients who smoke often have yellow teeth
is a	is a	a relation that indicates that one of the terms is more specific variation of the other	migraine is a kind of headache
part of	part of	an anatomical or structural sub-component	the left ventricle is part of the heart

Table 1: Set of medical relations.

to choose more than one relation at the same time. There were also options available for cases when the medical relation was other than the ones we provided (*other*), and for when there was no relation between the terms (*none*). The crowdsourcing was run on the Figure Eight[1] (formerly known as CrowdFlower) platform, with 15 workers per sentence, at a cost of $0.66 per sentence. Compared to a single expert judgment, the cost per sentence of the crowd amounted to 2/3 of the sum paid for the experts.

1 `https://www.figure-eight.com/`

All of the data that we have used, together with the templates for the crowdsourcing tasks, and the crowdsourcing implementation details are available online [41].

2.3.2 *CrowdTruth Metrics*

The crowd output was processed with the use of CrowdTruth metrics – a set of general-purpose crowdsourcing metrics [65], that have been successfully used to model ambiguity in annotations for relation extraction, event extraction, sounds, images, and videos [7]. These metrics model ambiguity in semantic interpretation based on the triangle of reference [99], with the vertices being the input sentence, the worker, and the seed relation. Ambiguity and disagreement at any of the vertices (e.g. a sentence with unclear meaning, a poor quality worker, or an unclear relation) will propagate in the system, influencing the other components. For example, if a sentence is unclear, we expect workers will be more likely to disagree with each other; if a worker is not doing a good job, we expect that worker to disagree with other workers across the majority of the sentences they worked on; and if a particular target relation is unclear, we expect workers to disagree on the application of that relation across all the sentences. By using multiple workers per sentence and requiring each worker to annotate multiple sentences, the aggregate data helps us isolate these individual signals and how they interact. Thus a high quality worker who annotates a low clarity sentence will be recognized as high quality. In our workflow, these metrics are used both to eliminate spammers, as detailed by [7], and to determine the clarity of the sentences and relations. The main concepts are:

- *annotation vector:* used to model the annotations of one worker for one sentence. For each worker i submitting their solution to a task on a sentence s, the vector $W_{s,i}$ records their answers. If the worker selects a relation, its corresponding component would be marked with '1', and '0' otherwise. The vector has 14 components, one for each relation, as well as *none* and *other*. Multiple choices (e.g. picking multiple relations for the same sentence) are modeled by marking all corresponding vector components with '1'.

- *sentence vector:* the main component for modeling disagreement. For every sentence s, it is computed by adding the annotation vectors for all workers on the given task: $V_s = \sum_i W_{s,i}$. One such vector was calculated for every sentence.

- *sentence-relation score:* measures the ambiguity of a specific relation in a sentence with the use of cosine similarity. The higher the score, the more clearly the relation is expressed in the sentence.

The sentence-relation score is computed as the cosine similarity between the sentence vector and the unit vector for the relation: $srs(s,r) = cos(V_s, \hat{r})$, where the unit vector $\hat{r}$ refers to a vector where the component corresponding to relation r is equal to '1', and all other components are equal to '0'. The reasoning is that the unit vector $\hat{r}$ corresponds to the clearest representation of a relation in a sentence – i.e. when all workers agree that relation r exists between the seed terms, and all other relations do not exist. As a cosine similarity, these scores are in the $[0,1]$ interval. Tab.2 shows the transformation of sentence vectors to the sentence-relation scores and then to the training scores using the threshold below.

- *sentence-relation score threshold:* a fixed value in the interval $[0,1]$ used to differentiate between a negative and a positive label for a relation in a sentence. Given a value t for the threshold, all sentences with a sentence-relation score less than t get a negative label, and the ones with a score greater or equal to t are positive. The results section compares the performance of the crowd at different threshold values. This threshold was necessary because our classifier required either a positive or negative label for each training example. Therefore, the sentence-relation scores must be re-scaled in the $[-1,0]$ interval for negative labels. An example of how the crowd scores for training the model were calculated is given in Tab.2.

2.3.3 *Training the Model*

The sentences together with the relation annotations were then used to train a manifold model for relation extraction [124]. This model was developed for the medical domain, and tested for the relation set that we employ. It is trained per individual relation, by feeding it both *positive* and *negative* data. It offers support for both discrete labels, and real values for weighting the confidence of the training data entries, with positive values in $(0,1]$, and negative values in $[-1,0)$. Using this system, we train several models using five-fold cross validation, in order to assess the performance of the crowd dataset. The training was done separately for the *treat* and *cause* relations. For each relation, we constructed four datasets, with the same sentences and term pairs, but with different labels for whether or not the relation is present in the sentence:

1. *baseline:* The distant supervision data is used to provide discrete (positive or negative) labels on each sentence - i.e. if a sentence contains two terms known (in UMLS) to be related by *treats*, the sentence is considered positive. Distant supervision does not

Relation	Sentence vector		Sentence-relation score		Crowd score used in model training	
	Sent.1	*Sent.2*	*Sent.1*	*Sent.2*	*Sent.1*	*Sent.2*
treat	0	3	0	0.36	-1	-0.64
prevent	0	1	0	0.12	-1	-0.88
diagnose	1	7	0.09	0.84	-0.91	0.84
cause	10	0	0.96	0	0.96	-1
location	1	0	0.09	0	-0.91	-1
symptom	2	0	0.19	0	-0.81	-1
manifestation	0	0	0	0	-1	-1
contraindicate	0	0	0	0	-1	-1
associated with	1	3	0.09	0.36	-0.91	-0.64
side effect	0	0	0	0	-1	-1
is a	0	0	0	0	-1	-1
part of	0	0	0	0	-1	-1
other	0	1	0	0.12	-1	-0.88
none	0	0	0	0	-1	-1

Table 2: Given two sentences, *Sent.1* and *Sent.2*, with term pairs in bold font, the table shows the transformation of the sentence vectors to sentence – relation scores, and then to *crowd* scores used for model training. The sentence-relation threshold for the train score is set at 0.5 for these examples.

Sent.1: **Renal osteodystrophy** is a general complication of chronic renal failure and **end stage renal disease**.
Sent.2: If **TB** is a concern, a **PPD** is performed.

extract negative examples, so in order to generate a negative set for one relation, we use positive examples for the other (non-overlapping) relations shown in Tab. 1. This dataset constitutes the baseline against which all other datasets are tested.

2. *expert:* Discrete labels based on an expert's judgment as to whether the *baseline* label is correct. The experts do not generate judgments for all combinations of sentences and relations – for each sentence, the annotator decides on the seed relation extracted with distant supervision. Similarly to the baseline data, we reuse positive examples from the other relations to increase the number of negative examples.

3. *single:* Discrete labels for every sentence are taken from one randomly selected crowd worker who annotated the sentence. This data simulates the traditional single annotator setting common in annotation environments.

4. *crowd:* Weighted labels for every sentence are based on the CrowdTruth *sentence-relation score*. Labels are separated into a positive and negative set based on the *sentence-relation score thresh-*

old, and negative labels are rescaled in the $[-1, 0]$ interval. An example of how the scores were processed is given in Tab.2.

For each relation, two experiments were run. First, we performed a comparison between the *crowd* and *expert* datasets by training a model using the subset of sentences that also has expert annotations. In total there are 975 unique sentences in this set. After we were able to determine the quality of the *crowd* data, we performed a second experiment comparing the performance of the classifier when trained with the *crowd* and *baseline* annotations from the full set of 3,984 sentences.

2.3.4 *Evaluation Data*

In order for a meaningful comparison between the crowd and expert models, the evaluation set needs to be carefully vetted. For each of the relations, we started by selecting the positive/negative threshold for *sentence-relation score* such that the crowd agrees the most with the experts. We assume that, if both the expert and the crowd agree that a sentence is either a positive or negative example, it can automatically be used as part of the test set. Such a sentence was labeled with the expert score.

The interesting cases appear when crowd and expert disagree. To ensure a fair comparison, our team adjudicated each of them to decide whether or not the relation is present in the sentence. The sentences where no decision could be reached were subsequently removed from the evaluation. There were 32 such sentences for *cause* (18 with negative expert labels, and 14 with positive), and 15 for *treat* (all for positive expert labels). Table 5 in the Appendix shows some example sentences that were removed from the evaluation set. This set constitutes of confusing and ambiguous sentences that our team could not agree on. Often these sentences contained a vague association between the two terms, but the relation was too broad to label it as a positive classification example. However, because a relation is nevertheless present, these sentences cannot be labeled as negative examples either. Eliminating these sentences is a disadvantage to a system like ours which was motivated specifically by the need to handle such cases, however the scientific community still only recognizes discrete measures such as precision and recall, and we felt it only fair to eliminate the cases where we could not agree on the correct way to map ambiguity into a discrete score.

For evaluation, we selected sentences through 5-fold cross-validation, but we obviously only used the test labels when a partition was chosen to be test. For the second evaluation over 3,984 sentences, we again selected test sets using cross-validation over the sentences with expert annotation, adding the unselected sentences with their training labels to the training set. This allows us to directly compare the learning curves between the 975 and 3,984 sentences experiments. The scores reported are the mean over the cross-validation runs.

2.3.5 *CrowdTruth-Weighted Evaluation*

We also explored how to incorporate CrowdTruth into the evaluation process. The reasoning of our approach is that the ambiguity of a sentence should also be accounted for in the evaluation – i.e. sentences that do not clearly express a relation should not count for as much as clear sentences. In this case, the *sentence-relation score* gives a real-valued score that measures the degree to which a particular sentence expresses a particular relation between two terms. Therefore, we propose a set of evaluation metrics that have been weighted with the *sentence-relation score* for a given relation. The metrics have been previously tested on a subset of our ground truth data, as detailed in [40].

We collect true and false positives and negatives in the standard way, such that $tp(s) = 1$ iff s is a true positive, and 0 otherwise, similarly for fp, tn, fn. The positive sentences (i.e true positive and false negative labels) are weighted with the sentence-relation score $srs(s)$ for the given sentence-relation pair, i.e. how likely it is that the relation is expressed in the sentence. Negative sentences (true negative and false positive labels) are weighted with $1 - srs(s)$, how likely it is that that the sentence does not express the relation. Based on this, we define the following metrics to be used in the evaluation:

- *weighted precision:* Where normally $P = tp/(tp + fp)$, weighted precision
$$P' = \frac{\sum_s srs(s) \cdot tp(s)}{\sum_s srs(s) \cdot tp(s) + (1 - srs(s)) \cdot fp(s)};$$
- *weighted recall:* Where normally $R = tp/(tp + fn)$, weighted recall
$$R' = \frac{\sum_s srs(s) \cdot tp(s)}{\sum_s srs(s) \cdot tp(s) + srs(s) \cdot fn(s)};$$
- *weighted F-measure:* Is the harmonic mean of weighted precision and recall:
$$F1' = 2P'R'/(P' + R').$$

2.4 RESULTS

2.4.1 *CrowdTruth vs. Medical Experts*

In the first experiment, we compare the quality of the crowd with expert annotations over the sentences that have been also annotated by experts. We start by comparing the crowd and expert labels to the adjudicated test labels on each sentence, without training a classifier, computing an F1 score that measures the *annotation quality* of each set, shown in Figure 1. Since the baseline, expert, and single sets are binary decisions, they appear as horizontal lines, whereas the crowd annotations are shown at different sentence-relation score thresholds. For both relations, the crowd labels have the highest annotation quality F1 scores, 0.907 for the *cause* relation, and 0.966 for *treat*. The expert data is close behind, with an F1 score of 0.844 for *cause* and 0.912 for *treat*. To calculate the statistical significance of the results, we used McNemar's test [89] over paired nominal data, by constructing a contingency table from the binary classification results (i.e. correct or incorrect classification) of paired datasets (e.g. crowd and expert). This difference between crowd and expert is not significant for *cause* ($p > 0.5$, $\chi^2 = 0.034$), and significant for *treat* ($p = 0.002$, $\chi^2 = 5.127$). The sentence – relation score threshold for the best annotation quality F1 is also the threshold where the highest agreement between crowd and expert occurs (Figure 2).

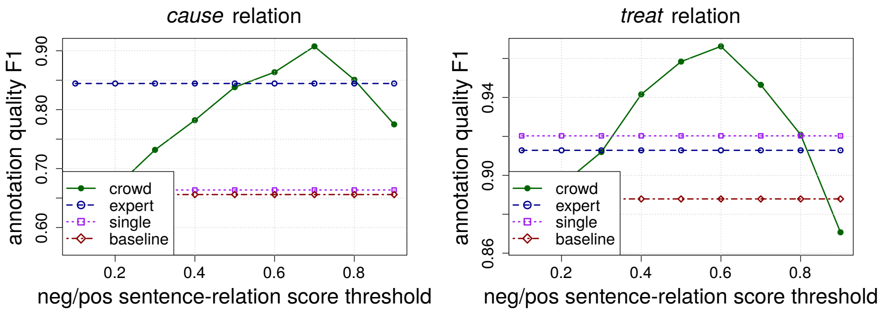


Figure 1: Annotation quality F1 scores.

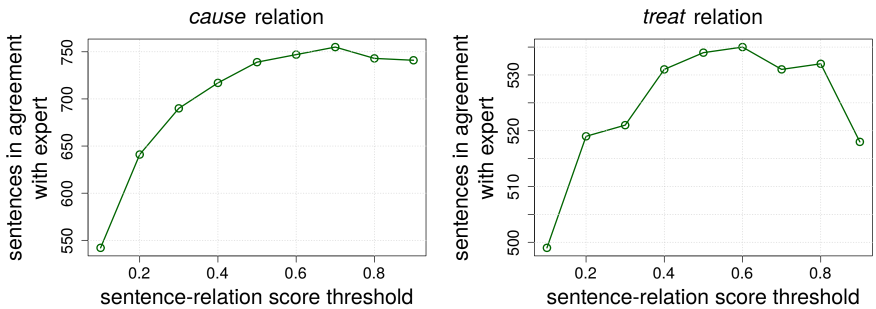


Figure 2: Crowd & expert agreement.

Next we compare the quality of the crowd and expert annotations by training the relation extraction model using each dataset. For the *cause* relation, the results of the evaluation (Fig.3) show the best performance for the crowd-trained model when the sentence-relation threshold is 0.5. Trained with this data, the classifier model achieves an F1 score of

0.642, compared to the expert-trained model which reaches 0.638. The difference is statistically significant with $p = 0.016$ ($\chi^2 = 5.789$).

Tab.3 shows the full results of the evaluation, together with the results of the CrowdTruth weighted metrics (P′, R′, F1′). In all cases, the F1′ score is greater than F1, indicating that ambiguous sentences have a strong impact on the performance of the classifier. Weighted P′ and R′ also have higher values in comparison with simple precision and recall.

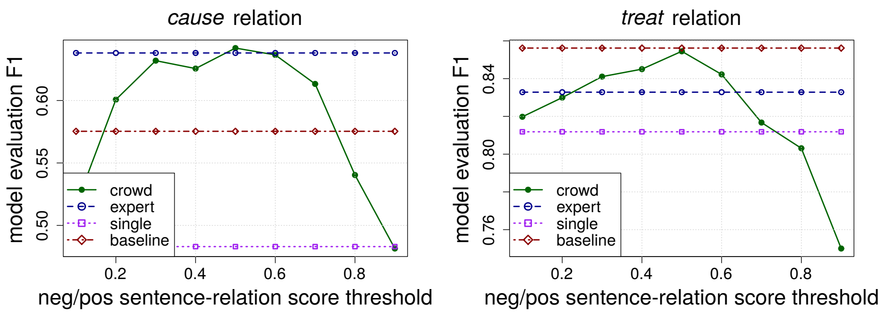


Figure 3: Model testing F1 scores.

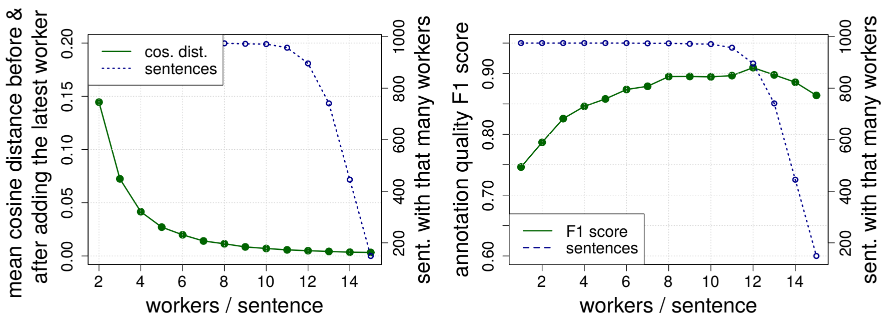
	Dataset	P	P′	R	R′	F1	F1′
	crowd	0.565	0.632	0.743	0.754	**0.642**	**0.687**
cause	*expert*	**0.672**	**0.711**	0.604	0.616	0.638	0.658
relation	*baseline*	0.436	0.474	**0.844**	**0.842**	0.575	0.606
	single	0.495	0.545	0.473	0.478	0.483	0.658
	crowd	0.823	0.843	0.891	0.902	0.854	0.869
treat	*expert*	**0.834**	**0.863**	0.833	0.84	0.832	0.85
relation	*baseline*	0.767	0.811	**0.968**	**0.968**	**0.856**	**0.882**
	single	0.774	0.819	0.856	0.866	0.811	0.84

Table 3: Model evaluation results over sentences with expert annotation. Crowd scores are shown at 0.5 negative/positive sentence-relation score threshold.

For the *treat* relation, the results of the evaluation (Figure!3) shows baseline as having the best performance, at an F1 score of 0.856. The crowd dataset, with an F1 score of 0.854, still out-performs the expert, scoring at 0.832. These three scores are not, however, significantly different ($p > 0.5$, $\chi^2 = 0.453$), as there are so few actual pairwise differences (a consequence of the higher scores and the size of the dataset).

For both *cause* and *treat* relations, the single annotator dataset performed the worst. It is also worth noting that the sentence – relation score threshold for the best classifier performance (0.5 for both relations) is different from the threshold for best annotation quality, and highest agreement with expert (0.7 for *cause* and 0.6 for *treat*, Figure 1).

Finally, we checked whether the number of workers per task was sufficient to produce a stable sentence-relation score. We did this in two ways, first by measuring the cosine distance between the sentence vectors at each incremental number of workers (Fig. 4a), and second by measuring the annotation quality F1 score for *treat* and *cause*, combined using the micro-averaged method (i.e. adding up the individual true positives, false positives etc.), against the number of workers annotating each sentence (Fig. 4b). For both plots, the workers were added in the order that they submitted their results on the crowdsourcing platform. Based on these results, we decided to ensure that each sentence has been annotated by at least 10 workers after spam removal. The plot of the mean cosine distance between sentence vectors before and after adding the latest worker shows that the sentence vector is stable at 10 workers. The annotation quality F1 score per total number of workers plot appears less stable in general, with a peak at 12 workers, and a subsequent drop due to sparse data – only 149 sentences had 15 or more total workers. However, after 10 workers there are no significant increases in the annotation quality. While it can be argued that both plots stabilize for a lower number of workers, we picked 10 as a threshold because it gives some room for improvement for sentences that might need more workers before getting a stable score, while still being economical.



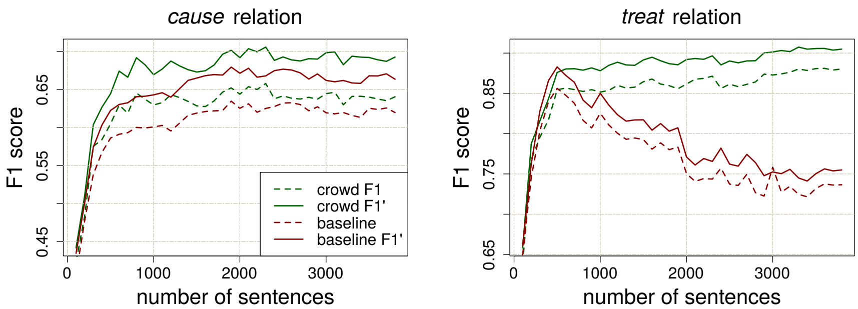
(a) Mean cosine distance for sentence vectors before and after adding the latest worker, shown per number of workers.

(b) Combined annotation quality F1 for *cause* and *treat* crowd, at their best pos./neg. thresholds (Figure 1).

Figure 4: Optimal number of workers analysis.

2.4.2 *CrowdTruth vs. Distant Supervision*

Distant supervision is a widely used technique in NLP, because its obvious flaws can be overcome at scale. We did not have enough time with the experts to gather a larger dataset from them, but the crowd is always available, so after we determined that the performance of the crowd matched the medical experts, we extended the experiments to 3,984 sentences. The crowd dataset in this experiment uses a fixed sentence-relation score threshold equal to 0.5, since this is the value where the crowd performed the best in the previous experiment, for both of the relations. As in the previous experiment, we employed five-fold cross validation to train the model. The test sets were kept the same as in the previous experiment, using the test partition labels as a gold standard. The goal was to compare the crowd to the distant supervision baseline, while scaling the number of training examples, until achieving a stable learning curve in the F1 score. Since the single annotator dataset performed badly in the initial experiment, it was dropped from this analysis. The full results of the experiment are available in Tab.4.



(a) *cause* relation (b) *treat* relation

Figure 5: Learning curves.

	Dataset	P	P′	R	R′	F1	F1′
cause	*crowd*	**0.538**	**0.61**	0.79	0.802	**0.64**	**0.692**
relation	*baseline*	0.475	0.53	**0.889**	**0.887**	0.619	0.663
treat	*crowd*	**0.876**	**0.913**	**0.887**	**0.898**	**0.88**	**0.904**
relation	*baseline*	0.808	0.858	0.678	0.673	0.736	0.754

Table 4: Model evaluation results over 3,984 sentences. Crowd scores are shown at 0.5 sentence-relation score threshold.

For both relations, the crowd consistently performs better than the baseline. In the case of the *cause* relation, crowd and baseline perform closer to each other, with an F1 score of 0.64 for crowd and 0.619 for baseline. This difference is significant with $p = 0.001$ and $\chi^2 = 10.028$. The gap in performance is even greater for accuracy, where the crowd model scored at 0.773 and baseline at 0.705. The learning curves for the *cause* relation (Fig.5a) show both datasets achieve stable performance.

For the *treat* relation, the crowd scores an F1 of 0.88, while baseline scores 0.736, with $p = 1.39 \times 10^{-10}$ significance, and $\chi^2 = 41.176$. The learning curves (Fig.5b) show that, while baseline out-performed crowd when training with less than 1,000 sentences, crowd performance became stable after 1,000, while baseline went down, significantly increasing the gap between the two datasets.

The gap in performance is also present in the weighted F1′ metrics. As is the case in the previous experiment, the F1′ scores higher than the regular F1 score for both crowd and baseline. The only weighted metric that does not increase is the baseline recall. This is also the only metric by which the baseline model performed better than the crowd.

2.5 DISCUSSION

2.5.1 *CrowdTruth vs. Medical Experts*

Our first goal was to demonstrate that, like the crowdsourced medical entity recognition work by [131], the CrowdTruth approach of having multiple annotators with precise quality scores can be harnessed to create gold standard data with a quality that rivals annotated data created by medical experts. Our results show this clearly, in fact with slight improvements, with a sizable dataset (975 sentences) on a problem (relation extraction) that *prima facie* seems to require more domain expertise (than entity recognition).

The most interesting result of the first experiment is that the sentence-relation score threshold that gives the best F1 score is the same for both *cause* and *treat* relations (Figure 3), at a value of 0.5. This shows that ambiguous data is indeed valuable in training of clinical NLP models, and that being too strict with what constitutes a positive (or negative) training example produces flawed ground truth data. It is also worth noting that the single crowd annotator performs the worst for each of the relations. This could be further indication that the crowd can only achieve quality when accounting for the choices of multiple annotators, and further calls into question the standard practice of using only one annotator per example.

A curious aspect of the results is that the sentence-relation score threshold that gives the highest annotation quality F1 score (i.e. F1 score calculated directly over the test data, without training the model), shown in Figure 1, is different from the best threshold for classifier performance (Figure 3). It is the lower threshold (equal to 0.5) that results in the best model. This is most likely due to the higher recall of the lower threshold, which exposes the classifier to more positive examples. F-score is the harmonic mean between precision and recall, and does not necessarily represent the best trade-off between them, as this experiment shows for annotation quality. Indeed F-score may not be the best trade-off between precision and recall for the classifier, either, but it is the most widely accepted and reported metric for relation extraction. Note also that for both relations, the annotation quality at the 0.5 threshold is comparable or better than expert annotation quality.

The fact that the experts performed slightly worse than the single crowd annotator on the *treat* annotation quality (Fig.1) is also a surprising finding. Although the difference is too small to draw significant conclusions from, it indicates that the *treat* relation was easier to interpret by the crowd and generated less disagreement – the single annotator had a better performance for *treat* than for *cause* also in the model evaluation (Fig.3). This result also shows that the experts we employed were fallible, and made mistakes when annotating the data. A better approach to gather the expert annotations would be to ask several experts per sentence, to account for the failures in a single person's interpretations.

In our error analysis of the annotation quality, we found that (as Figure 1 shows) experts and the crowd both make errors, but of different kinds. Experts tend to see relations that they know hold as being expressed in sentences, when they are not. For example, in, "He was the first to describe the relation between **Hemophelia** and **Hemophilic Arthropathy**," experts labeled the sentence as expressing the *cause* relation, since they know Hemophelia causes Hemophilic Arthropathy. Thus they are particularly prone to errors in sentences selected by distant supervision, since that is the selection criterion. Table 6 from the Appendix shows more such examples. Crowd workers, on the other hand, were more easily fooled by sentences that expressed one of the target relations, but *not between the selected arguments.* For example, in "Influenza treatments such as **antivirals** and **antibiotics** are sometimes recommended," some crowd workers will label the sentence with treats, even though we are looking for the relation between *antivirals* and *antibiotics*. More such examples are shown in Table 7 from the Appendix. The crowd achieves overall higher annotation quality due to redundancy, over the set of 15 workers, it is unlikely they will all make the same mistake.

In Figs. 4a & 4b we observe that we need at least 10 workers to get a stable crowd score. This result goes against the general practice for building a ground truth, where per task there usually are 1 to 5 annotators. Based on our results, we believe that the general practice is not applicable for the use case of medical relation extraction, and should perhaps be reconsidered for other annotation use cases where ambiguity can be present, as outside of a few clear cases, the input of more annotators per task can be very useful at indicating the ambiguities inherent in language, as well as all other interpretation tasks (e.g. images, audio, event processing, etc.). Even with this added requirement, we found that crowd data is still cheaper to acquire than annotation from medical experts, as the crowd is both cheap (the cost of the crowd was $\frac{2}{3}$ that of the expert) and always available via dedicated crowdsourcing platforms like Figure Eight.

A bottleneck in this analysis is the availability of expert annotations – we did not have the resources to collect a larger expert dataset, and this indeed is the main reason to consider crowdsourcing. In this context, the real value of distant supervision is that large amounts of data can be gathered rather easily and cheaply, since humans are not involved. Therefore, the goal of the second experiment was to explore the trade-off between quality and cost of crowdsourcing compared to distant supervision, while scaling up the model to reach its maximum performance.

2.5.2 *CrowdTruth vs. Distant Supervision*

The results for both relations (Fig.5a & Fig.5b) show that the crowd does out-perform the distant supervision baseline after the learning curves have stabilized, thus justifying its cost. From this we infer that not only is the crowd generating higher quality data than the automated baseline, but training the model with weights, as opposed to binary labels, does have a positive impact on the performance of the model.

The results of the CrowdTruth weighted F1′ consistently scored above the simple F1, for both baseline and crowd over both relations. This consolidates our assumption that ambiguity does have an impact on classifier performance, and weighting test data with ambiguity can account for this hidden variable in the evaluation.

The only weighted metric without a score increase is the baseline R′ for the *cause* relation (see Tab.4). Recall is also the only un-weighted metric for which the *cause* baseline model performed better than the crowd. Recall is inversely proportional to the number of false negatives, indicating that distant supervision, for this relation, is finding more positives at the expense of incorrectly labeling some of them. This appears to be a consequence of how the model performs its training – one of the features it learns is the UMLS type of the terms. For the *cause* relation, it seems that term types are often enough to accurately classify a positive example (e.g. an anatomical component will rarely be the effect of a causal relation).

Over-fitting on term types classification could also be the reason that baseline performs better than the crowd in the initial experiment for *treat* (Tab.3), where recall for baseline is unusually high. *treat* is also a relation that appears to favor a high recall approach – there are very few negative examples where the type constraint of the terms (drug - disease) is satisfied. In previous work [7] we observed that *treat* generates less ambiguity than *cause*, which explains why *treat* has overall higher F1 scores than *cause* in all datasets. However, the high F1 scores could also make the models for *treat* more sensitive to confusion from ambiguous examples, as a small number of confusing sentences would be enough to decrease such a high performance. Indeed, as more (potentially ambiguous) examples appear in the training set, both the F1 and the recall of the baseline for *treat* drop, while the crowd scores remain consistent (Fig.5b). This result emphasizes the importance of weighting training data with ambiguity, as a few ambiguous examples seem to have a strong impact in generating false negatives during classification.

Our experiment has two limitations: (1) because of the limited availability of domain experts, we could not collect more than one expert judgment per sentence, and (2) because the model used classifies data with either a positive or a negative label, we removed the examples from the evaluation set that could not fit into either label. We expect that adding more expert annotators per sentence will result in better quality annotations. However, disagreement will likely still be present – as indicated by our previous work [5] on a set of 90 sentences, two experts agreed only 30% of the time over what the correct relation is. Future work could explore whether disagreement between experts is consistent with the crowd disagreement. The second limitation lies with evaluation measures such as precision and recall that require discrete labels, which are the standard for classification models. The CrowdTruth method was designed specifically to represent ambiguous cases that are more difficult to fit into a positive or negative label, but to evaluate it in comparison with discrete data, we had to use the standard metrics. Now that we have shown the quality of the crowd data, it can be used to perform more detailed evaluations that take ambiguity into account through the use of weighted precision, recall and F1.

2.6 CONCLUSION

The standard data labeling practice used in supervised machine learning attempts to minimize disagreement between annotators, and therefore fails to model the ambiguity inherent in language. We propose the CrowdTruth method for collecting ground truth through crowdsourcing, that reconsiders the role of people in machine learning based on the observation that disagreement between annotators can signal ambiguity in the text.

In this work, we used CrowdTruth to build a gold standard of 3,984 sentences for medical relation extraction, focusing on the *cause* and *treat* relations, and used the crowd data to train a classification model. We have shown that, with the processing of ambiguity, *the crowd performs just as well as medical experts in terms of the quality and efficacy of annotations*, while being cheaper and more readily available. In addition, our results show that, when the model reaches maximum performance after training, the crowd also performs better than distant supervision. Finally, we introduced and validated new weighted measures for precision, recall, and F-measure, that account for ambiguity in both human and machine performance on this task. These results encourage us to continue our experiments by replicating this methodology for an increasing set of relations in the medical domain.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Chang Wang for support with using the medical relation extraction classifier, and Anthony Levas for help with collecting the expert annotations. The authors, Dr. Wang and Mr. Levas were all employees of IBM Research when the expert data collection was performed, and we are grateful to IBM for making the data freely available subsequently.

APPENDIX: EXAMPLE SENTENCES FROM THE EVALUATION SET

Sentence	Relation	Crowd Label	Expert Label
The physician should ask about a history of **diabetes** of long duration, including other manifestations of **diabetic neuropathy**.	*cause*	0.977	-1
If the oxygen is too low, the incidence of **decompression sickness** increases; if the **oxygen** is too high, oxygen poisoning becomes a problem.	*cause*	0.743	-1
Evidence: ? Vigilant intraoperative magement of **hypertension** is essential during resection of **pherochromocytoma**.	*cause*	-0.651	1
This is the first case of **Aicardi Syndrome** associated with lipoma and **metastatic angiosarcoma**.	*cause*	-0.909	1
Will giving **Acetaminophen** prevent the **pain** of the immunization?	*treat*	0.995	-1
FDA approves **Thalidomide** for **Hansen's disease** side effect, imposes unprecedented restrictions on distribution.	*treat*	0.913	-1

Table 5: Example sentences removed from the evaluation (term pairs in bold font).

Sentence	Relation	Crowd Label	Expert Label
Patients with a history of **bee sting allergy** may have a higher risk of a **hypersensitivity reaction** with paclitaxel treatment.	*cause*	0.9	-1
In contrast, we did not find a definite increase in the LGL percentage within 6 months postpartum in patients with **Grave's disease** who relapsed into **Grave's thyrotoxicosis**.	*cause*	0.737	-1
Hepatoma in one patient was correctly identified by both methods, as well as the presence of **ascites**.	*cause*	-0.579	1
The diagnosis of **gyrate atrophy** was confirmed biochemically and clinically; **hyperornithinemia** and a deficiency of ornithine ketoacid transamise were confirmed biochemically.	*cause*	-0.863	1
Thirdly the evidence of the efficacy of **Clomipramine** in **OCD without concomitant depression** reported by Montgomery 1980 and supported by other studies suggests that 5 HT uptake inhibitors have a specifically anti obsessiol effect.	*treat*	0.905	-1
The 1 placebo controlled trial that found black cohosh to be effective for **hot flashes** did not find **estrogen** to be effective, which casts doubt on the study's validity.	*treat*	0.73	-1
Graft Versus Host Disease (GVHD) Prophylaxis was methotrexate (1 patient), cyclosporine (2 patients), methotrexate + **cyclosporine** (3 patients), cyclosporine + physical removal of T cells (2 patients).	*treat*	-0.657	1
Patients with severe forms of **Von Willebrands' Disease (VWD)** may have frequent haemarthroses, especially when **Factor VIII (FVIII)** levels are below 10 U/dL, so that some of them develop target joints like patients with severe haemophilia A.	*treat*	-1	1

Table 6: Example sentences where the expert was wrong (term pairs in bold font).

SENTENCE	RELATION	CROWD LABEL	EXPERT LABEL
Instability of **bone** fragments is regarded as the most important factor in pathogenesis of **pseudoarthrosis**.	*cause*	0.928	-1
Atopic conditions include allergic rhinitis, atopic eczema, **allergic conjunctivitis** and asthma.	*cause*	0.507	-1
The histological finding of **Psammoma bodies** is important in the diagnosis of **duodel stomatostatinomas**.	*cause*	-0.558	1
A retrospective review of 64 patients with haematuria and subsequent histologically proven **carcinoma of the bladder** revealed that bladder tumours could be diagnosed pre operatively in 34 of 46 (76%) of patients with gross **haematuria** and 12 of of 18 (67%) of those with microhaematuria.	*cause*	-0.658	1
Hypersecretion of insulin increases the chance of the incidence of **diabetes type I and II** while inhibiting **insulin** secretion helps prevent diabetes.	*treat*	0.949	-1
To determine whether late asthmatic reactions and the associated increase in airway responsiveness induced by toluene diisocyate (TDI) are linked to **airway inflammation** we investigated whether they are inhibited by **Prednisone**.	*treat*	0.52	-1
In one group of four pigs sensitive to **Malignant Hyperthermia (MHS)** a dose response to **intravenous Dantrolene** was determined by quantitation of toe twitch tension..	*treat*	-0.575	1
Deficiency diseases include night blindness and keratomalacia (caused by lack of vitamin A); beriberi and polyneuritis (lack of thiamine); **pellagra** (lack of **niacin**); scurvy (lack of vitamin C); rickets and osteomalacia (lack of vitamin D); pernicious anemia (lack of gastric intrinsic factor and vitamin B 12.	*treat*	-1	1

Table 7: Example sentences where the crowd was wrong (term pairs in bold font).

3 DATA QUALITY FROM DISAGREEMENT

Whenever you find yourself on the side of the majority, it is time to reform (or pause and reflect).

– Mark Twain, NOTEBOOK, 1904

The process of gathering ground truth data through human annotation is a major bottleneck in the use of information extraction methods for populating the Semantic Web. Crowdsourcing-based approaches are gaining popularity in the attempt to solve the issues related to volume of data and lack of annotators. Typically these practices use inter-annotator agreement as a measure of quality. However, in many domains, such as event detection, there is ambiguity in the data, as well as a multitude of perspectives of the information examples. While in Chapter 2 we have shown that the preservation of inter-annotator disagreement still results in high quality ground truth data, in this chapter we investigate more concretely how does disagreement influences data quality across a variety of tasks, as compared with voting aggregation methods that only take into account the opinion of the majority.

We present an empirically derived methodology for efficiently gathering of ground truth data in a diverse set of use cases covering a variety of domains and annotation tasks. Central to our approach is the use of CrowdTruth metrics that capture inter-annotator disagreement. We show that measuring disagreement is essential for acquiring a high quality ground truth. We achieve this by comparing the quality of the data aggregated with CrowdTruth metrics with majority vote, over a set of diverse crowdsourcing tasks: Medical Relation Extraction, Twitter Event Identification, News Event Extraction and Sound Interpretation. We also show that an increased number of crowd workers leads to growth and stabilization in the quality of annotations, going against the usual practice of employing a small number of annotators.

This chapter will appear in publication as *Empirical Methodology for Crowdsourcing Ground Truth* in the Semantic Web Journal and was co-authored by Oana Inel, Benjamin Timmermans, Carlos Ortiz, Robert-Jan Sips, Lora Aroyo and Chris Welty [49].

3.1 INTRODUCTION

Knowledge base curation, or the task of populating knowledge bases, is one of the main research challenges of crowdsourcing the Semantic Web [112]. Knowledge base curation can be done either manually, by asking annotators to populate the knowledge graph by manually extracting triples from unstructured data, or automatically by using information extraction methods that are trained and evaluated on ground truth collected from human annotators. In both cases, the process of gathering the human annotations is the a bottleneck in the entire knowledge base population process. The traditional approach to gathering human annotation is to employ experts to perform annotation tasks [127], which is a costly and time consuming process. Additionally, in order to prevent high disagreement among expert annotators, strict annotation guidelines are designed for the experts to follow. On the one hand, creating such guidelines is a lengthy and tedious process, and on the other hand, the annotation task becomes rigid and not reproducible across domains. And, as a result, the entire process needs to be repeated over and over again in every domain and task. Finally, expert annotators are not always available for specific tasks such as open domain question-answering or news events, while many annotation tasks can require multiple interpretations that a single annotator cannot provide [4].

As a solution to those problems, crowdsourcing has become a mainstream approach. It has proved to provide good results in multiple domains: annotating cultural heritage prints [100], medical relation annotation [6], ontology evaluation [98]. Following the central feature of volunteer-based crowdsourcing introduced by [123] that majority voting and high inter-annotator agreement [20] can ensure truthfulness of resulting annotations, most of those approaches are assessing the quality of their crowdsourced data based on the hypothesis [97] that there is only one right answer to each question.

However, in Chapter 2 we have shown that the preserving inter-annotator disagreement as part of the ground truth can still result in high quality data, at least for the case of medical relation extraction. In this chapter, we build on these findings and investigate more precisely *how does allowing disagreement in crowdsourcing tasks influence the quality of the data* (**RQ2**). To answer this question, we investigate across a variety of tasks and domains, and we perform an evaluation in comparison with voting aggregation methods that only take into account the opinion of the majority.

Recent work in collecting annotations for text [24, 106], sounds [55] and images [31, 113] found that disagreement between annotators is not just a result of poor quality work, and can actually be an indicator for other properties of the data, such as ambiguity and uncertainty [9]. Previous experiments we performed [5] also identified issues with the assumption of the one truth: inter-annotator disagreement is usually never captured, either because the number of annotators is too small to capture the full diversity of opinion, or because the crowd data is aggregated with metrics that enforce consensus, such as majority vote. These practices create artificial data that is neither general nor reflects the ambiguity inherent in the data.

To address these issues, we proposed the *CrowdTruth* methodology for crowdsourcing human annotation by harnessing inter-annotator disagreement, i.e representing the diversity of human interpretations in the ground truth. This is a novel approach for crowdsourcing human annotation that, instead of enforcing agreement between annotators, captures the ambiguity inherent in semantic annotation through the use of ambiguity-aware metrics for aggregating crowdsourcing responses. Based on this principle, we have implemented the CrowdTruth methodology as part of a framework [65] for machine-human computation, that first introduced the ambiguity-aware metrics and built a pipeline to process crowdsourcing data with these metrics.

To capture disagreement across different tasks, we extend the definition of our ambiguity-aware methodology (CrowdTruth version 1.0 [65]) to work both with crowdsourcing tasks that are *closed*, i.e. the annotations that can occur in the data are already known, and the workers are asked to validate their existence (e.g. given a news event, decide whether it is expressed in a tweet), and tasks that are *open*, i.e. the annotation space is not known, and workers can freely select all the choices that apply (e.g. given a news piece, select all events that appear in the text). The code for the extended CrowdTruth version 1.1 methodology and metrics is available at: `https://git.io/fA3Mq`.

We investigate tasks of text and sound annotation, in both domains that typically require expertise from annotators (e.g. medical) and those that don't (open domain). In particular, we look at four crowdsourcing tasks: *Medical Relation Extraction*, *Twitter Event Identification*, *News Event Extraction* and *Sound Interpretation*. The aim is to investigate the role of inter-annotator disagreement as part of the crowdsourcing system by applying the CrowdTruth methodology to collect data over a set of diverse use cases.

Through the use of CrowdTruth aggregation metrics, the interpretations collected from the crowd are transformed into explicit semantics for the various tasks presented in this chapter – i.e. relations expressed in sentences, topics / events expressed in tweets and news articles, words describing sounds – thus enabling knowledge base curation for these specific tasks. Furthermore, we prove that capturing disagreement is essential for acquiring high quality semantics. We achieve this by comparing the quality of the data aggregated with CrowdTruth metrics with majority vote, a method which enforces consensus among annotators. By applying our analysis over a set of diverse tasks we show that, even though ambiguity manifests differently depending on the task (e.g. each task has an optimal number of workers necessary to capture the full spectrum of opinions), our theory of inter-annotator disagreement as a property of ambiguity is generalizable for any semantic annotation crowdsourcing task.

The chapter makes the following contributions:

1. *comparative analysis of crowdsourcing aggregation methods:* we compare the performance of *ambiguity-aware metrics* and *consensus - enforcing metrics* over a diverse set of crowdsourcing tasks (Sections 3.4, 3.5);

2. *stability of crowd results:* we show in several crowdsourcing tasks that *an increased number of crowd workers leads to growth and stabilization in the quality of annotations*, going against the usual practice of employing a small number of annotators (Sections 3.4, 3.5);

3. *measuring quality in open-ended tasks:* we present an extension to the CrowdTruth methodology that allows the ambiguity-aware metrics to deal *both with open-ended and closed tasks* (Sections 3.2, 3.3), as opposed to the initial version of the CrowdTruth metrics which only processed closed tasks;

4. *semantics of ambiguity:* applying the CrowdTruth methodology we collect richer data that allows to reason about ambiguity of content (in all modality formats, e.g. images, videos and sounds), which is intrinsically relevant to the Semantic Web community.

3.2 CROWDTRUTH METHODOLOGY

In this section, we describe the CrowdTruth *methodology* version 1.1, for aggregating crowdsourcing data, which offers methods to aggregate both closed an open-ended tasks. Version 1.1 presented in this chapter is a generalization of the initial version 1.0 of CrowdTruth [65].

In Section 3.4 we use a number of annotation tasks in different domains to illustrate its use and gather experimental data to prove the main claim of this research - CrowdTruth methodology provides a viable alternative to traditional consensus-based majority vote crowdsourcing and expert-based ground truth collection. The elements of the CrowdTruth methodology are:

- annotation modeling with the *triangle of disagreement*;

- quality *metrics* for media units (input data), annotations and crowd workers;

- identification of workers with low quality annotations.

Each of these elements is applicable across a variety of domains, content modalities, *e.g.*, text, sounds, images and videos and annotation tasks, *e.g.*, closed and open-ended annotations. The following sub-sections briefly introduce the overview of the methodology elements.

3.2.1 *CrowdTruth quality metrics*

Measuring quality in CrowdTruth is done with the triangle of disagreement model (based on the triangle reference [77]), which links together media units, workers, and annotations, as seen in Fig.6. It allows us to assess the quality of each worker, the clarity of each media unit, and the ambiguity, similarity and frequency of each annotation. This model makes it possible to express how the ambiguity in any of the corners disseminates and influences the other components of the triangle. For example, an unclear sentence or an ambiguous annotation scheme would cause more disagreement between workers [7], and thus, both need to be accounted for when measuring the quality of the workers.

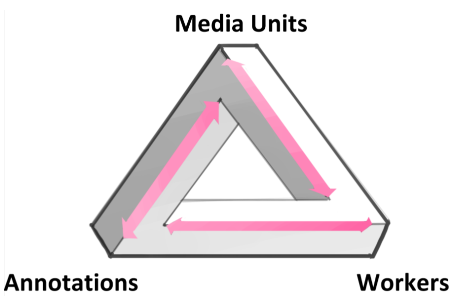


Figure 6: Triangle of Disagreement

The CrowdTruth quality metrics [7] are designed to capture inter-annotator disagreement in crowdsourcing. The metrics were introduced for *closed tasks*, i.e. multiple choice tasks, where the annotation set is known before running the crowdsourcing task. In this chapter, we present an extended version of these metrics (version 1.1), that can be used for both *closed tasks* as well as *open-ended tasks* (i.e. the annotation set is not known beforehand, and the workers can freely select all the choices that apply). The code for the CrowdTruth version 1.1 metrics is available at: `https://git.io/fA3Mq`.

The quality of the crowdsourced data is measured using a *vector space representation* of the crowd annotations. For *closed tasks*, the annotation vector contains the given answer options in the task template, which the crowd can choose from. For example, the template of a *closed task* can be composed of a multiple choice question, which appears as a list checkboxes or radio buttons, thus, having a finite list of options to choose from.

While for closed tasks the number of elements in the annotation vector is known in advance, for *open-ended tasks* the number of elements in the annotation vector can only be determined when all the judgments for a media unit have been gathered. An example of such a task can be highlighting words or word phrases in a sentence, or as an input text field where the workers can introduce keywords. In this case the answer space is composed of all the unique keywords from

WORKER ANNOTATIONS	*dog barking*	*walking*	*animal*	*echo*	*loud*
MEDIA UNIT – ANNOTATION SCORE	0.47	0.31	0.79	0.15	0.15
MEDIA UNIT VECTOR	3	2	5	1	1
MAJORITY VOTE	0	0	1	0	0

Table 8: Consider an open-ended sound annotation task where 10 workers have to describe a given sound with keywords. The media unit for this task is a sound, the annotation set contains all the keywords workers provide for a sound. The table shows the media unit metrics, as well as the majority vote score for the media unit.

all the workers that solved that media unit. As a consequence, all the media units in a closed task have the same answer space, while for open-ended tasks the answer space is different across all the media units.

Although the answer space for open-ended tasks is not known from the beginning, it is still possible to deduce a finite answer space. To achieve this, we added an *answer space dimensionality reduction step* to the methodology for open-ended tasks. Additional goals of this step are to reduce redundancy in the answer space through similarity clustering (e.g. by making sure that synonymous words do not count as disagreement between annotators), and to keep the vector space representation small enough so that the CrowdTruth quality metrics still produce meaningful values. The method for performing dimensionality reduction is dependent on the annotation task itself.

In the annotation vector, each answer option is a boolean value, showing whether the worker annotated that answer or not. This allows the annotations of each worker on a given media unit to be aggregated, resulting in a *media unit vector* that represents for each option how often it was annotated.

Three core *worker metrics* are defined to differentiate between low-quality and high-quality workers. *Worker-Worker Agreement* (*wwa*) measures the pairwise agreement between two workers across all media units they annotated in common - indicating how close a worker performs compared to workers solving the same task. *Worker-Media Unit Agreement* (*wma*) measures the similarity between the annotations of a worker and the aggregated annotations of the rest of the workers. The average of this metric across all the media units solved gives a measure of how much a worker disagrees with the crowd in the context of all media units. *Average annotations per media unit* (*na*) measures for each worker the total number of annotations they chose per media unit, averaged across all media units they annotated. Since in many tasks workers can choose all the possible annotations, a low quality worker can appear to agree more with the rest of the workers by repeatedly choosing multiple annotations, thus increasing the chance of overlap.

Two *media unit metrics* are defined to assess the quality of each unit. In this chapter, we focus on the *Media Unit-Annotation Score* – the core CrowdTruth metric, used to measure the clarity with which the media unit expresses a given annotation. This metric is computed for each media unit and each possible annotation as the cosine between the media unit vector and the unit vector for each possible annotation. This metric is used in evaluating the quality of the CrowdTruth annotations.

3.2.2 *Spam Removal*

After collecting the crowd annotations, but before the evaluation of the data, we perform spam removal. The purpose of this step is to identify the adversarial and low quality workers – e.g. those workers that always pick the same annotations, regardless of the unit. Once identified, the spam workers are removed from the dataset, and their annotations are not used in the evaluation. The methodology for spam removal is based on our previous work in [118], extended in this chapter to work also for open-ended tasks.

We identify the low quality workers by applying the core CrowdTruth worker metrics, the worker-worker agreement (*wwa*), worker-media unit agreement (*wma*) and the average number of annotations (*na*) submitted by a worker for one sentence. The first two metrics are used to model the extent to which a given worker agrees with the other annotators. The purpose is not to penalize disagreement with the majority, but rather to identify outliers, *i.e.*, workers that are in constant disagreement. For *closed tasks* where the semantics of the annotations in the answer space could rarely overlap, it is unlikely that a large number of possible annotations will occur for the same media unit. Therefore, the number of annotations per sentence can also indicate spam behavior.

In *open-ended tasks* we apply the same approach. However, we need to acknowledge the fact that open-ended tasks are more prone to disagreement due to the large answer space and thus, the overall agreement between the workers can occur with lower values. Thus, we do not have predefined values for identifying the low-quality workers, but for every task or job we use the following main heuristic: given worker w, if the agreement $wwa(w)$, $wsa(w)$ and optionally, annotations per sentence $na(w)$, parameters do not fall within the standard deviation for the task, then worker w is marked as a spammer. To confirm the validity of this metrics we also perform manual evaluation based on sampling of the results.

Based on the specificity of each task, closed or open-ended, the effort required to pick different annotations might vary. For instance, when no good annotation exists in the media unit, the time to complete the annotation is considerably reduced. This can bias the workers towards selecting the option that requires the least work. In order to prevent this, we introduce *in-task effort consistency checks*. Such annotations do not count towards building the ground truth, and are used to reduce the bias from picking the quickest option. For instance, when stating that no annotation is possible in the media unit, the workers also have to write an explanation in a text box for why no annotation were provided.

Task	Type	Media Unit	Annotations
Medical Relation Extraction	closed	sentence	medical relations: *cause*, *treat*, *prevent*, *symptom*, *diagnose*, *side effect*, *location manifestation*, *contraindicate*, *is a*, *part of*, *associated with*, *other*, *none*
Twitter Event Identification	closed	tweet	tweet events: *FIFA World Cup 2014*, *Davos world economic forum 2014*, *Islands disputed between China and Japan*, *2014 anti-China protests in Vietnam*, *Korean MV Sewol ferry ship sinking*, *Japan whaling and dolphin hunting*, *Disappearance of flight MH370*, *Ukraine crisis 2014*, *none of the above*
News Event Extraction	open-ended	sentence	words in the sentence
Sound Interpretation	open-ended	sound	tags describing sound

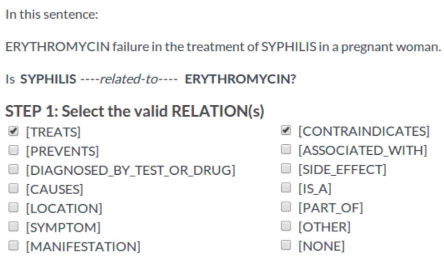
Table 9: Crowdsourcing Task Details

3.3 EXPERIMENTAL SETUP

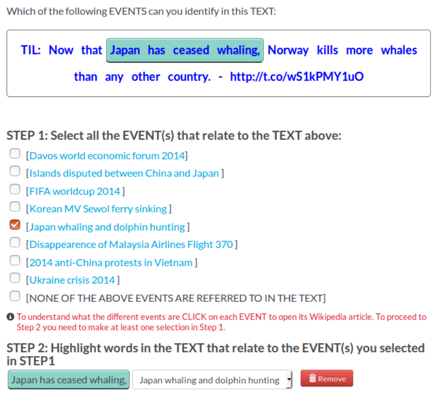
The aim of the crowdsourcing experiments described and analyzed in this chapter is to show that the CrowdTruth ambiguity-aware crowdsourcing approach produces data with a higher quality than the traditional majority vote where consensus among annotators is enforced. In order to show this, we perform an experiment over a set of four diverse crowdsourcing tasks:

- two closed tasks, i.e. *Medical Relation Extraction*, *Twitter Event Identification*,
- two open-ended tasks, i.e. *News Event Extraction* and *Sound Interpretation*.

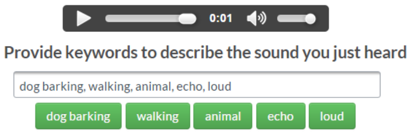
These tasks were picked from diverse domains (medical, sound, open), to aid in the generalization of our results. To evaluate the quality of the crowdsourcing data, we constructed a trusted judgments set by combining expert and crowd annotations. The rest of this section describes



(a) Medical Relation Extraction



(b) Twitter Event Identification



(c) Sound Interpretation

(d) News Event Extraction

Figure 7: Templates of the Crowdsourcing Tasks

the details of the crowdsourcing tasks, trusted judgments acquisition process, as well as the evaluation methodology we employed.

3.3.1 *Crowdsourcing Overview*

Tables 9 and 10 present an overview of the crowdsourcing tasks, as well as the datasets used. The results of the crowdsourcing tasks were processed with the use of CrowdTruth metrics (Sec. 3.2.1), and we removed consistently low quality workers based on the spam removal procedure (Sec 3.2.2). The tasks were implemented and ran on Figure Eight[1] (formerly known as CrowdFlower). The templates are available on the CrowdTruth platform[2].

Task	Source	Has Expert	Media Units	Workers/ Unit	Cost/ Judgment
Medical Relation Extraction	PubMed	yes	975	15	$0.05
Twitter Event Identification	Twitter	no	3,019	7	$0.02
News Event Extraction	TimeBank	yes	200	15	$0.02
Sound Interpretation	Freesound	yes	284	10	$0.01

Table 10: Crowdsourcing Task Data

1 https://figure-eight.com/
2 tasks marked with ∗: https://github.com/CrowdTruth/CrowdTruth/wiki/Templates

The payment per judgment was determined through a series of pilot runs of the tasks where we started with a $0.01 cost per judgment, and then gradually increased the payment until a majority of Figure Eight workers rated our tasks as having fair payments. As a result, we were able to get a constant stream of workers to participate in the tasks. The values shown in Table 10 show the final cost per judgment we reached after the pilot runs. Since crowd pay has a complex effect on the quality of the annotation [86], and in order to remove confounding factors, judgments collected with costs lower than those in Table 10 were left out of this evaluation. In total, it took two months to perform the pilot runs and then collect the judgments for all of the tasks.

The number of workers per media unit was determined experimentally with the goal of capturing all possible results from the crowd and stabilizing the quality of the annotations; this process is explained at length further on in Section 3.4, with the results of the experiment shown in Figure 9.

The **Medical Relation Extraction dataset** consists of 975 sentences extracted from PubMed[3] article abstracts. The sentences were collected using distant supervision [93], a method that picks positive sentences from a corpus based on whether known arguments of the seed relation appear together in the sentence (*e.g.*, the *treat* relation occurs between the terms *antibiotics* and *typhus*, so find all sentences containing both and repeat this for all pairs of arguments that hold). The MetaMap parser [3] was used to extract medical terms from the corpus and the UMLS vocabulary [15] was used for mapping terms to categories, and relations to term types. The intuition of distant supervision is that since we know the terms are related, and they are in the same sentence, it is more likely that the sentence expresses a relation between them (than just any random sentence). We started with a set of 8 UMLS relations important for clinical decision making [124], that became the seed in distant supervision, but this chapter only discusses results for the relations *cause* and *treat*, as these were the only relations for which we could also collect expert annotations. The expert judgment collection is detailed in Section 3.3.3.

The *medical relation extraction task* (see Figure 7a) is a *closed task*. The crowd is given a medical sentence with the two highlighted terms collected with distant supervision, and is then asked to select from a list all relations that are expressed between the two terms in the sentence. The relation list contains eight UMLS[4] relations, as well as *is a*, *part of*, *associated with*, *other*, *none* relations, added to make the choice list complete. Multiple choices are allowed in this task. To reduce the bias of selecting *none*, we also added an in-task effort consistency check by asking workers to explain in a text box why no relation is possible between the terms. The task results are processed into an

3 http://www.ncbi.nlm.nih.gov/pubmed
4 https://www.nlm.nih.gov/research/umls/

annotation vector containing a component for each of the relations. A detailed description of the crowdsourcing data collection is given in [45].

The **Twitter Event Identification dataset** consists of 3,019 English tweets from 2014, crawled from Twitter. The tweets are selected as been relevant to eight events, such as, "Japan whale hunt", "China Vietnam relation" among other controversial events. The dataset was created by querying a Twitter dataset from 2014 with relevant phrases for each of the eight events, *e.g.*, "Whaling Hunting", "Anti-Chinese in Vietnam". The *Twitter event identification task* (see Figure 7b) is a *closed task*. The crowd is asked to choose for each tweet the relevant events out of the list of eight, as well as to highlight for each of the relevant events the event mentions in the tweet. The crowd could also pick that none of the events was present in the tweet. Multiple choices of events were permitted. Since tweets and tweet annotations typically are not done by experts, we did not collect expert data for this task. To reduce the bias of selecting no event, we also added an in-task effort consistency check by asking workers to explain in a text box why none of the events is present in the tweet. The task results are processed into an annotation vector containing a component for each of the events.

The **News Event Extraction dataset** consists of 200 randomly selected English sentences from the English TimeBank corpora [108], which were also presented in [21]. The *news event extraction* (see Figure 7d) is an *open-ended task*. The crowd receives an English sentence, and is asked to highlight words or word phrases (multiple words) that describe an event or a time expression. For each sentence, the crowd is allowed to highlight a maximum of 30 event expressions or time expressions. For the purpose of this research we only focus on evaluating the extraction of event expressions. We define an *event* as something that happened, is happening, will or happen. On this dataset we employed expert annotators as described in Section 3.3.3. To reduce the bias of selecting fewer events than actually expressed in the task, we implemented an in-task effort consistency check by asking workers that annotated 3 events or less to explain in a text box why no other events are expressed in the sentence. As part of the *answer set dimensionality reduction step*, we removed the stop words from the sentence (we consider that the stop words are not meaningful for our analysis and they could add unsubstantial disagreement), and split the expressions collected from the crowd into words. The annotation vector is composed of the words in the sentence, where a word is selected in the worker vector if it appears in at least one of the expressions identified by the worker.

The **Sound Interpretation dataset** consists of 284 unique sounds sampled from the Freesound[5] online database. All these recordings

5 `https://www.freesound.org/`

and their metadata are freely accessible through the Freesound API[6]. We focused on SoundFX sounds, *i.e.*, sound effects category, as classified by [56]. The *Sound interpretation task* (see Figure 7c) is an *open-ended task*, where the crowd is asked to listen to three sounds and provide for each sound a comma separated list of keywords that best describe what they heard. For each sound, any number of answers is possible. In the *answer set dimensionality reduction step*, the annotated keywords were clustered syntacticly using spell checking and stemming, and semantically using a word2vec model [90] pre-trained on the Google News corpus. The annotation vector contains a component for each of the keywords used to describe the sound, after clustering. A detailed description of the crowdsourcing data collection and processing is given in [92]. For this dataset we also collected expert annotations from the sound creators as described in Section 3.3.3.

3.3.2 *Evaluation Methodology*

The purpose of the evaluation is to determine the quality of the annotations generated with CrowdTruth ambiguity-aware aggregating metrics. To this end, we label each media unit and annotation pair with its media unit-annotation score (see Section 3.2.1), and compare it with three other methods for labeling the data, as described below:

- **Majority vote**: Each media unit-annotation pair receives either a positive or a negative label, according to the decision of the majority of crowd workers. For each annotation performed by a crowd worker over a given media unit, we calculate the ratio of workers that have selected this annotation over the total number of workers that have annotated the unit, and assess whether it is greater or equal to 0.5. This allows for multiple annotations to be picked for one media unit. For some units, however, none of the annotations were picked by half or more of the workers. This is especially the case for open-ended tasks, such as sound interpretation, where workers put in a large number of annotations, and agreement is seldom. In these situations, we picked the annotations that were selected by the most workers (even if they do not constitute more than half). An example of the majority vote aggregation is shown in Table 8.
- **Single**: Each media unit-annotation pair receives either a positive or a negative label, according to the decision of a single crowd worker. For every media unit, this score was randomly sampled from the set of workers annotating it. Judgments from workers labeled as spammers were not employed. While a single annotator is not used as often as the majority vote in traditional crowdsourcing, we use this dataset as a baseline for the crowd,

6 `https://www.freesound.org/docs/api/`

to show that having more annotators generates better quality data.

- **Expert**: Each media unit-annotation pair receives either a positive or a negative label, according to the expert decision. The details of how expert data was collected for each tasks are discussed in Section 3.3.3.

The *evaluation of the quality of the CrowdTruth method* was done by computing the micro-F1 score over each task. The micro-F1 score was used in order to treat each case equally, without giving advantage to annotations that appear less frequently in our datasets. Using the trusted judgments collected according to Section 3.3.3, we evaluate each media unit – annotation pair as either a true positive, false positive etc. We compute the value of the micro-F1 score using the following formulas for the micro precision (Equation 1) and micro recall (Equation 2):

$$P_{micro} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FP_i} \tag{1}$$

$$R_{micro} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FN_i} \tag{2}$$

where TP_i, FP_i, FN_i, with i from 1 to n (the number of media units in the dataset), represent the number of true positive, false positive and false negative annotations for media unit i. Finally, the micro-F1 score is computed as the harmonic mean of the micro-precision and micro-recall.

An important variable in the evaluation is the *media unit-annotation score threshold* for differentiating between a negative and a positive classification. Traditional crowdsourcing aims at reducing disagreement, and therefore corresponds to high values for this threshold. Lower values means accepting more disagreement in the classification of positive answers by the crowd. In our experiments, we tried a range of threshold values for each task, to investigate with which one we achieve the best results. The media unit-annotation score threshold was also used in gathering the set of trusted judgments for the evaluation (Section 3.3.3). All the data used in this chapter can be found in our data repository[7].

3.3.3 *Trusted Judgments Collection*

To perform the evaluation, a set of trusted judgments is necessary to assess the correctness of crowd annotations. For each dataset, we

7 https://github.com/CrowdTruth/Cross-Task-Majority-Vote-Eval

manually evaluated the correctness of all the media unit annotations that were generated by the crowd and the experts. Depending on the task, the number of media unit-annotation pairs can become quite high, so we explored methods to make the manual evaluation more efficient.

For the datasets that contain expert annotation, we calculated the thresholds which yielded the maximum agreement in number of annotations between the crowd and expert annotations. These annotations were then added to the trusted judgments collection, as the judgment in this case is unambiguous. The interesting cases appear when crowd and expert disagree. Previous work we performed in crowdsourcing *Medical Relation Extraction* [8] has indicated that experts might not always provide better annotations than crowd workers. Additionally, for the *Sound Interpretation* task we noticed that experts provided considerably fewer tags than the crowd, and there was a large discrepancy between annotations of crowds and experts, with a very small overlap between their annotations. Therefore, instead of simply relying on expert judgment, the annotations where crowd and expert disagree were manually relabeled by exactly one of the authors, and then added to the trusted judgments set, which is also published in our data repository. In Appendix 3.7 we present a selection of examples where the expert judgment is different from the trusted judgment. While these cases might call into question the level of expertise of the domain experts, inconsistencies and disagreement in expert annotation are regularly reported in various annotation tasks [27, 63, 88]. Furthermore, in Section 3.4 we will show that using the trusted judgments for evaluation still results in the expert performing the best for 2 out of 3 tasks. The only task where the expert underperforms is *Sound Interpretation*, where the set of annotations provided by the expert is much smaller than the one provided by the crowd.

We collected expert annotations for the *Medical Relation Extraction* data by employing medical students. Each sentence was annotated by exactly one person. The annotation task consisted of deciding whether or not the UMLS seed relation discovered by distant supervision is present in the sentence for the two selected terms.

For the *Sound Interpretation* task, each sound in the dataset contains a description and a set of keywords that were provided by the authors of the sounds. We consider the keywords provided by the sounds' authors as trusted judgments given by domain experts.

The *news event extraction* data was annotated with events by various linguistic experts. In total, 5 people annotated each sentence but we only have access to the final annotations, a consensus among the annotators. In the annotation guidelines described in [108], events are defined as situations that happen or occur, but are not generic situations. In contrast to the crowdsourcing task, where the workers had very loose instructions, the experts had very strict rules for identifying events, strictly based on linguistic features: *(i)* tensed verbs: has called, will leave, was captured, *(ii)* stative adjectives: sunken, stalled, on board and *(iii)* event nominals: merger, Military Operation, Gulf War.

The only task without expert annotation is *Twitter Event Identification* – as it is in the open domain, no experts exist for this type of data.

3.4 RESULTS

We begin by evaluating *how the majority vote method compares with CrowdTruth*, by calculating the precision/recall metrics using the gold standards we collected for each of the four crowdsourcing tasks. Figure 8 shows the F1 score for CrowdTruth over the four tasks. The results are calculated for different media unit-annotation score thresholds for separating the data points into positive and negative classifications. Table 11 shows the detailed scores for CrowdTruth, given the highest F1 media unit-annotation score threshold.

Across all four tasks, the CrowdTruth method performs better than both majority vote and the single annotator dataset. While majority vote unsurprisingly performs the best on precision, as a consequence of its lower rate of positive labels, CrowdTruth consistently scores the best for both recall, F1 score and accuracy. These differences in classification are statistically significant, as shown in Table 12 – this was calculated using McNemar's test [89] over paired nominal data.

The evaluation of CrowdTruth compared with the expert is more nuanced. For the *Medical Relation Extraction* and *news event extraction tasks*, CrowdTruth performs as well as the expert annotators, with

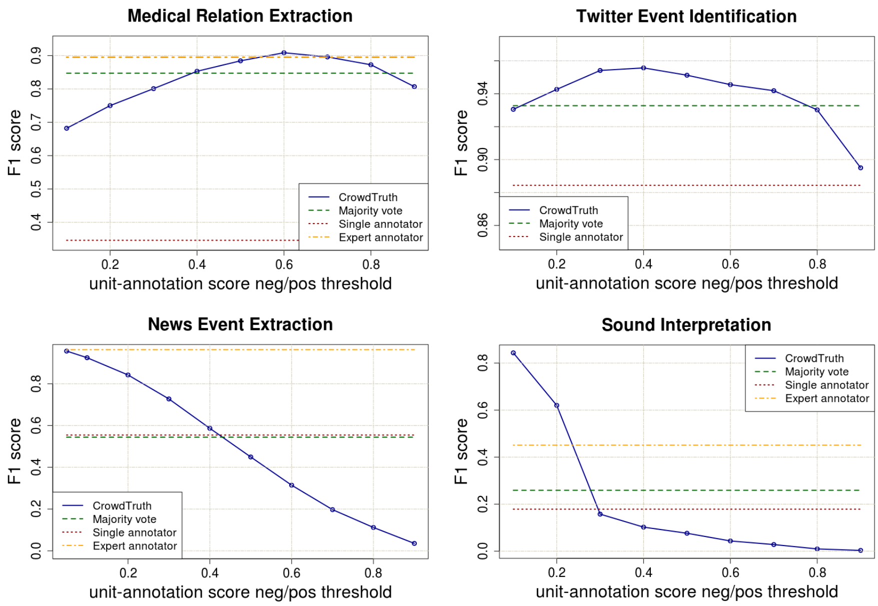


Figure 8: CrowdTruth F1 scores for all crowdsourcing tasks.

Task	Dataset	Precision	Recall	F1 score	Accuracy	Threshold
Medical Relation Extraction	CrowdTruth	0.86	0.962	0.908	0.932	0.6
	expert	0.899	0.89	0.895	0.927	
	majority vote	0.924	0.781	0.847	0.902	
	single	0.222	0.776	0.346	0.748	
Twitter Event Identification	CrowdTruth	0.965	0.945	0.955	0.995	0.4
	majority vote	0.984	0.885	0.932	0.984	
	single	0.959	0.819	0.884	0.972	
News Event Extraction	CrowdTruth	0.984	0.929	0.956	0.931	0.05
	expert	0.983	0.944	0.963	0.942	
	majority vote	0.985	0.375	0.544	0.492	
	single	0.99	0.384	0.554	0.501	
Sound Interpretation	CrowdTruth	1	0.729	0.843	0.815	0.1
	expert	1	0.291	0.45	0.515	
	majority vote	1	0.148	0.258	0.418	
	single	1	0.098	0.178	0.383	

Table 11: CrowdTruth evaluation results; the **Threshold** column shows the highest F1 media unit-annotation score threshold for each task, for which the evaluation was done.

Task	Maj. Vote	Expert	Single
Medical Relation Extraction	0.0001	0.629	$< 2.2 \times 10^{-16}$
Twitter Event Identification	0.0001	N/A	6.145×10^{-15}
News Event Extraction	$< 2.2 \times 10^{-16}$	0.505	$< 2.2 \times 10^{-16}$
Sound Interpretation	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$

Table 12: p-values for McNemar's test of statistical significance in the CrowdTruth classification, compared with the others.

p-values indicating there is no statistically significant difference in the classifications. In contrast, for the task of *Sound Interpretation*, CrowdTruth performs better than the expert by a large margin.

The second evaluation shows the *influence of the number of workers on the quality of the CrowdTruth data*. Figure 9 shows the CrowdTruth F1 score in relation to the number of workers. Given one task, the number of workers per unit varies because of spam removal, so the F1 score was calculated using at most the number of workers at every point in the graph. The number of units annotated with the given number of workers is also shown in the graph.

The effects of the number of workers on the CrowdTruth F1 is clear – more workers invariably leads to a higher F1 score. For the tasks of *Medical Relation Extraction*, *Twitter Event Identification* and *News Event Extraction*, the CrowdTruth F1 grows into a straight line, showing that the opinions of the crowd stabilize after enough workers. For the *Sound Interpretation* task, the CrowdTruth F1 score is still on an

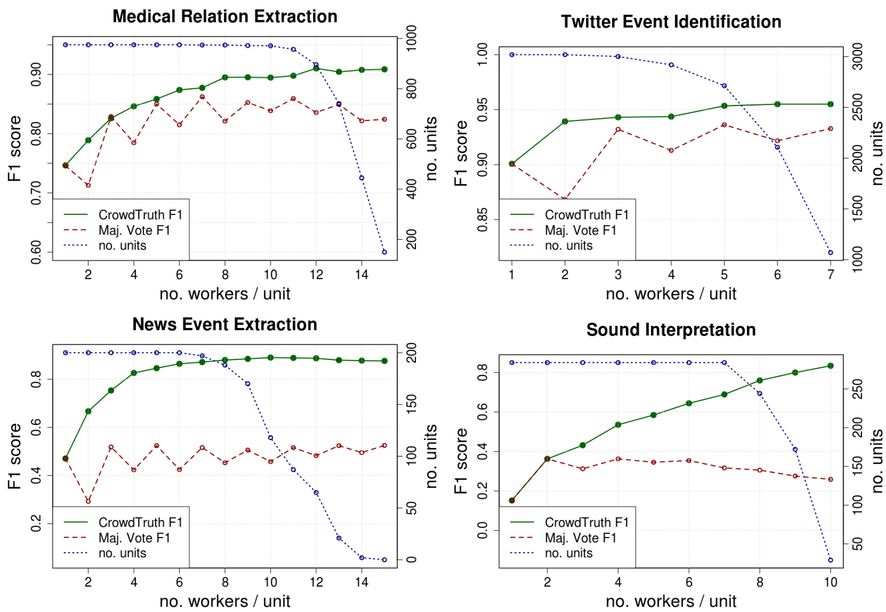


Figure 9: The effect of the number of workers per unit on the F1 score, calculated at the best media unit-annotation score threshold (Table 11). For every point, the F1 is calculated with at most the given number of workers. The number of units used in the calculation of the F1 is shown in the y-axis on the right.

upwards trend after 10 workers, possibly indicating that more workers are necessary to get the full spectrum of annotations.

Figure 9 also shows that CrowdTruth performs better than majority vote regardless of the number of workers per task. For closed tasks, increasing the number of workers has a positive impact on the majority vote F1 score. For open tasks, adding more workers has less of an effect – more workers increase the size of the annotation set for a unit, which is typically larger than for closed tasks, but the agreement is low because opinions are split between possible annotations.

Finally, Figure 10 shows an evaluation of CrowdTruth using only the expert annotations as ground truth (the *Twitter Event Identification* task does not have experts, so it could not be evaluated). The F1 scores are lower than in the evaluation over the trusted judgments collection. For the *Medical Relation Extraction Task*, majority vote performs essentially the same as CrowdTruth, whereas for the open-ended tasks, CrowdTruth still performs better. However, as we have shown in the Appendix, the expert annotations contain errors and are sometimes incomplete, particularly in the case of open-ended tasks. The evaluation using expert ground truth was done to show that the trusted judgments set is not biased in favor of CrowdTruth.

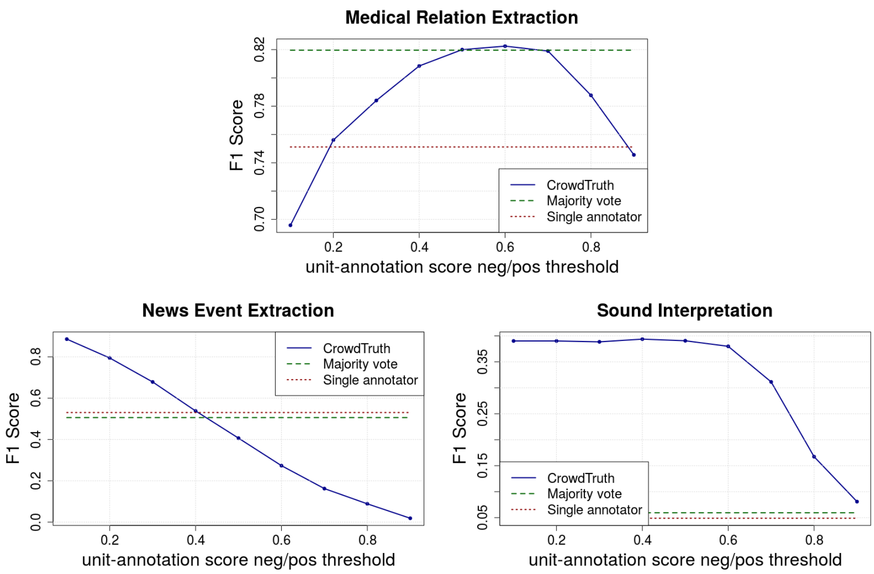


Figure 10: CrowdTruth F1 score evaluation, using expert annotation as ground truth.

3.5 DISCUSSION

The first goal in this chapter was to show that the *ambiguity-aware CrowdTruth approach with multiple annotators and disagreement-based quality scores can perform better than majority vote*, a method that enforces consensus among annotators. Our results over several crowdsourcing tasks, as seen in Figure 8, show this clearly.

The gap in performance between CrowdTruth and majority vote is the most striking for open tasks (*News Event Extraction* and *Sound Interpretation*). These tasks also require the lowest agreement threshold for achieving the best performance with CrowdTruth. During the trusted judgments collection process, we observed how these tasks are prone to a wide range of opinions – for instance, in the case of *Sound Interpretation*, there are frequent examples of labels that are semantically dissimilar, but could reasonably be applied to the same sound (e.g. the same sound was annotated with the tag `balloon popping` by one worker, and with `gunshot` by another worker). Because of this, enforcing consensus does not work for these tasks, and ambiguity-aware annotation aggregation appeared to be a viable solution.

Our evaluation also shows that processing crowd data with ambiguity-aware metrics performs at least as well as expert annotators, which is not the case for majority vote. Crowdsourcing annotation is significantly cheaper in cost than experts – e.g. even with 15 workers per unit, crowdsourcing for the task of *Medical Relation Extraction* cost 2/3 of what the experts did. The crowd also has the advantage of being readily available on platforms such as Figure Eight, while the process of finding and hiring expert annotators can incur significant time costs. As our results showed, in order for the crowdsourcing to produce results comparable in quality to that of experts, appropriate processing with ambiguity-aware metrics is a necessity.

The variation in the optimal media unit-annotation score thresholds across the tasks shows that the level of ambiguity is dependent on the crowdsourcing task, thus supporting our triangle of disagreement model (Section 3.2.1). It is not surprising that the task with the highest agreement threshold (*Medical Relation Extraction*) also has the most exact definition of a correct answer (i.e. whether a medical relation is expressed or not in a given sentence). The definition of a medical relation is fairly clear; in contrast, the definition of an event is more subjective, therefore workers were able to come up with a wider range of correct annotations.

The experimental setup provides an empirical method for selecting the optimal threshold for media unit-annotation score. However, if performing an evaluation with trusted judgments is not possible, selecting the optimal threshold becomes more difficult. For open-ended tasks, the experiments indicate that almost all opinions matter, and the agreement threshold should be as low as possible. In these cases, spam workers can be successfully eliminated by in-task effort consistency checks, and there is no need to enforce agreement beyond that. In contrast, the experiments for closed tasks show higher agreement thresholds tend to work better. The difficulty as well as the subjectivity of the domain also appear to have an impact. The threshold should grow together with the difficulty, and inversely with subjectivity. However, both difficulty and subjectivity might be difficult to measure in practice. In the end, the tuning of the threshold should be regarded similarly to a precision-recall trade-off analysis, where the optimal value depends on the requirements of the ground truth (high precision but many false negative crowd labels, or high recall but more false positives). The high variability for optimal threshold values also shows the limitations of traditional evaluation metrics like precision and recall that rely on discrete labels. CrowdTruth metrics were constructed to measure ambiguity on a continuous scale, but the use of standard metrics resulted in losing this information by forcing the conversion to either positive or negative. Ultimately, our goal is to move away from a binary ground truth that needs to be calculated using a fixed threshold, and instead to use the CrowdTruth metrics to express ambiguity on a continuous scale.

The second goal of the experiment was to show *the effect of the number of workers on the quality of CrowdTruth annotations*. The results in Figure 9 clearly show the increase in F1 score for CrowdTruth as more workers contribute to the tasks. This combined with the poor performance of the single annotator dataset proves the importance in considering a large enough pool of workers to be able to accurately capture the full spectrum of opinions.

The stabilization of the F1 score for *Medical Relation Extraction*, *Twitter Event Identification* and *News Event Extraction* is an indication that we have indeed managed to collect the entire set of opinions for these tasks. The fact that the scores all stabilize at different points in the graph (around 8 workers for *Medical Relation Extraction*, 5 for *Twitter Event Identification*, and 10 for *News Event Extraction*) indicates that the optimal number of workers is dependent on the task type, thus also confirming our hypothesis that more workers than what is typically being considered in crowdsourcing studies are necessary for acquiring a high quality ground truth.

There exists a trade-off between cost and quality of annotations that should also be considered when optimizing the number of workers. The higher cost was justified for these tasks, as the expert annotation was three times more expensive than the crowdsourced annotations at expert quality level.

An interesting observation is that the optimal number of workers per task does not seem to influence the optimal media unit-annotation score threshold for the task. The *News Event Extraction* requires a high number of workers, but the optimal media unit-annotation score threshold is low, while the *Twitter Event Identification* requires a low number of workers, and also a low media unit-annotation score threshold, at least compared to *Medical Relation Extraction*.

While four tasks is a small sample to draw conclusions from, our findings seem to indicate that ambiguity in the crowdsourcing system has an impact on both the optimal number of workers per task, as well as the clarity of the media units. These observations will form the basis for our future research in modeling crowd disagreement.

Finally, it is worth discussing the outlier characteristics of the *Sound Interpretation* task. It is the only task that does not achieve a stable F1 curve (Figure 9) possibly due to insufficient workers assigned to it. It is also unique in its lack of false positive examples – precision is 1 for the optimal media unit-annotation score threshold (Table 11), meaning that all labels collected from the crowd were accepted as part of the trusted judgments, with the exception of the spam workers that were removed from the set. *Sound Interpretation* is also the only task for which the expert annotator performed comparatively poor, with a statistically significant difference from CrowdTruth. As mentioned in the beginning of this section, after collecting the trusted judgments for this task, it became clear that the main challenge for the *Sound Interpretation* task is not to achieve consensus between annotators, but to collect the entire spectrum of annotations that describe a sound, given that this spectrum is so large (e.g. the tags `balloon popping` and `gunshot` can both reasonably apply to the same sound). For this reason, it was difficult to label tags as false positives, and the annotations of the workers, experts included, were largely non-overlapping, as they tended to interpret the sounds quite differently. The *Sound*

Interpretation task is therefore an extreme example of subjective ground truth.

3.6 RELATED WORK

3.6.1 *Crowdsourcing Ground Truth*

Crowdsourcing has grown into a viable alternative to expert ground truth collection, as crowdsourcing tends to be both cheaper and more readily available than domain experts. Experiments have been carried out in a variety of tasks and domains: medical entity extraction [53, 122, 131], medical relation extraction [73, 122], open-domain relation extraction [78], clustering and disambiguation [81], ontology evaluation [98], web resource classification [22] and taxonomy creation [17]. [116] have shown that aggregating the answers of an increasing number of unskilled crowd workers with majority vote can lead to high quality NLP training data. The typical approach in these works is to assume the existence of a universal ground truth. Therefore, disagreement between annotators is considered an undesirable feature, and is usually discarded by using either of the following methods: restricting annotator guidelines, picking one answer that reflects some consensus usually through majority voting, or using a small number of annotators.

3.6.2 *Disagreement and Ambiguity in Crowdsourcing*

Besides CrowdTruth, there exists some research on how disagreement in crowdsourcing should be interpreted and handled. In assessing the OAEI benchmark, [27] found that disagreement between annotators (both crowd and expert) is an indicator for inherent uncertainty in the domain knowledge, and that current benchmarks in ontology alignment and evaluation are not designed to model this uncertainty. [105] found similar results for the task of crowdsourced part-of-speech tagging – most inter-annotator disagreement was indicative of debatable cases in linguistic theory, rather than faulty annotation. [14] also investigate the role of inter-annotator disagreement as a possible indicator of ambiguity inherent in natural language. [79] propose a method for crowdsourcing ambiguity in the grammatical correctness of text by giving workers the possibility to pick various degrees of correctness, but inter-annotator disagreement is not discussed as a factor in measuring this ambiguity. [113] propose a framework for dealing with uncertainty in ground truth that acknowledges the notion of ambiguity, and uses disagreement in crowdsourcing for modeling this ambiguity. For the task of word sense disambiguation, [71] show that, in modeling ambiguity, the crowd was able to achieve expert-level quality of annotations. [23] implemented a workflow of tasks for collecting and correcting labels for text and images, and found that ambiguous cases cannot simply be resolved by better annotation guidelines or through worker quality control. Finally, [84] shows that often, machine learning classifiers can achieve a higher accuracy when trained with noisy crowdsourcing data. To our knowledge, this chapter presents the first experiment across several tasks and domains that explores ambiguity as a property of crowdsourcing systems, and how it can be interpreted to improve the quality of ground truth data.

3.6.3 *Crowdsourcing Aggregation beyond Majority Vote*

The literature on alternative crowdsourcing aggregation metrics typically focuses on analyzing worker performance – identifying spam workers [16, 68, 75], and analyzing workers' performance for quality control and optimization of the crowdsourcing processes [115]. [129] and [125] have used a latent variable model for task difficulty, as well as latent variables to measure the skill of each annotator, to optimize crowdsourcing for image labels. [128] use on-the-job learning with Bayesian decision theory to assign the most appropriate workers for each task, for both text and image annotation. Finally, [107] show that the surprisingly popular crowd choice (i.e. the answer that most workers thought would not be picked by other workers, even though it is correct) gave better results than the majority vote for a variety of tasks with unambiguous ground truths (state capitals, trivia questions and price of artworks).

All of these approaches show promising improvements over the use of majority vote as an aggregating method. These methods were developed only for closed tasks, primarily dealing with classification. However, the novel approach of CrowdTruth allows to explore both closed and open-ended tasks. Furthermore, our focus is on modeling ambiguity as a latent variable in the crowdsourcing system, as well as its role in generating inter-annotator disagreement, which these approaches currently do not take into account. We believe an optimal crowdsourcing approach would combine both ambiguity modeling, as well as specialized task assignment to workers. For instance, [51] developed a generative model to aggregate crowd scores that incorporates features of the data (e.g. number of words), although they do not evaluate the performance of specific features. Ambiguity as measured with CrowdTruth, like the media unit-annotation score, could be used as a data feature in such a system.

3.7 CONCLUSIONS

Gathering human annotation is a major bottleneck in the process of knowledge base curation. Crowdsourcing-based approaches are gaining popularity in the attempt to solve the issues related to volume of data and lack of annotators. Typically these practices use inter-annotator agreement as a measure of quality. However, by ignoring inter-annotator disagreement, these practices tend to create artificial data that is neither general nor reflects the ambiguity inherent in the source.

In this chapter, we investigated what is the impact of inter-annotator disagreement on the quality of data across a variety of crowdsourcing tasks. To capture inter-worker disagreement, we presented an empirically derived methodology for efficiently gathering of human annotation by aggregating crowdsourcing data with CrowdTruth metrics, which harness the inter-annotator disagreement. We applied this methodology over a set of diverse crowdsourcing tasks: closed tasks (*Medical Relation Extraction*, *Twitter Event Identification*), and open-ended tasks (*News Event Extraction* and *Sound Interpretation*).

Our results showed that *preserving disagreement in the annotations allows us to collect richer data*, which enables reasoning about the ambiguity of the content being annotated. This is intrinsically relevant to the Semantic Web community, i.e. to identify the semantics of ambiguity across all modalities, e.g. text, images, videos and sounds. In all the tasks we considered, ambiguity-aware quality scores provide better ground truth data than the traditional majority vote. Moreover, we have shown that CrowdTruth annotations have at least the same quality, even better in the case of *Sound Interpretation*, as expert annotations. Finally, we showed that, contrary to the common crowdsourcing practice of employing a small number of annotators, adding more crowd workers actually can lead to significantly better annotation quality.

In the future, we plan to expand our methodology to more complex annotation tasks, that require multiple or combined types of input beyond the closed/open-ended categorization we presented in this chapter. We are also working on expanding the CrowdTruth metrics for ambiguity to incorporate the state-of-the art in modeling crowd worker and data features [51]. Finally, we want to use the CrowdTruth data in practice for training and evaluating information extraction models used to populate the Semantic Web.

ACKNOWLEDGEMENTS

We would like to thank Emiel van Miltenburg for assisting with the exploration of feature analysis of sounds, Chang Wang and Anthony Levas for providing and assisting with the medical data, Zhaochun Ren for the help in gathering the Twitter dataset, Tommaso Caselli for providing the news dataset, and the anonymous crowd workers for their contributions to our crowdsourcing tasks.

APPENDIX: EXAMPLE MEDIA UNITS WHERE THE EXPERT JUDGMENT IS DIFFERENT FROM THE TRUSTED JUDGMENT

Media Unit	Annotation	Expert Judgment	Crowd Score	Trusted Judgment
The *epidermal nevus syndrome* is a neurocutaneous disorder characterized by *distinctive skin lesions* and often serious somatic and central nervous system (CNS) abnormalities.	*cause*	no	0.98	yes
For empiric *treat*ment of epididymitis, especially when gonococcal or *chlamydial infection* is likely Ofloxacin or *levofloxacin* should be used only if epididymitis is not *cause*d by gonorrhea.	*treat*	no	0.966	yes
In contrast, we did not find a definite increase in the LGL percentage within 6 months postpartum in patients with *Graves' disease* who relapsed into *Graves' thyrotoxicosis*.	*cause*	no	0.738	yes
The 1 placebo controlled trial that found black cohosh to be effective for *hot flashes* did not find *estrogen* to be effective, which casts doubt on the study's validity.	*treat*	no	0.73	yes
Multicentric reticulohistiocytosis (MR) is a *systemic disease* of unknown *cause* characterized by the presence of a heavy macrophage infiltrate in skin and synovial tissues and the development of an erosive polyarthritis.	*cause*	yes	0.697	no
Urokise versus *tissue plasminogen activator* in *pulmonary embolism*.	*treat*	yes	0.365	no
The principal differences between these vaccines are the transmission of live vaccine viruses from recipients to their contacts and the occurrence of occasional cases of *paralytic poliomyelitis* associated with use of *live poliovirus vaccine*	*treat*	yes	0.1	no
These cases highlight the importance of considering *PTLD* in the differential diagnosis of *lymphadenopathy*.	*cause*	yes	0.09	no

Table 13: Example sentences from the *Medical Relation Extraction* task where the expert judgment is different from the trusted judgment. The pair of terms that express the medical relation are shown in italic font in the media unit.

MEDIA UNIT	ANNOTATION	EXPERT JUDGMENT	CROWD SCORE	TRUSTED JUDGMENT
The plan provides for the *distribution* of one common stock-purchase right as a dividend for each share of common outstanding	*distribution*	no	0.95	yes
Two Middle East terrorists with records of successful *attacks* against Western targets Abu Nidal and Abu Abbas have ties to Baghdad.	*attacks*	no	0.73	yes
Secretary of State James Baker said on ABC-TV's "This Week With David Brinkley" that the series of UN resolutions condemning Iraq's *invasion* of Kuwait "imply that the restoration of peace and stability in the Gulf would be a heck of a lot easier if he and that leadership were not in power in Iraq."	*invasion*	no	0.53	yes
The company also said it continues to explore all options concerning the possible *sale* of National Aluminum's 54.5% stake in an aluminum smelter in Hawesville Ky.	*sale*	no	0.24	yes
Yield on the issue was 7.88%	*no event*	yes	0.14	no
Har-Shefi said she heard Amir talk about killing Rabin but did not tell the police because she did not believe he was *serious*.	*serious*	yes	0	no
The American hope is that someone from within Iraq perhaps from the army 's professional ranks will step forward and push Saddam Hussein aside so that the country can begin recovering from the disaster.	*no event*	yes	0	no

Table 14: Example sentences from the *News Event Extraction* task where the expert judgment is different from the trusted judgment. The annotation is shown in italic font in the media unit.

Media Unit URL	Media Unit Description	Annotation	Expert Judgment	Crowd Score	Trusted Judgment
https://freesound.org/data/previews/21/21266_88803-hq.mp3	jazz	cymbals	no	0.272	yes
		bangle	no	0.136	yes
		rhythmic	no	0.136	yes
https://freesound.org/data/previews/26/26086_11477-hq.mp3	chicken	birds	no	0.538	yes
		geese	no	0.359	yes
		horns	no	0.359	yes
https://freesound.org/data/previews/35/35823_317782-hq.mp3	weird drums	music	no	0.875	yes
		band	no	0.145	yes
		disco	no	0.145	yes
https://freesound.org/data/previews/39/39329_404624-hq.mp3	trip hop	beat	no	0.371	yes
		percussion	no	0.371	yes
		chimes	no	0.371	yes
https://freesound.org/data/previews/41/41462_78779-hq.mp3	beer glasses	clicks	no	0.242	yes
		clink	no	0.242	yes
		ding	no	0.242	yes

Table 15: Example sounds from the *Sound Interpretation* task where the expert judgment is different from the trusted judgment.

4 LEARNING RELATION CLASSIFICATION FROM THE CROWD

> *It is not humanly possible to gather immediately from it what the logic of language is. Language disguises thought.*
> – Ludwig Wittgenstein, TRACTATUS LOGICO-PHILOSOPHICUS

In this chapter, we investigate whether disagreement-preserving crowdsourcing data can be used to improve the performance of natural language processing models, focusing on the use case of relation extraction from sentences. Distant supervision (DS) is a well-established method for relation extraction from text, based on the assumption that when a knowledge-base contains a relation between a term pair, then sentences that contain that pair are likely to express the relation. We use the results of a crowdsourcing relation extraction task to identify two problems with DS data quality: the widely varying degree of false positives across different relations, and the observed causal connection between relations that are not considered by the DS method. The crowdsourcing data aggregation is performed using ambiguity-aware CrowdTruth metrics, that are used to capture and interpret inter-annotator disagreement. We also explore the problem of propagating human annotation signals gathered for open-domain relation classification through the CrowdTruth methodology for crowdsourcing. We present preliminary results of using the crowd to enhance DS training data for a relation classification model, without requiring the crowd to annotate the entire set. Finally, we present a method that propagates crowd annotations to sentences that are similar in a low dimensional embedding space, expanding the number of labels by two orders of magnitude. Our experiments show significant improvement in a sentence-level multi-class relation classifier.

This chapter is based on the following publications:

- *False Positive and Cross-relation Signals in Distant Supervision Data* in the Automated Knowledge Base Construction Workshop at NeurIPS 2017, co-authored by Lora Aroyo and Chris Welty [42];
- *Crowdsourcing Semantic Label Propagation in Relation Classification* in the Fact Extraction and Verification Workshop at EMNLP 2018, co-authored by Lora Aroyo and Chris Welty [46].

4.1 INTRODUCTION

Distant supervision (DS) [93, 126] is a well-established semi-supervised method for performing relation extraction from text. It is based on the assumption that, when a knowledge-base contains a relation between a pair of terms, then any sentence that contains that pair is likely to express the relation. This approach can generate false positives, as not every mention of a term pair in a sentence means a relation is also expressed [52]. Furthermore, dependencies between the semantics of the relations such as causality or contradiction are also not considered by the DS methodology. It is often assumed that these disadvantages are compensated for by the scale of the data a DS method can produce, or can be largely overcome with crowdsourced human annotation [2, 85].

Previously, we have shown that preserving disagreement in training data for relation extraction results in performance that is comparable to that of models trained with data from domain experts, and better than for models just trained on DS (Chapter 2). However, the main advantage of DS is that it is cheap to acquire, and therefore easy to scale up in order to train the state-of-the-art models based on neural networks [69, 133]. In contrast, crowdsourcing data is more expensive to acquire, especially when collecting a multitude of perspectives so as to capture disagreement. In this chapter, we investigate whether *disagreement-preserving crowdsourcing data can be used at the scale needed to improve the performance of a relation classification neural network model* (**RQ3**), when the model requires hundreds of thousands of training examples.

To achieve this, we present two experiments in correcting DS data with crowdsourcing. In the first experiment, we identify the DS issues that crowdsourcing is able to solve: the widely varying degree of false positives across different TAC-KBP relation types, and the observed causal connection between relations missing from DS. We expose these problems using the CrowdTruth [5, 7, 8] approach to gathering human annotated data, analyze them, and offer preliminary heuristic and statistical approaches to incorporating them back into DS-based training, that provides better sentence-level relation classification results.

The second experiment explores the possibility of automatically expanding smaller human-annotated datasets to DS scale using semantic label propagation. Sterckx et al. [119] first proposed this method to correct labels of sentence dependency paths by using expert annotators, and then propagating the corrected labels to a corpus of DS sentences by calculating the similarity between the labeled and unlabeled sentences in the embedding space of their dependency paths. We adapt and simplify semantic label propagation to propagate labels without computing dependency paths, and using the crowd instead of experts, which is more scalable. Our simplified algorithm propagates crowdsourced annotations from a small sample of sentences to a large DS corpus. To evaluate our approach, we perform an experiment in open domain relation classification in the English-language, using a corpus of sentences [44] whose labels have been collected using the CrowdTruth method.

This chapter makes the following contributions:

1. a comparison between crowdsourced and distant-supervision data quality, highlighting the *distant-supervision issues fixed with by the crowd;*
2. a *label propagation methodology* to use crowdsourcing data at the scale needed for training neural relation classification models;
3. a *d*ataset of 4,100 sentences annotated with relations in the open domain, that have been processed with disagreement analysis to capture ambiguity [44].

4.2 RELATED WORK

In recent years, researchers have explored unsupervised methods for correcting DS data. For the task of knowledge base completion, [52] applied memory networks both to correct false positives in the data, and to capture dependencies between relations. For the same task, [70] developed a loss function that works with multi-label data, in order to capture co-occurring relations. For relation classification from sentences, [111] learn embeddings that capture cross-signals between relations. However, these approaches are dependent on training data that can express relation semantics with at least some accuracy. The initial experiments presented in this chapter show the error rate in the DS data can be so high that unsupervised learning becomes unreliable when it comes to capturing cross-relation signals.

Crowdsourcing is a well-used approach to correcting the mistakes in DS by scaling out cheap human annotation. Angeli et al. [2] present an active learning approach to select the most useful sentences that need human re-labeling using a query by committee. Zhang et al. [132] show that labeled data has a statistically significant, but relatively low impact on improving the quality of DS training data, while increasing the size of the DS corpus has a more significant impact. In contrast, Liu et al. [85] prove that a corpus of labeled sentences from a pool of highly qualified workers can significantly improve DS quality. All of these methods employ large annotated corpora of 10,000 to 20,000 sentences. In our experiment, we show that a comparatively smaller corpus of 2,050 sentences is enough to correct DS errors through semantic label propagation. Levy et al. [82] have shown that a small crowdsourced dataset of questions about relations can be exploited to perform zero-shot learning for relation extraction. Pershina et al. [104] use a small dataset of hand-labeled data to generate relation-specific guidelines that are used as additional features in the relation extraction.

We have been studying the problem of collecting human annotations from the crowd using the CrowdTruth methodology [5]. Our method differs in that it gathers many annotations for the same examples, to better reflect properties like ambiguity, human error and spam, and the target semantics [7]. As discussed in Chapter 2, we have used this method successfully to improve DS results for the task of medical relation extraction, achieving annotation quality equivalent to that provided by medical experts, at less than half the cost.

The label propagation method was introduced by Xiaojin and Zoubin [130], while Chen et al. [29] first applied it to correct DS, by calculating similarity between labeled and unlabeled examples an extensive list of features, including part-of-speech tags and target entity types. In contrast, our approach calculates similarity between examples in the word2vec [90] feature space, which it then uses to correct the labels of training sentences. This makes it easy to reuse by the state-of-the-art in both relation classification and relation extraction – convolutional [69] and recurrent neural network methods [133] that do not use extensive feature sets. To evaluate our approach, we used a simple convolutional neural network to perform relation classification in sentences [96].

4.3 EXPERIMENTAL SETUP

4.3.1 *Data and Crowdsourcing Setup*

For the *relation-based correction experiment*, we asked the crowd annotate 2,500 sentences from the NIST TAC-KBP 2013 English Slotfilling data that were annotated with DS. For the *semantic label propagation experiment*, we augmented this dataset by another 2,050 sentences picked at random from the corpus of Angeli et al. [2]. For both datasets, we collected annotations for 16 popular relations from the open domain that occur between terms of types *Person*, *Organization* and *Location*, as shown in in Figure 11,[1]. The resulting corpus also contains candidate term pairs and DS seed relations for each sentence. As some relations are more general than others, the relation frequency in the corpus is slightly unequal – e.g. *places of residence* is more likely to be in a sentence when *place of birth* and *place of death* occur, but not the opposite.

We ran a multiple-choice crowdsourcing task (Figure 11), asking *15 workers to annotate each sentence* with the appropriate relations, or choose the option *none* if none of the relations presented apply. Workers were encouraged to select all relations that apply. Each worker was paid \$0.05 per sentence. The task was run on the Figure Eight[2]

1 The *alternate names* relation appears twice in the list, once referring to alternate names of persons, and the other referring to organizations.

2 `https://www.figure-eight.com/`

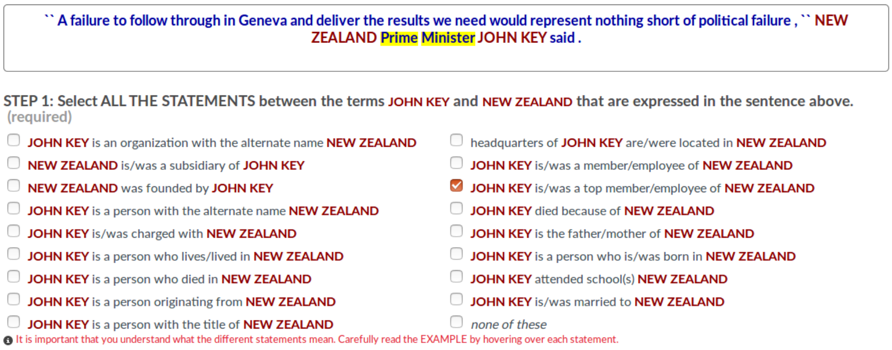


Figure 11: Fragment of the crowdsourcing task template.

and Amazon Mechanical Turk[3] crowdsourcing platforms. The data is available online [44].

4.3.2 *CrowdTruth Metrics*

Crowdsourcing annotations are aggregated usually by measuring the consensus of the workers (e.g. using majority vote). This is based on the assumption that a single right annotation exists for each example. In the problem of relation classification, the notion of a single truth is reflected in the fact that a majority of proposed solutions treat relations as mutually exclusive, and the objective of the classification task is usually to find the best relation for a given sentence and term pair. In contrast, the CrowdTruth methodology proposes that crowd annotations are inherently diverse [8], due to a variety of factors such as the ambiguity that is inherent in natural language. We use a comparatively large number of workers per sentences (15) in order to collect inter-annotator disagreement, which results in a more fine-grained ground truth that separates between clear and ambiguous expressions of relations. This is achieved by labeling examples with the inter-annotator agreement on a continuous scale, as opposed to using binary labels.

To aggregate the results of the crowd, we use CrowdTruth metrics[4] [48] to capture and interpret inter-annotator disagreement as quality metrics for the workers, sentences, and relations in the corpus. The annotations of one worker over one sentence are encoded as a binary worker vector with 17 components, one for each relation and including *none*. The quality metrics for the workers, sentences and relations, are based on average cosine similarity over the worker vectors – e.g. the quality of a worker w is given by the average cosine similarity between the worker vector of w and the vectors of all other workers that annotated the same sentences. These metrics are mutually dependent (e.g. the sentence quality is weighted by the relation

3 https://www.mturk.com/
4 https://github.com/CrowdTruth/CrowdTruth-core

quality and worker quality), the intuition being that low quality workers should not count as much in determining sentence quality, and ambiguous sentences should have less of an impact in determining worker quality, etc. Among the CrowdTruth measures discussed in this chapter, we calculate the per-relation *false positive (FP) rate*, the *causal power* between relation pairs (RCP), and the *sentence-relation score*. Spam removal was performed as well, but the details of this process are not relevant for the chapter.

For each sentence-relation pair, we compute the *sentence-relation score (srs)* as the ratio of workers that picked that relation over the total of number of workers, weighted by the worker and relation quality. The *srs* measures how clearly the relation is expressed in the sentence (the higher the score, the more likely the relation is expressed), and is used as a continuous truth measure. In order to make our results compatible with discrete evaluation metrics (e.g. P, R, F1), we have chosen a threshold of 0.5 per relation, corresponding to the majority vote, that allows for multiple relations to be considered correct in a sentence. False positive rates are then computed per relation using this threshold.

Causal power [30] is an estimate of the probability that the presence of one relation implies the presence of another. Given two relations i and j, $RCP(R_i, R_j) = [P(R_j|R_i) - P(R_j|\neg R_i)]/[1 - P(R_j|\neg R_i)]$, where $P(R_i)$ is the probability that relation R_i is annotated in the sentence. This probability can be calculated on a micro basis giving us the probability of one worker annotating two relations together; the *macro RCP* calculates the probabilities in the sentence vectors, capturing causality as a result of two relations being annotated together in the same sentence, but not necessarily by the same workers. We found micro RCP to be vastly inferior to macro RCP, which is further evidence of the value of having multiple workers per sentence, and only include the macro RCP results in this chapter.

4.3.3 *Label Propagation*

Inspired by the semantic label propagation method [119], we propagate the vectors of *srs* scores on each crowd annotated sentence to a much larger set of distant supervised (DS) sentences (see datasets description in Section 4.3.4), scaling the vectors linearly by the distance in low dimensional word2vec vector space [90]. One of the reasons we chose the CrowdTruth set for this experiment is that the annotation vectors give us a score *for each relation* to propagate to the DS sentences, which have only one binary label.

Similarly to Sultan, Bethard, and Sumner [120], we calculate the vector representation of a sentence as the average over its word vectors, and like Sterckx et al. [119] we get the similarity between sentences using cosine similarity. Additionally, we restrict the sentence representation to only contain the words between the term pair, in order to reduce the vector space to the one that is most likely to express the relations. For each sentence s in the DS dataset, we find the sentence l' from the crowd annotated set that is most similar to s: $l' = \arg\max_{l \in Crowd} cos\ sim(l,s)$. The score for relation r of sentence s is calculated as the weighted average between the $srs(l',r)$ and the original DS annotation, weighted by the cosine similarity to s ($cos\ sim(s,s) = 1$ for the DS term, and $cos\ sim(s,l')$ for the srs term):

$$DS^*(s,r) = \frac{DS(s,r) + cos\ sim(s,l') \cdot srs(l',r)}{1 + cos\ sim(s,l')} \tag{3}$$

where $DS(s,r) \in \{0,1\}$ is the original DS annotation for the relation r on sentence s.

4.3.4 *Training the Relation Classification Model*

The relation classification model employed is based on Nguyen and Grishman [96], who implement a convolutional neural network with four main layers: an embedding layer for the words in the sentence and the position of the candidate term pair in the sentence, a convolutional layer with a sliding window of variable length of 2 to 5 words that recognizes n-grams, a pooling layer that determines the most relevant features, and a softmax layer to perform classification.

We have adapted this model to be both multi-class and multi-label – we use a sigmoid cross-entropy loss function instead of softmax cross-entropy, and the final layer is normalized with the sigmoid function instead of softmax – in order to make it possible for more than one relation to hold between two terms in one sentence. The loss function is computed using continuous labels instead of binary positive/negative labels, in order to accommodate the use of the *srs* in training. The features of the model are the word2vec embeddings of the words in the sentences, together with the position embeddings of the two terms that express the relation. The word embeddings are initialized with 300-dimensional word2vec vectors pre-trained on the Google News corpus[5]. Both the position and word embeddings are nonstatic and become optimized during training of the model. The values of the other hyper-parameters are the same as those reported by Nguyen and Grishman [96]. The model was implemented in Tensorflow [1], and trained in a distributed manner on the DAS-5 cluster [13].

4.4 RESULTS AND DISCUSSION

In this section, we discuss the results of our experiments on improving the performance of relation classification models with CrowdTruth.

5 `https://code.google.com/archive/p/word2vec/`

First, we evaluate DS data quality using crowdsourced data as ground truth. Next, we present experimental results for two methods to enhance DS training data for relation classification: (1) a preliminary experiment with *relation-based correction* of the DS data, that shows the potential of disagreement-aware crowdsourcing to correct DS data at scale, without requiring the crowd to annotate the entire set, and (2) an experiment with *semantic label propagation* that shows a robust way of propagating the information in a small crowdsourced corpus to the scale needed for training relation classification models.

4.4.1 *Evaluating DS with CrowdTruth*

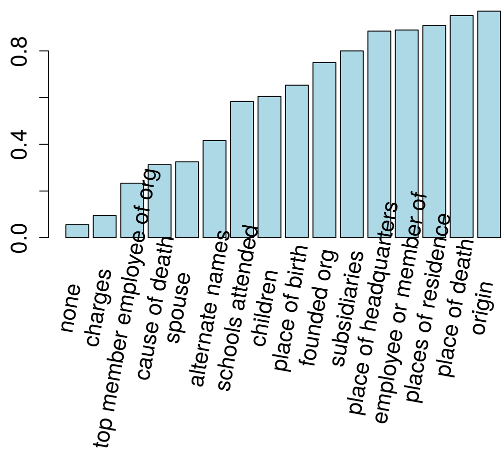


Figure 12: DS ratio of false positive over all positive labels, using the crowd as ground truth.

Using the *srs* as a ground truth at a 0.5 threshold, Figure 12 shows the correctness of the DS labels on the initial dev set of 1,025 sentence. There is *considerable variation in DS data quality across relations*. The *origin* and *place of death* relations scored particularly badly, with more than 90% false positives. With such a high error rate in some relations, it is arguable that any classifier could learn anything meaningful, regardless of algorithm or quantity of data.

Manual error analysis on the initial dev set showed that many sentences contain a *Person - Location* pair, where freebase specified both that the person resided in and died at that location. This makes intuitive sense, people tend to die in the places they live. In most of these cases, the sentence expressed only the *places of residence* relation, leading to the false positives. The *origin* relation data suffers from the same problem. Table 18 in the Appendix 4.5 shows several examples of these sentences. This led us to consider a heuristic solution to this problem as a headroom study as well as a statistical solution. Both are discussed in Section 5.

	PoB	O	PoR	PoD	FO	EoM	TEoM
PoB	1	**0.64**	0.17	-0.12	-0.19	-0.2	-0.21
O	**0.88**	1	0.31	-0.16	-0.29	-0.22	-0.22
PoR	0.42	**0.56**	1	**-0.1**	-0.59	0.12	0.13
PoD	-0.03	-0.03	-0.01	1	-0.04	-0.05	-0.05
FO	-0.07	-0.07	-0.09	-0.06	1	0.1	0.13
EoM	-0.45	-0.36	0.11	-0.47	0.62	1	0.82
TEoM	-0.5	-0.38	0.13	-0.45	0.86	**0.86**	1

(a) Crowd-based RCP

	PoB	O	PoR	PoD	FO	EoM	TEoM
PoB	1	**-0.6**	0.55	-0.14	-0.54	-0.48	-0.57
O	**-0.02**	1	-0.11	-0.16	-0.16	0.19	-0.15
PoR	0.65	**-0.33**	1	**0.45**	-0.7	-0.68	-0.75
PoD	-0.06	-0.18	0.17	1	-0.18	-0.13	-0.19
FO	-0.08	-0.06	-0.09	-0.06	1	0.09	0.09
EoM	-0.35	0.35	-0.42	-0.21	0.46	1	0.66
TEoM	-0.16	-0.1	-0.17	-0.12	0.34	**0.24**	1

(b) DS-based RCP.

Table 16: RCP for relation subset: *place of birth* (PoB), *origin* (O), *places of residence* (PoR), *place of death* (PoD), *founded organization* (FO), *employee or member* (EoM), *top employee or member* (TEoM). The scores show the causal power $RCP(R_i, R_j)$ of relations R_i in the rows, over the relations R_j in the columns. Significant changes between crowd annotation based causal power and distant supervision are in bold.

The results of the macro RCP analysis for six of the relations we analyzed (Table 16) shows that the *place of birth* relation has a high causal power (0.64) over *origin*, meaning that when *place of birth* is annotated in a sentence, *origin* is also likely to appear, with the inverse causal power at 0.88. This high co-causality seems to indicate a confusion between the two relations. Note also that these two relations have significant differences in causal power in the DS-based data. In contrast, *place of death* has a high causal power over *places of residence* in the DS data (0.45), reflecting the high error rate of *place of death* caused by the overlap in the KB with *places of residence*.

In the crowd data we see a much higher co-causality for *employee or member* and *top employee or member*, with only a slight preference in the data for what we expect to be the "correct" causal direction (that *top employee or member* causes *employee or member*), but in the DS-based analysis, the incorrect interpretation drops a lot. In manual error analysis we observed that these are properties of the data set, which talk about more famous people who tend to be leaders and founders, not "regular" employees. Table 19 in the Appendix shows several examples sentences with false negative DS labels due to missing causality.

Among the non-symmetric causal pairs we see that *top employee or member* causes *founded organization*, *employee or member* causes *founded org*, and *top employee or member* causes *founded org*. These again appear to be properties of the data set.

4.4.2 *Relation-Based Correction Experiment*

We expect that the metrics from CrowdTruth annotation can be used to systematically enhance DS data at scale, without requiring the crowd to annotate the entire set. As a preliminary headroom exercise, we trained three models to test a few simple heuristic characterizations of our analysis, and compared them to a baseline trained purely on DS data. In each model, we changed only the training set (using the methods described below). Each model was trained for 20,000 iterations, after the point of stabilization for the train loss. We used the data in our initial held-out test set as an evaluation target, again processing the continuous SRS scores with a threshold of 0.5 to yield discrete truth values for calculating P, R, and F. To evaluate the relation classification model on CrowdTruth data with discrete metrics, we set a comparable threshold of 0.5 on the model confidence score, separating between negative and positive labels. Results are shown in Table 17.

1. **DS:** The baseline of 235,000 sentences annotated by DS from freebase relations, used in Riedel et al. [109]. The per-relation training labels are binary (1 and 0), based on the results of DS.

2. **DS merged:** Based on the results of the causality analysis, the training set is augmented to reflect the highest cross-relation signals. We merge relations with symmetric RCP (*origin* and *place of birth*), and add the implied relation in the case of asymmetric RCP (*employee or member* and *top employee or member*). To merge, the **DS** baseline data is updated so that the symmetric relations always co-occur, and adding caused relation whenever the caused relation appears. This approach shows a huge improvement across the board over the baseline, with the overall highest P and F.

3. **DS_RCP:** Instead of manually identifying merged relations, the training data is augmented by using the RCP scores. When a relation i has a positive **DS** label for a given sentence, the labels of all other relations $j \neq i$ are updated by adding the macro RCP that i has over j. The maximum value for the label is clipped at 1, to keep scores in the $[0, 1]$ interval. The training labels

in this set have continuous values, as opposed to the binary values in the previous two sets. The formula for updating the training label for relation j in sentence s is: $DS\ RCP(s,j) = max[1, DS(s,j) + \sum_{i \neq j} RCP(i,j) \cdot DS(s,i)]$, where $DS(s,i)$ is the DS label of relation i in sentence s. This method was comparable in precision to the baseline, but scored a huge win in recall. The recall increase makes sense, though we have yet to investigate or explain the lack of increase in precision.

4. **DS_FP:** Our analysis showed that the *place of death* relation was a large source of false positives in the DS data, because most of the positives were actually expressing *places of residence*. In every sentence in the DS training set that had a 1 for *place of death*, we updated the score by subtracting its false positive ratio, which was used in the loss function as described above. This did not impact the results over the baseline, mainly because there were not many *place of death* relations in the DS data nor the test set, and any improvement did not impact the overall result. We are confident that more systematic treatment of false positive rates will improve performance.

	Precision	Recall	F1 score
DS	0.19	0.22	0.2
DS merged	**0.43**	0.33	**0.37**
DS_RCP	0.19	**0.48**	0.27
DS_FP	0.21	0.22	0.21

Table 17: Precision & Recall at 20,000 training steps.

The differences (in bold in Table 16) between the crowd and DS-based causal power accounts for some of the classification errors in our trained system, and we expect them to be a significant cause of error in systems that try to learn cross-relation signals from DS data alone.

The preliminary results are not overwhelming, but highly indicative. There is considerable headroom in cross-relation signals, and a more robust approach holds promise to eliminate manual analysis, and work as part of an overall pipeline that includes partial crowd data.

4.4.3 *Label Propagation Experiment*

Building on the results from the previous section on, we studied *label propagation* as a more robust method of using a small crowdsourced corpus to augment a DS dataset larger by several orders of magnitude. As opposed to relation-based correction methods, label propagation takes into account the information contained in the sentences themselves, providing a more fine-grained method to correct errors in DS.

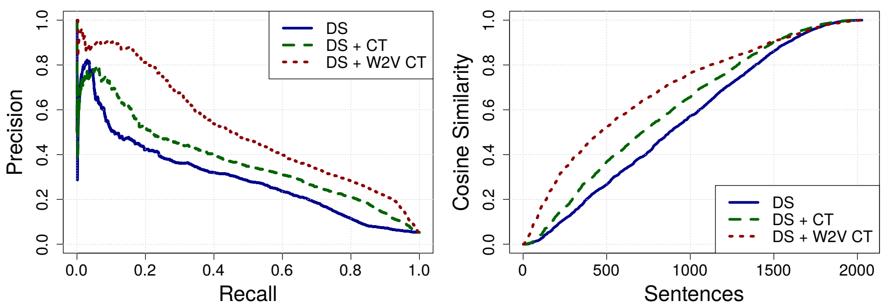
For this experiment, we split the full 4,100 crowd sentences into a dev and a test set of equal size, and trained three models to compare with the baseline on the held-out test set. Each model is trained for 25,000 iterations, after the point of stabilization for the train loss. The models were trained by the following datasets:

1. **DS:** The baseline of 235,000 sentences annotated by DS.

2. **DS + CT:** The 2,050 crowd dev annotated sentences added directly to the DS dataset.

3. **DS + W2V CT:** The DS* dataset (Eq. 3), with relation scores propagated over the 2,050 crowd dev sentences.

To evaluate the performance of the label propagation method, we calculate the micro precision and recall (Figure 13a), as well as the cosine similarity per sentence with the test set (Figure 13b). In order to calculate the precision and recall, a threshold of 0.5 was set in the *srs*, and each sentence-relation pair was labeled either as positive or negative. However, for calculating the cosine similarity, the *srs* was used without change, in order to better reflect the degree of agreement the crowd had over annotating each example. We observe that **DS + W2V CT**, with a precision/recall $AUC = 0.512$, significantly outperforms **DS** (P/R $AUC = 0.294$). **DS + CT** (P/R $AUC = 0.372$) also does slightly better than **DS**, but not enough to compete with the semantic label propagation method. The cosine similarity result (Figure 13b) shows that **DS + W2V CT** also produces model predictions that are closer to the different agreement levels of the crowd. Take advantage of the agreement scores in the CrowdTruth corpus, the cosine similarity evaluation allows us to assess relation confidence scores on a continuous scale. The crowdsourcing results and model predictions are available online [44].



(a) Precision / Recall curve, calculated for each sentence-relation pair.

(b) Distribution of sentence-level cosine similarity with test set values.

Figure 13: Label propagation evaluation results.

One reason for which the semantic label propagation method works better than simply adding the correctly labeled sentences to the train set is the high rate of incorrectly labeled examples in the DS training data, as discussed in Section 4.4.1. The success of the **DS + W2V CT** comes in part because the method relabels all sentences in DS. Adding correctly labeled sentences to the train set would require a significantly larger corpus in order to correct the high false positive rate, but semantic label propagation only requires a small corpus (two orders of magnitude smaller than the train set) to achieve significant improvements.

4.5 CONCLUSION

This chapter explores how to *improve the performance of open-domain relation classification models with disagreement-aware ground truth data*, by propagating human annotation signals in distant supervision training data. We have shown a very significant variation in the false positive rate in distant supervision data, and it seems extremely likely that this can be exploited to improve training. We also presented experimental results for two methods to enhance DS training data for relation classification: (1) a preliminary experiment with *relation-based correction* of the DS data, that shows the potential of disagreement-aware crowdsourcing to correct DS data at scale, without requiring the crowd to annotate the entire set, and (2) an experiment with *semantic label propagation* that shows a robust way of propagating the information in a small crowdsourced corpus to the scale needed for training relation classification models.

Our version of the label propagation approach passes on the information in human annotations to sentences that are similar in a low dimensional embedding space, using a small crowdsourced dataset of 2,050 sentences to correct training data labeled with distant supervision. We present experimental results from training a relation classifier, where our method shows significant improvement over the DS baseline, as well as just adding the labeled examples to the train set. Unlike Sterckx et al. [119] who employ experts to label the dependency path representation of sentences, our method uses the general crowd to annotate the actual sentence text, and is thus easier to scale and not dependent on methods for extracting dependency paths, so it can be more easily adapted to other languages and domains. Also, since the semantic label propagation is applied to the data before training is completed, this method can easily be reused to correct train data for any model, regardless of the features used in learning.

In future work, we plan to use the label propagation method to correct training data for state-of-the-art models in relation classification, but also relation extraction and knowledge-base population. We also plan to explore different ways of collecting and aggregating data from the crowd. CrowdTruth [42] proposes capturing ambiguity through inter-annotator disagreement, which necessitates multiple annotators per sentence, while Liu et al. [85] propose increasing the number of labeled examples added to the training set by using one high quality worker per sentence. We will compare the two methods to determine whether quality or quantity of data are more useful for semantic label propagation. To achieve this, we will investigate whether disagreement-based metrics such as sentence and relation quality can also be propagated through the training data. We believe a more continuous truth measure as opposed to the rather arbitrary discrete measure will be productive for this evaluation. Finally, we are particularly excited about the possibility of using our approach in conjunction with logical reasoning approaches such as those reported in [35]. In this case, we are looking at more informed data that reflects human understanding and properties of the data set, to discover candidate relation pairs for investigating rules.

APPENDIX: EXAMPLE SENTENCES WITH WRONG DISTANT SUPERVISION LABELS

SENTENCE	RELATION	CROWD SRS	DS LABEL
After growing up on Cat Island, **Tony McKay** moved to **New York City** at age 17 to study architecture.	*place of death*	0.004	1
	places of residence	0.995	1
The film is based very loosely on the lives of **Wolfgang Amadeus Mozart** and Antonio Salieri, two composers who lived in **Vienna, Austria**.	*place of death*	0.074	1
	places of residence	0.865	1
Marku Ribas is the side more Black music of this group and was **Bob Marley**'s friend in the 1970s, **Jamaica**, where he lived.	*origin*	0	1
	places of residence	0.87	1
Osama bin Laden had moved from **Saudi Arabia** to Sudan during the 1990-91 Gulf War.	*origin*	0.3	1
	places of residence	0.74	1

Table 18: Example sentences with false positive *place of death* and *origin* DS labels due to multiple relations in the KB over *Person - Location* term types.

Sentence	Relation	Crowd SRS	DS label
China on Monday officially appointed **Donald Tsang** as **Hong Kong**'s chief executive for a second term.	*employee or member*	0.623	0
	top employee or member	0.753	1
More than 3,000 taxi drivers blocked **Rome**'s historic centre Wednesday to protest extra licences given by mayor **Walter Veltroni**.	*employee or member*	0.529	0
	top employee or member	0.841	1
Early years **Joey Harrington** was born and raised in **Portland, Oregon**, where he has resided his entire life.	*origin*	0.645	0
	place of birth	0.867	1
Nelli Zhiganshina (born March 31, 1987 in **Moscow**, Russia) is a Russian ice dancer who currently represents Germany.	*origin*	0.555	0
	place of birth	0.791	1

Table 19: Example sentences with false negative *employee or member* and *origin* DS labels due to missing causal connections.

5 FINDING AMBIGUITY FROM DISAGREEMENT

Everything is vague to a degree you do not realize until you have tried to make it precise, and everything precise is so remote from everything that we normally think, that you cannot for a moment suppose that is what we really mean when we say what we think.
– Bertrand Russell, The Philosophy of Logical Atomism

In this chapter, we investigate how inter-annotator disagreement can be used as an indicator for language ambiguity, using the task of FrameNet frame disambiguation as a use case. FrameNet is a computational linguistics resource composed of semantic frames, high-level concepts that represent the meanings of words. In this chapter, we present an approach to gather frame disambiguation annotations in sentences using a crowdsourcing approach with multiple workers per sentence to capture inter-annotator *disagreement*. We perform an experiment over a set of 433 sentences annotated with frames from the FrameNet corpus, and show that the aggregated crowd annotations achieve an F1 score greater than 0.67 as compared to expert linguists. . This methodology was then scaled up to collect a frame disambiguation resource over 5,000 sentence-word pairs from Wikipedia – the largest corpus of this type outside of FrameNet.

A qualitative examination of the disagreement in our data revealed cases where the crowd annotation was correct even though the expert is in disagreement, arguing for the need to have multiple annotators per sentence. Most importantly, we examine cases in which crowd workers could not agree, and demonstrate that these cases exhibit ambiguity, either in the sentence, frame, or the task itself, and argue that collapsing such cases to a single, discrete truth value (i.e. correct or incorrect) is inappropriate, creating arbitrary targets for machine learning.

This chapter is based on the following publications:

- *Capturing Ambiguity in Crowdsourcing Frame Disambiguation*, in the AAAI Conference on Human Computation and Crowdsourcing, co-authored by Lora Aroyo and Chris Welty [43];
- *A Crowdsourced Frame Disambiguation Corpus with Ambiguity*, in submission, co-authored by Lora Aroyo and Chris Welty.

5.1 INTRODUCTION

We have shown that preserving inter-annotator disagreement can result in ground truth data of a high quality (Chapters 2 & 3), that can be used to improve the performance of natural language processing systems (Chapter 4). Based on these results, it appears that inter-annotator disagreement is a useful property to have in ground truth data. We argue that is because *disagreement is often times indicative of ambiguity that is inherent to natural language* (**RQ4**). In this chapter, we explore how disagreement can be used as an indicator for language ambiguity, using the task of FrameNet frame disambiguation as a use case.

FrameNet is a computational linguistics resource based on the frame semantics theory [12]. A semantic *frame* is an abstract representation of a word sense, describing a type of entity, relation, or event, and identifies the associated *roles* implied by the frame. The FrameNet resource offers a collection of semantic frames, together with a corpus of documents annotated with these frames. In the corpus, individual words are mapped to the single frame that represents the meaning of that word in the sentence.

Since many words have multiple possible meanings, the task of obtaining these annotations is called *frame disambiguation*, similarly to word-sense disambiguation. It is a complex task that typically is performed by linguistic experts, subjected to strict annotation guidelines and quality control [11]. As such, this task typically does not scale sufficiently in order to meet the annotation requirements of modern machine learning methods. Moreover, the annotation is typically performed by only one expert, which makes it impossible to capture any diversity of perspectives.

There have been a number of attempts at using crowdsourcing for frame disambiguation in sentences, such as those by Hong and Baker [60] and Chang et al. [25], offering a creative way to deal with the complexity of the annotation task. This chapter addresses the considerable problem of *ambiguity* in frame annotation, which we show to be a prominent feature in frame semantics. We adapt the CrowdTruth framework, which encourages using multiple crowd annotators to perform the same work, and processes the disagreement between them to signal low quality workers, sentences, and frames.

This chapter presents the following contributions:

1. *annotated corpus*: 433 FrameNet sentences [47], and 5,000 Wikipedia sentences with crowd annotations;

2. *crowd vs. expert evaluation*: the crowd achieves comparative quality with trained FrameNet experts (F1 $>$ 0.67), and we provide examples of typical cases where the crowd annotation is correct despite the expert disagreement;

3. *metrics for frame and sentence quality*: a qualitative evaluation showing that inter-annotator disagreement is an indicator of ambiguity in both frames and sentences.

4. *ambiguity-aware annotation methodology*: we demonstrate that the cases in which the crowd workers could not agree exhibit ambiguity, either in the sentence, frame, or the task itself; we argue that collapsing such cases to a single, discrete truth value (i.e. correct or incorrect) is inappropriate, creating arbitrary targets for machine learning;
5. *evaluation of several frame disambiguation models*: using evaluation metrics that leverage the multiple possible frames per sentence and their confidence scores, we show that even a model that always predicts the top crowd answer will not always have the best performance.

5.2 RELATED WORK

This work relates to the state of the art in two areas of research: (1) various crowdsourcing approaches for FrameNet related tasks, and (2) dealing with ambiguity and disagreement in crowdsourcing. Below we provide an overview of the research on which we base or inspire our approach.

5.2.1 *Crowdsourcing FrameNet*

Hong and Baker [60] first experimented with applying crowdsourcing for frame disambiguation, where the authors were able to achieve an accuracy of 0.982 as compared to the expert annotators. We replicate the performance of the crowd from this research in our experiments. Moreover, we also measure the inter-annotator disagreement which we show is a useful indicator of ambiguity in both sentences and frames. Fossati, Giuliano, and Tonelli [57] extend the frame disambiguation task with identifying frame roles (roles are the elements of the semantic frame, e.g. participants in an event).

More recently, Chang et al. [25] proposed a method for supervised crowdsourcing of frame disambiguation, where after an initial step of picking the best frame for a word in a sentence, the crowd worker receives feedback from the other annotators, and can then decide if they want to change their annotation or not. This serves to correct misunderstandings of the frame definition by the crowd. Pavlick et al. [102] use automatic paraphrasing to increase the lexical coverage of FrameNet, where crowdsourcing is employed to manually filter out bad paraphrases.

Similarly to our claim, Jurgens [71] argues that ambiguity is an inherent feature of frame/word sense disambiguation, and that crowdsourcing can be used to capture it. The crowd is asked to annotate on a Likert scale the degree to which a sense applies to a word. As Likert scales have been shown to be unreliable for capturing subjective measures [74], our annotation task is composed of quantifiable binary questions (i.e. does the frame apply to the word in the sentence or not?), and the ambiguity is captured by giving the same examples to multiple workers and measuring disagreement [7].

In our experiments we found between 10-15 workers provided the most reliable results (the more complex the task, the more workers are needed). Thus, we employ 15 annotators per task in our experiments in order to ensure we capture sufficient diversity of interpretations, compared to 10 by Hong and Baker [60] and 3 by Jurgens [71].

5.2.2 *Disagreement and Ambiguity in Crowdsourcing*

Our work is part of a continuous effort in exploring the inter-annotator disagreement as an indicator for (1) inherent uncertainty in the domain knowledge as Cheatham and Hitzler [27] found when assessing the Ontology Alignment Evaluation Initiative (OAEI) benchmark, (2) debatable cases in linguistic theory, rather than faulty annotation, as Plank, Hovy, and Søgaard [105] found in their part-of-speech tagging task, and (3) ambiguity inherent in natural language [14].

In our own work, we have primarily been interested in ambiguity at the sentence level and in the target semantics [45]. The CrowdTruth project has made software available [65] to process vector representations of crowd gathered data that *encourages disagreement*, in a more continuous representation of truth. We replicated our approach from other semantic interpretations tasks to the frame disambiguation task.

Finally we note recent efforts to consider in ground truth corpora (1) the notion of uncertainty, where Schaekermann et al. [113] also use disagreement in crowdsourcing for modeling it, (2) the notion of ambiguity, where Chang, Amershi, and Kamar [23] found that ambiguous cases cannot simply be resolved by better annotation guidelines or through worker quality control, and (3) the notion of noise, where Lin and Weld [84] show that machine learning classifiers can often achieve a higher accuracy when trained with noisy crowdsourcing data.

5.3 CROWDSOURCING SETUP

5.3.1 *Dataset*

The dataset used in this experiment consists of sentence-word pairs from the FrameNet corpus from release 1.7 (the latest one at the time of writing), where the given word in the sentence has been labeled with a frame by expert annotators. We selected a word in each sentence and constructed a list of candidate frames to show to the crowd (Fig. 14). To do this, we used the Framester corpus [58], which maps FrameNet semantic frames to synonym sets from WordNet [91]. First, the sentences were processed with tokenization, sentence splitting, lemmatization and part-of-speech tagging. Then each word with a frame attached to it was matched with all of its possible synonym sets from WordNet, while making sure that the part-of-speech constraint of the synonym set is fulfilled. Using the WordNet mapping, we constructed a list of possible frames for each word with an expert annotation. From this dataset, we randomly selected 433 sentence-word pairs, containing 341 unique frames and 300 unique words after lemmatization, that respect the following conditions:

- The word has a part-of-speech of either a *noun* or a *verb*.
- Each word has *at least two and no more than 20 candidate frames*.

The restriction on the maximum number of frames was done so as not to overwhelm the crowd with too many choices. However, annotating words that have more than 20 frames can easily be adapted for our template, by fragmenting the candidate frame list into several parts and running the task multiple times. Also, having just one frame per word means that the crowdsourcing task becomes one of validation, not disambiguation, so the restriction on the minimum number of frames was put in place.

For simplicity, we refer to the sentence-word pairs as sentences in the rest of the chapter. This dataset, as well as the crowdsourcing results and aggregated metrics are available online [47].

5.3.2 *Task Template*

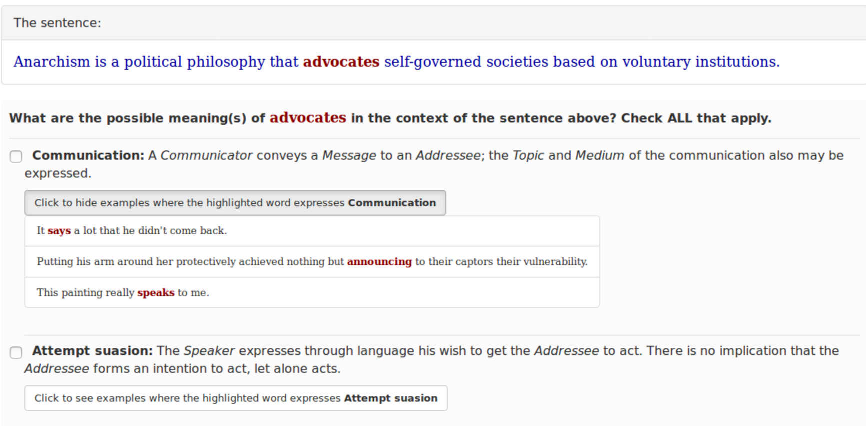


Figure 14: Fragment of the crowdsourcing task template.

The crowdsourcing task was run on the Amazon Mechanical Turk platform[1]. The task template is shown in Figure 14. The workers were given a sentence with the word highlighted, and then asked to perform the multiple choice task of selecting all frames that fit the

1 `https://mturk.com/`

sense of the highlighted word, or that none of the frames fit. The most challenging part of the frame disambiguation task design is making sure that the crowd can understand the meaning of the frame. For each frame, we show the definition, as well as a list of sentences exemplifying the usage of the frame. These example sentences can be accessed by the workers by clicking a button next to each frame, so that the workers do not become overwhelmed with the information on the task page. In order to make sure we capture diverse worker opinions, we increased the number of annotators per sentence from 10 (the number recommended by Hong and Baker [60]), to 15. The cost of the task varied from $0.08 per annotation at the start of the task, in order to attract a sizable pool of workers, to $0.06 at the end, as workers became quicker at solving the task.

5.3.3 *Disagreement Metrics*

To aggregate the results of the crowd, while also capturing inter-annotator disagreement, we use a modified version of the Crowd-Truth [7] metrics. The first step is to construct the *worker vectors*, which are a set of binary vectors encoding the decision of one worker for one sentence. The vector has $n+1$ components, where n is the number of frames shown together with the sentence. If the worker selects a frame from the multiple-choice list, its corresponding component would be marked with '1', and '0' otherwise. The decision to pick none of the frames also corresponds to a component in the vector. Using these worker vectors, we then calculate the following disagreement metrics:

- **frame-sentence score (FSS):** the degree with which a frame matches the sense of the word in the sentence. It is the ratio of workers that picked the frame to all the workers that read the sentence, weighted by the worker quality (WQS). A higher FSS should indicate that the frame is more clearly expressed in a sentence.

- **sentence quality (SQS):** the overall worker agreement over one sentence. It is the average cosine similarity over all worker vectors for one sentence, weighted by the worker quality (WQS) and frame quality (FQS). A higher SQS should indicate a clear sentence.

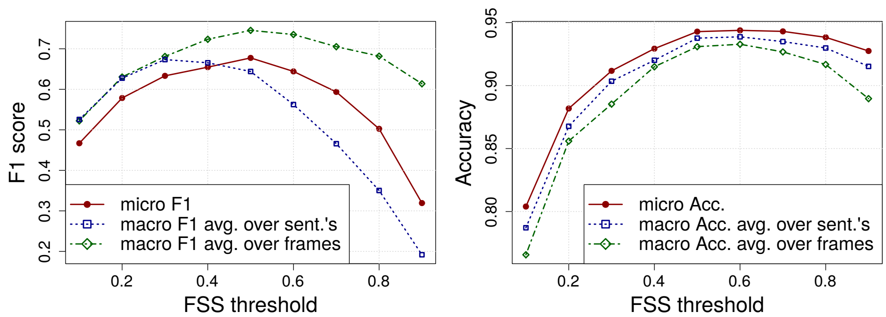
- **frame quality (FQS):** the agreement on a frame in all sentences that it appears. Given frame f, $FQS(f) = avg(FSS(f,s) | FSS(f,s) > 0)$. FQS is also weighed by the quality of the workers and the sentences. A higher FQS should indicate a clear frame semantics.

- **worker quality (WQS):** the overall agreement of one crowd worker with the other workers, calculated using average cosine

#	Sentence	Frame	FSS
S1	Shops *aimed* at the tourist market are interspersed with the more workaday ironmongers.	*aiming*	0.808
		purpose(*)	0.288
S2	The major *changes* were not to daily tasks and routines , but to the political power base.	*cause change*	0.804
		undergo change(*)	0.305
S3	This *investigation* has been stymied stopped, obstructions thrown up every step of the way.	*criminal investigation*	0.898
		scrutiny(*)	0.377
S4	Does supersizing *cause* obesity?	*cause to start*	0.804
		causation(*)	0.608
S5	The loud, raucous Jamaican English dialect and the *waving* hands reflect the joy with which social relations are conducted here.	*body movement*	0.861
		gesture(*)	0.463
S6	The Intifada *heralded* the rise of the Muslim fundamentalism.	*heralding*	0.777
		omen(*)	0.227
S7	Fish (heads discreetly *wrapped* in paper) are still hung out to dry in the sun.	*adorning*	0.31
		filling(*)	0.278

Table 20: Example sentence-word pairs where the top crowd frame choice is different than the expert. The targeted word appears in italics font in the sentence. The frame picked by the expert is marked with (*).

similarity with other workers per sentence, and weighted by the sentence and frame qualities.



(a) F1 score (b) Accuracy

Figure 15: Crowd evaluation results, using expert annotation as correct.

These definitions are mutually dependent, e.g. the definition of SQS depends on the FQS and WQS, the intuition being that low quality workers should not make sentences look bad, and low quality sentences should not make workers look bad, etc. The mutual dependence requires an iterative dynamic programming approach, which converged in numerous applications in fewer than 8 iterations.

5.4 CROWD VS. EXPERTS

To evaluate the quality of the crowd annotations, we iterate through different values of thresholds in the FSS to classify a frame-sentence pair as either positive or negative, then compare the results with the annotations of the FrameNet experts. The results for both the micro (i.e. each frame-sentence pair is counted as either true positive, false positive etc. and used in the calculation of the F1 and accuracy) and macro (the F1 and accuracy are calculated for each sentence and each frame, and then averaged into the final values) scores are presented in Figures 15a & 15b.

At the best FSS threshold, the accuracy scores are comparable to those presented by Hong and Baker [60], who report an average accuracy of 0.928, although on a different dataset. However, accuracy in multi-class classification problems are unreliable as there are high numbers of true negatives. The F1 score is likely a more reliable metric of the performance of the crowd, with scores > 0.67 for all 3 versions of the F1. Finally, an ANOVA test over the paired FSS and expert decision for a frame-sentence pair resulted in the $F - value = 4597$ and $p < 2e^{-16}$, proving that there is a statistically significant relationship between the crowd FSS and the decision of the expert.

While the majority of expert choices have high FSS scores, there are some exceptions. We observed 3 different causes for this disagreement, which are exemplified in Table 20:

1. The crowd *misunderstood the frame definition*. For instance, in $S1$, the crowd mistook the *aiming* frame to mean purpose, instead of the more literal meaning of the frame of adjusting an instrument to reach a target. In $S2$, the crowd correctly identifies a causal sense, but the correct interpretation is a passive change (*changes [...] to the political power*) instead of the active change (i.e. a subject is doing the changing) that is picked by the crowd.

2. The *information in the sentence is incomplete* to identify the correct frame. $S3$ does not express whether the investigation is criminal in nature, although that is a possible interpretation. This represents a limitation in the design of the crowdsourcing task – in some versions of the expert task, annotators had the full context of the document available when performing the annotations. This could be fixed or reduced by providing the sentence before and after, without overloading the workers.

3. The crowd offers a *legitimate alternative interpretation* of what the correct frame should be. In $S5$ the crowd picks the more general frame *body movement* for *waving*, while in $S4$ and $S6$, the crowd picks more specific interpretations than the expert (*cause to start* for the *obesity* effect instead of the broader sense of *causation* in $S4$, and *heralding* instead of *omen* for the word *heralding* in

#	Sentence	SQS	Frame	FSS
P1	Egypt has provided no evidence demonstrating the *elimination* of its biological warfare ability, which has existed since at least 1972.	0.841	*removing*$^{(*)}$ *cause change* *event*	0.938 0.175 0.032
P2	First, he forbade seeking the aid of infidels when the Syrian Mujahiddin asked Saddam Hussein to *overthrow* the regime of Hafiz Al-Assad in Syria.	0.669	*change of leadership*$^{(*)}$ *removing* *eventive cognizer affecting* *people*	0.847 0.539 0.087 0.005
P3	Their influence helped draw a line in the desert sand between legitimate operations and mob casinos, where illegal *skimming* of profits was rampant.	0.366	*removing*$^{(*)}$ *theft* *committing crime* *misdeed* *cause change*	0.532 0.494 0.459 0.431 0.273
P4	The above mentioned protection *procedures* are only for observation purposes, while patrols check the fences, the barriers, and the towers.	0.786	*means*$^{(*)}$ *being employed*	0.889 0.11
P5	We've expanded Goodwill's proven *methods* to towns and neighborhoods where they are needed most.	0.364	*means*$^{(*)}$ *expertise* *domain* *fields*	0.601 0.342 0.173 0.131
P6	The latest *approach* is perhaps the best of the post-mob era : the comprehensive resort.	0.208	*means*$^{(*)}$ *conduct* *path traveled* *communication*	0.457 0.225 0.159 0.121
P7	Prime Minister Ariel Sharon of Israel *urged* President Bush to step up pressure on Iran to give up all elements of its nuclear program.	0.528	*attempt suasion*$^{(*)}$ *request* *communication* *cause to start*	0.81 0.387 0.337 0.115
P8	The security team should *urge* everyone to take precautions and guard their homes tightly.	0.358	*attempt suasion*$^{(*)}$ *request* *cause to start* *communication*	0.605 0.321 0.256 0.213
P9	The security team should publish a periodic bulletin and distribute to all residents, *advising* them how to safely store gaz and logs.	0.386	*attempt suasion*$^{(*)}$ *communication* *expertise* *request*	0.576 0.567 0.167 0.156

Table 21: Different FSS values for the frames *removing* (P1, P2, P3), *means* (P4, P5, P6), *attempt suasion* (P7, P8, P9). The targeted word appears in italics font in the sentence. The frame picked by the expert is marked with $^{(*)}$.

*S*6). *S*7 shows an example where the expert made a mistake, as *filling* refers to the action of covering an area with something, whereas *adorning* refers to the passive act of being covered.

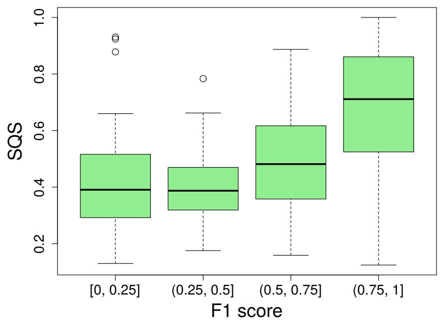
5.5 CAPTURING AMBIGUITY

The cases where the experts and crowd disagree exemplify how difficult frame disambiguation can be when dealing with ambiguity, both in sentences and in the frame definition. Currently in the FrameNet corpus, the expert annotations lack the level of granularity necessary to differentiate between clear expressions of the frames, and more ambiguous ones. We propose the FSS metric as a method to capture the degree of ambiguity with which a frame captures a word sense in a sentence. In Table 21, we show how the FSS metric varies together with the clarity with which a frame is expressed across different sentences. We demonstrate this across 3 different frames:

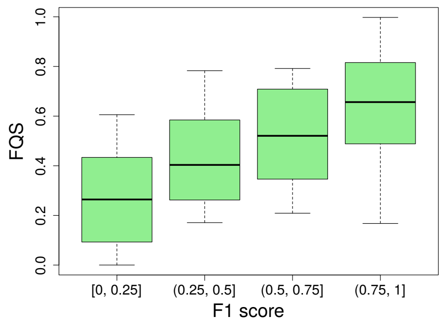
- *removing*: *P*1 is an unambiguous expression of the frame, as reflected by the high agreement score. In *P*2, the top crowd frame as well as the expert choice frame *change of leadership* refers to overthrowing the government, and *removing* can be read as a generalization of this sense (i.e. removing the government by overthrowing it) – *removing* is a valid interpretation, but less specific, and the lower FSS seems justified. *P*3 is an even more ambiguous case – it is not clear whether the word *skimming* refers to generally *committing crime*, or to the more specific crime of *theft*, and *removing* is a generalization for the sense of *theft*, however skimming here is a common metaphor, and not the actual act of skimming. We claim the rank ordering of uses of the *removing* frame here is sensible, moreover it is far more useful to capture this information than require a single discrete truth value - the third case is simply not as clear a usage of the frame as the first. There is a certain arbitrariness to determining which of these is "truly removing" and which is not.

- *means*: This frame refers to the means used by an agent to achieve a purpose. While *P*4 offers an unambiguous expression of the frame, in *P*5 the means with which to achieve a goal becomes confused with the expertise and knowledge required to achieve it. In *P*6 the goal is not mentioned, therefore creating confusion about the purpose of the *approach*, and whether it might refer to a way of communicating or behaving. Again, we claim this rank ordering is more informative than requiring a discrete judgment on each case.

- *attempt suasion*: This frame refers to a speaker attempting to influence the addressee to act. Sentences *P*7 to *P*9 express various



(a) SQS in relation to F1 score (with expert annotations as true positives), shows that in higher quality sentences, the crowd tends to agree with experts.



(b) FQS in relation with F1 score (with expert annotations as true positives), shows that in higher quality frames, the crowd tends to agree with the experts.

Figure 16: SQS & FQS evaluation.

degrees of persuasion, from obviously to weakly expressed. In *P7*, it is clear that the attempt at persuasion is an event that has occurred (*Sharon [...] urged*). *P8* expresses an obligation at an attempt to persuade (*should urge*), whereas in *P9* the persuasion is weaker, merely *advice*.

In addition to the ranking, the method of collecting data from multiple crowd workers yields alternate interpretations, which are also quite useful. Consider that a common motivation for collecting annotated data is to train and evaluate deep learning models, many of which produce vectors of output (frame disambiguation can be implemented as a multi-class problem). Our methods of gathering annotations are naturally suited to multi-class objectives.

The SQS and FQS metrics can additionally be used to express the overall ambiguity in the sentence and frame, respectively. Figures 16a & 16b show that sentences with higher SQS and frames with higher FQS also have higher F1 values, demonstrating that the SQS and FQS metrics can be useful in determining data quality. This result, in combination with the correlation between FSS and expert annotations, shows that when there is agreement in the crowd, then the crowd also agrees with the experts, but when there is disagreement, it may be because something is wrong: with the workers, the sentence, or the frames.

In Table 22, we show some examples of how SQS captures the clarity for the sense of a word in a sentence, by taking the same word (and therefore same list of candidate frames) in different sentences:

- Sentences *Q1*, *Q2* and *Q3* all contain the word *construct*, with different degrees of clarity. When the object being constructed is a building (i.e. the *Temple* in *Q1*), there is no ambiguity in selecting the *building* frame, but when the object is a *road bump* (*Q2*),

#	Sentence	SQS	Frame	FSS
Q1	Although David bought the land for the Temple and carefully assembled its building materials, he was deemed unworthy of *constructing* the Temple.	0.711	*building*$^{(*)}$	0.925
			manufacturing	0.183
			create physical artwork	0.056
Q2	Passageways for cars and pedestrians should be designated 4-Road bumps: Six successive bumps should be *constructed* at 500 meters from the location.	0.542	*building*$^{(*)}$	0.768
			manufacturing	0.326
			create physical artwork	0.089
Q3	*Constructed* in wood, brick, stone, ceramic, and bronze, this is a work of extravagant beauty, uniting many ancient art forms.	0.351	*building*$^{(*)}$	0.515
			create physical artwork	0.335
			manufacturing	0.237
Q4	U.S. Congressman Tony Hall arrived here Sunday evening, *becoming* the first U.S. lawmaker to visit Iraq since the 1991 Gulf War.	0.901	*becoming*$^{(*)}$	0.995
			cause change	0.24
			undergo change	0.212
Q5	Cheung Chau *becomes* the center of Hong Kong life once a year, usually in May , during the Bun Festival, a folklore extravaganza.	0.562	*becoming*$^{(*)}$	0.783
			undergo change	0.783
			cause change	0.402
Q6	Are there any *efforts* to bring back small investors?	0.811	*attempt*$^{(*)}$	0.926
			commitment	0.178
Q7	At AOL there was a conscious *effort* to develop other "characters," for lack of a better word.	0.588	*attempt*$^{(*)}$	0.739
			commitment	0.468

Table 22: Sentence Quality Score Examples. The targeted word appears in italics font in the sentence. The frame picked by the expert is marked with $^{(*)}$.

the sense of the building *frame* becomes difficult to separate from *manufacturing*. In Q3, the object of the construction is not expressed, but the construction materials imply a precious object, therefore *building*, *manufacturing* and *create physical artwork* are all possible interpretations. Sentences

- Q4 and Q5 illustrate the variation in clarity for the word *become*. While in Q4, the sense *becoming* is the unambiguous choice, in Q5 it is difficult to choose between the frames *becoming* and *undergo change* (it is arguable that *Cheung Chau* needs to undergo some form of change in order to become a center).

Frame	FQS	Definition	Example Sentences	FSS
killing	0.954	A Killer or Cause causes the death of the Victim.	*F*1: Older kids left homeless after a recent murder-*suicide* in Indianapolis claimed Mom and Dad.	0.8
			*F*2: The incident at Mayak was the third *shooting* in recent weeks involving nuclear weapons or facilities in Russia.	0.75
food	0.838	Words referring to items of food.	*F*3: Lamma Island is perfect for sitting back to watch *bananas* grow.	1.0
			*F*4: Along with the usual *chickens*, you will see for sale snakes, dogs, and sometimes monkeys - all highly prized delicacies .	0.838
			*F*5: You can browse among antiques, flowers, *herbs*, and more.	0.503
assistance	0.634	A Helper benefits a Benefited party by enabling the culmination of a Goal of the Benefited party.	*F*6: Your support *helps* provide real solutions.	0.955
			*F*7: Unemployment *provides* benefits that many entry-level jobs don't.	0.467
			*F*8: Your support of Goodwill will *provide* job training.	0.401
purpose	0.63	An Agent wants to achieve a Goal. A Means is used to allow the Agent to achieve a Goal.	*F*9: The *objective* of having kiosks is they serve as communication points between the guards	0.94
			*F*10: They are antiviral drugs *designed* to shorten the flu.	0.476
			*F*11: It seems that the city produced artists of this stature by accident, even against its *will*.	0.241
subjective influence	0.366	An Agent has influence on a Cognizer. The influence may be general, manifested in an Action as a consequence of the influence.	*F*12: There have been changes, many of them due to economic progress, new construction, and other factors that *influence* cities.	0.54
			*F*13: The Cycladic culture was *influenced* by societies in the east.	0.46
			*F*14: Their complaint: the system *discourages* working.	0.364
undergo change	0.313	An Entity changes, either in its category membership or in terms of the value of an Attribute.	*F*15: The animosity between these two traditional enemies is beginning to *diminish*.	0.805
			*F*16: The *shift* in the image of Gates has been an interesting one for me to watch.	0.351
			*F*17: The settlements of Thira and Akrotiri *thrived* at this time.	0.256

Table 23: Frame Quality Score Examples. The targeted word appears in italics font in the sentence.

- *Q*6 and *Q*7 both deal with the word *effort*. In *Q*7, however, the *conscious* qualifier for the word *effort*, as well as the goal to *develop*, implies a sustained, long-term action that can be understood as either an *attempt* or a *commitment* to achieve a goal. In contrast, *Q*6 expresses a short-term, concrete action (to *bring*), which more closely fits the sense of the frame *attempt*.

Again, our claim is that these scores and ranking are far more sensible and informative than requiring a discrete truth decision, which seems more arbitrary as the scores decrease.

As the examples above indicate, one possible cause for sentence ambiguity is missing context information (e.g. in *Q*3). This was also one of the causes for disagreement between crowd and expert. A solution to this problem would be to expand the input text for the crowdsourcing task, to include the full paragraph, or even just one sentence before and one after the one we want the crowd to annotate.

Another reason for sentence ambiguity is frames that overlap in meaning (e.g. in *Q*5 and *Q*7). While providing more context could help with this, it is often the case that even the definitions of the frames are very close. The FQS metric is a useful indicator for these case.

Table 23 shows varying FQS values for different frames, from very clear to ambiguous. The frame *subjective influence*, with an FQS of 0.366, has a low score compared to the others. From looking at the sentences, we observed that the crowd had difficulty distinguishing between this frame and *objective influence*. The difference between these two frames is very small – *subjective influence* means a general, vague type of influence, whose effect cannot be measured, whereas *objective influence* refers to a more concrete type of influence. However, as we see from the example sentences in Table 23, these cases can be very difficult to separate in natural language (e.g. in *F*13 is *cultural influence* subjective or objective?).

Another feature we observed was the correlation of FQS with how abstract the sense of the frame is. Frames with high FQS, such as *killing* and *food*, tend to refer to concrete events or objects. These frames can still appear in ambiguous contexts (e.g. in *F*5, it is not clear whether *herbs* classify as a type of *food*), but overall these frames refer to specific and particular senses that are unambiguous. As the value of the FQS metric goes down, the frames become more abstract. *assistance* and *purpose* both have example sentences where they are expressed unambiguously (*F*6 and *F*9), but their definitions are more abstract, and therefore have more room for interpretation. For instance, providing benefits (in *F*7) or expertise (in *F*8) can be regarded as a type of help, or *assistance*, even though the expert picked the more literal sense of the frame *supply* for both of these cases. Likewise the frame *purpose* can be understood in *F*10 as the purpose of a design (the expert picked the more literal *coming up with*), or in *F*11 as the goal of the desire/will (the expert picked *desiring*). *undergo change*, the frame with the lowest FQS in Table 23 has a very broad meaning, and is a generalization of other more specific frames: *change position on a scale* in *F*16, and *thriving* in *F*17.

As we have seen from these examples, ambiguity in frames is connected to ambiguity in sentences. Frames with abstract or overlapping definitions are likely to appear in ambiguous sentences, and missing context from sentences is likely to result in more ambiguous scores for the frames. While workers misunderstanding the task is also a confounding factor that adds to the noise in the data, it is clear that there are many instances where inter-annotator disagreement is legitimately a by-product of ambiguity. This is an issue with the FrameNet dataset, as it does not allow for expressing the various degrees with which a sense applies to a word in a sentence, and instead relies on binary labels (i.e. the frame is expressed or not). This results in a loss of information that could impact the various natural language processing and machine learning applications that make use of this corpus, as it sets false targets for optimization – i.e. it seems unfair to expect a model to differentiate between highly ambiguous examples, when even human annotators are having such difficulty with them.

5.6 A FRAME DISAMBIGUATION CORPUS WITH AMBIGUITY

Following from the encouraging results of crowdsourcing the FrameNet corpus, we scaled up our method and collected a corpus of 5,000 sentence-word pairs. More than 1,000 of these are lexical units not part of FrameNet. To our knowledge, it is the largest corpus of this type outside of FrameNet. To perform the collection, we re-used the crowdsourcing methodology described in Section 5.3, using Wikipedia as a source for the sentences. This corpus was then used to perform an evaluation of several frame disambiguation models. Our proposed evaluation methodology uses evaluation metrics that leverage the multiple answers and their confidence scores, showing that even a model that always predicts the top crowd answer will not always have the best performance.

5.6.1 *Ambiguity in the Corpus*

An analysis of the corpus found many examples of inter-annotator disagreement, of which a few examples are shown in Table 24. For 720 sentences, a majority of the workers picked at least 2 frames (examples 1-3 in Tab.24). And for 1,514 sentences, no one frame has been picked by a majority of the workers (examples 4-7 in Tab.24). Disagreement is also more prominent in the sentences where the lexical unit is not a part of FrameNet (Fig.17).

The disagreement comes from a variety of causes: a parent-child relation between the frames (*statement* and *communication* in #3), an

#	Sentence	*SQS*	Frames (*FSS*)
1	Domestication of plants has, over the centuries **improved** disease resistance.	0.652	*improvement or decline* (0.823), *cause to make progress* (0.683)
2	He is the 5th of 8 male players in history to **achieve** this.	0.626	*accomplishment* (0.764), *successful action* (0.709)
3	Albertus Magnus, a Dominican monk, **commented** on the operations and theories of alchemical authorities.	0.511	*communication* (0.522), *statement* (0.703)
4	He **slices** at Hector's armor, throwing him off guard and spinning him around.	0.319	*part piece* (0.499), *cause harm* (0.4), *cutting* (0.394), *attack* (0.254), *hit target* (0.227)
5	Another 46 steps **remain** to climb in order to reach the top, the "terrasse", from where one can enjoy a panoramic view of Paris.	0.308	*left to do* (0.497), *remainder* (0.478), *state continue* (0.319), *existence* (0.155)
6	Borzoi males frequently **weigh** more.	0.283	*assessing* (0.421), *dimension* (0.402), *importance* (0.128)
7	The dance includes bending and **straightening** of the knee giving it a touch of Cuban motion.	0.24	*reshaping* (0.495), *arranging* (0.356), *body movement* (0.298), *cause motion* (0.249)

Table 24: Example sentences with disagreement over the frame annotations (candidate word in bold).

overlap in the definition of the frames (*accomplishment* and *successful action* in #2), the meaning of the word is expressed by a composition of frames (in #7, "straightening of the knee" is a combination of *reshaping* the form of the knee, *arranging* the knee in the right position, and *body movement*), and combinations of all of these reasons (in #4, "slices" is a combination of *part piece* and *cause harm*, and the other frames are their children). More example sentences for each type of disagreement are available in the appendix. The sentences themselves are not difficult to understand, and it can be argued that all of them have one frame that applies the best for the word. The goal of this corpus is to show that next to this best frame for the word, there are other frames that apply to a lesser degree, or capture a different part of the meaning. When evaluating a model for frame disambiguation, it seems unfair to penalize misclassifications of frames that still apply to the word, but with less clarity, in the same way we would penalize a frame that captures a wrong meaning. Also, we argue that models should take into account that annotators do not agree over some examples, and treat them differently than clear expressions of frames. Disagreement can also be caused by worker mistakes (in #6, *dimension* refers to the size of the object, not the act of measuring the size). While we try to mitigate for this by weighing confidence scores with the worker quality, the mistakes still appear in the corpus. This type of disagreement could be useful in future work to identify examples that workers need to be trained on.

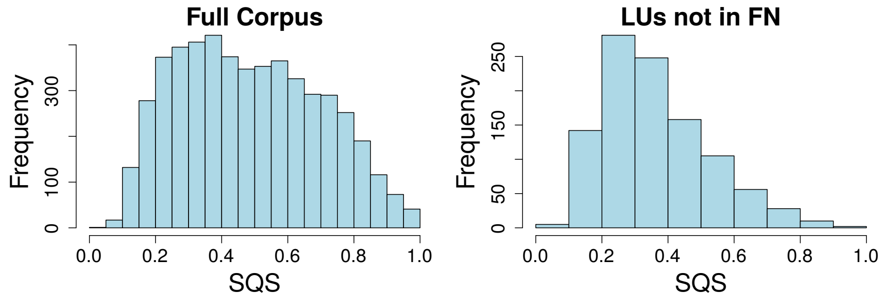


Figure 17: Histogram of SQS values - the quality scores in sentences where the lexical unit is not in FrameNet skew lower.

5.6.2 *Systems Tested*

As an example usage of our corpus, we used it to evaluate these frame disambiguation models:

1. **OS:** The Open-Sesame [121] classifier, pre-trained on the FrameNet corpus (release 1.7). Given a word-sentence pair, OS uses a BiLSTM model with a softmax final layer to predict a single frame for the word. If the lexical unit is not in FrameNet, it cannot make a prediction.

2. **OS+:** We modified the OS classifier to perform multi-label classification. To calculate the confidence score for candidate frame f, we removed the softmax layer and passed the output of the BiLSTM model $v(f)$ through the following transformation: $c(f) = [1 + tanh\ v(f)]/2$. This gave a score $c(f) \in [0, 1]$ expressing the confidence that frame f is expressed in the sentence.

3. **FS:** Framester includes a tool for rule-based multi-class multi-label frame disambiguation [58]. While for the dataset preprocessing (Sec. 5.3) we considered the frames for all synsets a word is part of, FS performs an additional word-sense disambiguation step to return a more precise list of frames. We used the tool with *profile T* as it was shown to have the overall better performance. FS can only predict FrameNet frames from the 1.5 release, which is missing 202 frames from version 1.7.

While OS+ produces confidence scores, the other methods produce binary labels for each frame-sentence pair. These models do not have state-of-the-art performance [54, 59], we picked them because they were accessible and allowed testing on a novel corpus. Finally, we evaluate the quality of the **TC** corpus, containing only the top frame picked by the crowd for every sentence. This test shows what is the best possible performance over our corpus that can be expected from a system such as OS that selects a single frame per sentence.

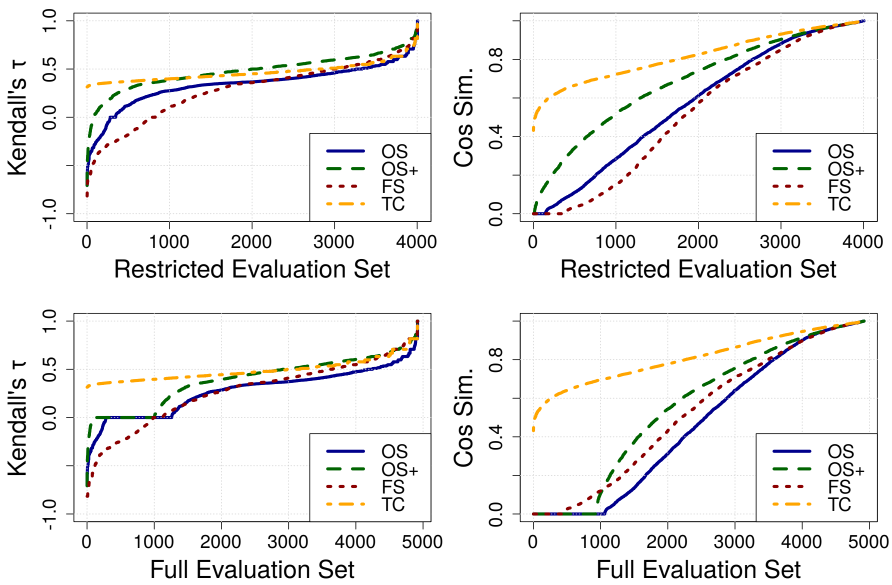


Figure 18: Baselines evaluation results.

5.6.3 *Evaluation Metrics & Results*

Instead of traditional evaluation metrics that require binary labels, we propose an evaluation methodology that is able to consider multiple candidate frames for each sentence and their quality scores. We use **Kendall's** τ list ranking coefficient [72] and **cosine similarity** to calculate the distance between the list of frames produced by the crowd labeled with the FSS, and the frames predicted by the baselines in each sentence. Whereas Kendall's τ only accounts for the ranking of the FSS for each frame, cosine similarity uses the actual FSS values in the calculation of the similarity. Both metrics compute a score per sentence (Kendall's $\tau \in [-1, 1]$, and cosine similarity $\in [0, 1]$). This is similar to the method used in [39]. Using these metrics, we produce two aggregate statistics over our test corpus: (1) the area-under-curve (AUC) for each metric, normalized by the corpus size, and (2) the SQS-weighted average of each metric ($w - avg$), which also accounts for the ambiguity of the sentence as expressed by the SQS. We evaluate on two versions of the corpus: (1) the restricted set (R-Set) of 4,000 sentences with lexical units from the FrameNet corpus, and (2) the full set (F-Set) of 5,000 sentences.

The results (Fig.18 & Tab.25) show that OS+ performs best out of all the models, even taking into account sentences with lexical units not in FrameNet for which OS+ cannot disambiguate. FS performs the worst out of all models on R-Set, because it cannot find newly added frames from the latest FrameNet release, but improves on the F-Set (FS can find candidate frames for lexical units not in FrameNet). The scores on the F-Set were lower for all baselines, suggesting that sentences with lexical units not in FrameNet are more difficult to classify – this

	Eval. Metric	OS	OS+	FS	TC
R-Set	Kendall's τ AUC	0.339	**0.477**	0.279	0.466
	Kendall's τ w-avg	0.362	**0.497**	0.3	0.48
	Cos Sim AUC	0.57	**0.685**	0.518	0.818
	Cos Sim w-avg	0.608	**0.717**	0.545	0.854
F-Set	Kendall's τ AUC	0.269	**0.379**	0.253	0.491
	Kendall's τ w-avg	0.307	**0.421**	0.284	0.501
	Cos Sim AUC	0.453	**0.544**	0.511	0.810
	Cos Sim w-avg	0.515	**0.607**	0.539	0.849

Table 25: Aggregated evaluation results.

could be because FrameNet is missing frames that can express the full meaning of these lexical units. TC has a good performance, but is far from being unbeatable – when measuring Kendall's τ over the R-Set, OS+ performs better than TC.

5.7 CONCLUSION

In this chapter, we explored how *inter-annotator disagreement can be used as an indicator for language ambiguity* for the task of FrameNet frame disambiguation. To achieve this, we employed the CrowdTruth [7] method, using multiple workers per sentence in order to capture and interpret inter-annotator disagreement. We modified CrowdTruth metrics in order to capture frame-sentence agreement (FSS), sentence quality (SQS) and frame quality (FQS). We performed an experiment over a set of 433 sentences annotated with frames from FrameNet corpus, and showed that the aggregated crowd annotations achieve an F1 score greater than 0.67 compared to expert linguists, and an accuracy that is comparable to the state of the art [60]. Afterwards, we scaled up the methodology to collect a frame disambiguation resource over 5,000 sentence-word pairs from Wikipedia, out of which 1,000 have lexical units that are new to FrameNet. This is the largest corpus of this type outside of FrameNet.

We showed cases where the crowd annotation is correct even though the expert is in disagreement, arguing for the need to have multiple annotators per sentence. Most importantly, we examined the cases in which crowd workers could not agree. We found that disagreement is caused by one or more of the following: workers misunderstanding the task, missing context from the sentences, frames with overlapping or abstract definitions. The results show a clear link between inter-annotator disagreement and ambiguity, either in the sentence, frame, or the task itself. We argue that collapsing such cases to a single, discrete truth value (i.e. correct or incorrect) is inappropriate, creating brittle, incomplete datasets, and therefore arbitrary targets for machine learning. We further argued that ranking examples by a score is informative, and that the crowd offers alternate interpretations that are often sensible.

Finally, we proposed an evaluation method that uses the scores for multiple frames, and is thus able to differentiate between frames that still apply to the word, but with less clarity, and frames that capture the wrong meaning. Our goal was to build a resource that recognizes different levels of ambiguity in the expression of the frames in the text, and allows a more fair evaluation of performance of frame disambiguation systems.

ACKNOWLEDGMENTS

We would like to thank Luigi Asprino, Valentina Presutti and Aldo Gangemi for their assistance with using the Framester corpus, as well as their advice in better understanding the task of frame disambiguation. We would also like to thank the anonymous crowd workers for their contributions to our crowdsourcing tasks.

APPENDIX: EXAMPLE AMBIGUOUS SENTENCES FROM THE FRAME DISAMBIGUATION CORPUS

#	Sentence	*SQS*	Frames (*FSS*)
1	These Articles have historically shaped and **continue** to direct the ethos of the Communion.	0.795	*activity ongoing* (0.862) *process continue* (0.86)
2	"A Modest Proposal" is **included** in many literature programs as an example of early modern western satire.	0.771	*inclusion* (0.89) *cause to be included* (0.813)
3	The states often **failed** to meet these requests in full, leaving both Congress and the Continental Army chronically short of money.	0.628	*endeavor failure* (0.826) *success or failure* (0.8)
4	This is a chart of trend of nominal gross domestic product of Angola at market prices **using** International Monetary Fund data.	0.598	*using resource* (0.831) *using* (0.554) *tool purpose* (0.336)
5	The Asian tigers have now all **received** developed country status, having the highest GDP per capita in Asia.	0.504	*receiving* (0.751) *getting* (0.556)
6	MasterCard has released Global Destination Cities Index 2013 with 10 of 20 are **dominated** by Asia and Pacific Region Cities.	0.467	*dominate situation* (0.638) *dominate competitor* (0.579) *being in control* (0.327)

Table 26: Ambiguity because of parent-child relation between frames.

#	SENTENCE	*SQS*	FRAMES (*FSS*)
1	Kournikova then **withdrew** from several events due to continuing problems with her left foot and did not return until Leipzig.	0.725	*withdraw from participation* (0.955), *removing* (0.61)
2	Some aikido organizations use belts to **distinguish** practitioners' grades.	0.68	*differentiation* (0.867) *distinctiveness* (0.703)
3	Since then, it has focused on **improving** relationships with Western countries, cultivating links with other Portuguese-speaking countries, and asserting its own national interests in Central Africa.	0.654	*improvement or decline* (0.787) *cause to make progress* (0.732)
4	To **emphasize** the validity of the Levites' claim to the offerings and tithes of the Israelites, Moses collected a rod from the leaders of each tribe in Israel and laid the twelve rods over night in the tent of meeting.	0.65	*emphasizing* (0.764) *convey importance* (0.638)
5	He not only had enough food from his subjects to **maintain** his military, but the taxes collected from traders and merchants added to his coffers sufficiently to fund his continuous wars.	0.453	*cause to continue* (0.7) *activity ongoing* (0.602)
6	He **spent** the later part of his life in the United States, living in Los Angeles from 1937 until his death.	0.29	*taking time* (0.41) *expend resource* (0.365)

Table 27: Ambiguity because of overlapping frame definitions.

#	SENTENCE	*SQS*	FRAMES (*FSS*)
1	These writings lack the mystical, philosophical elements of alchemy, but do contain the works of Bolus of Mendes (or Pseudo-Democritus), which **aligned** these recipes with theoretical knowledge of astrology and the classical elements.	0.284	*arranging* (0.474) *adjusting* (0.4) *assessing* (0.298) *compatibility* (0.254) *undergo change* (0.169)
2	However, commercial application of this fact has challenges in **circumventing** the passivating oxide layer, which inhibits the reaction, and in storing the energy required to regenerate the aluminium metal.	0.239	*dodging* (0.477) *compliance* (0.248) *surpassing* (0.204) no frame (0.148)
3	This had the effect of **inculcating** the principle of "Lex orandi, lex credendi" (Latin loosely translated as 'the law of praying [is] the law of believing') as the foundation of Anglican identity and confession.	0.201	*education teaching* (0.384) *communication* (0.35) no frame (0.153)
4	Legal segregation ended in the states in 1964, but Jim Crow customs often continued until specifically **challenged** in court.	0.172	*difficulty* (0.372) *competition* (0.283) *taking sides* (0.257) *communication* (0.154)
5	When Washington's army arrived outside Yorktown, Cornwallis prematurely abandoned his outer position, **hastening** his subsequent defeat.	0.134	*speed description* (0.39) *assistance* (0.209) *self motion* (0.165) *travel* (0.16) *causation* (0.124)

Table 28: Ambiguity because the meaning of the word is expressed by a composition of frames.

6 CONCLUSION

Man must not attempt to dispel the ambiguity of his being but, on the contrary, accept the task of realizing it.
– Simone de Beauvoir, The Ethics of Ambiguity

Have patience with everything that remains unsolved in your heart. ...live in the question.
– Rainer Maria Rilke, Letters to a Young Poet

This chapter summarizes the research presented in this thesis, by revisiting the research questions from the introduction. We also discuss the limitations of the current work, and identify future research directions to extend and compliment our findings on how to handle disagreement in ground truth for natural language processing.

6.1 RESEARCH QUESTIONS REVISITED

In this section, we consider again the research questions introduced at the beginning of this thesis. For each question, we provide possible answers, based on the research presented in this thesis.

RQ1: *Does allowing disagreement in crowdsourcing ground truth yield the same quality as asking domain experts?*

In Chapter 2, we studied this research question for the task of medical relation extraction. Using the CrowdTruth methodology for disagreement-preserving crowdsourcing, we collected a gold standard of 3,984 sentences expressing medical relations, focusing on the *cause* and *treat* relations. This data was used to train a sentence-level classification model. We have shown that allowing the disagreement in the crowd data does not mean that the quality of the ground truth has to suffer – the relation extraction models trained on crowd data performed just as well as the ones trained on annotations from medical experts, while the cost of collecting the data from the crowd was cheaper than for the experts.

In addition, our results show that, when the model reaches maximum performance after training, the crowd also performs better than distant supervision. Finally, we introduced and validated new weighted measures for precision, recall, and F-measure, that account for ambiguity in both human and machine performance on this task.

RQ2: *How does allowing disagreement in diverse crowdsourcing tasks influence the quality of the data?*

In Chapter 3, we studied the impact of inter-annotator disagreement on data quality for a set of diverse crowdsourcing tasks: closed tasks (*Medical Relation Extraction*, *Twitter Event Identification*), and open-ended tasks (*News Event Extraction* and *Sound Interpretation*). To do this, we employed an empirically derived methodology for efficiently gathering of human annotation by aggregating crowdsourcing data with CrowdTruth metrics. Our results showed that preserving disagreement in the annotations allows us to collect richer data, which enables reasoning about the ambiguity of the content being annotated. In all the tasks we considered, ambiguity-aware quality scores provide better ground truth data than the traditional majority vote. Finally, we showed that, contrary to the common crowdsourcing practice of employing a small number of annotators, adding more crowd workers actually can lead to significantly better annotation quality.

RQ3: *Can we improve the performance of natural language processing models by using disagreement-aware ground truth data?*
In Chapter 4 we perform several experiments using disagreement-aware ground truth to train and evaluate models for open-domain relation classification in sentences. Using the crowd data as ground truth, we have shown a very significant variation in the false positive rate in distant supervision data, and it seems extremely likely that this can be exploited to improve training. An initial experiment showed that cross-relation signals that were identified by the crowd can be used correct training data for relation classification. Next, we explored a more robust approach that propagates human annotations to sentences that are similar in a low dimensional embedding space. We showed that a small crowdsourced dataset of 2,050 sentences, collected and aggregated with the disagreement-preserving CrowdTruth methodology, can be successfully used to correct training data labeled with distant supervision, using a technique called "semantic label propagation". We have shown experimental results from training a relation classifier, where our method shows significant improvement over the distant supervision baseline, as well as just adding the labeled examples to the train set. Since the semantic label propagation is applied to the data before training is completed, this method can easily be reused to correct train data for other related models (e.g. to perform knowledge base completion), regardless of the features used in learning.

RQ4: *Is inter-annotator disagreement an accurate indicator for ambiguity in natural language?*
In Chapter5, we explore the relation between inter-annotator disagreement and natural language ambiguity for the task of frame disambiguation annotations in sentences. We performed an experiment over a set of 433 sentences annotated with frames from FrameNet

corpus, and showed that the crowd annotations aggregated with disagreement-preserving CrowdTruth metrics are comparable in quality to domain experts – the crowd achieves an F1 score greater than 0.67 compared to expert linguists, and an accuracy that is comparable to the state of the art [60]. Next, we scaled up the methodology to collect a resource of 5,000 sentence-word pairs, and 1,000 lexical units that are new to FrameNet – the largest corpus of this type outside of FrameNet. Finally, we proposed an evaluation method that uses the scores for multiple frames, and is thus able to differentiate between frames that still apply to the word, but with less clarity, and frames that capture the wrong meaning.

We also showed cases where the crowd annotation is correct even though the expert is in disagreement, arguing for the need to have multiple annotators per sentence. Most importantly, we examined the cases in which crowd workers could not agree. We found that disagreement is caused by one or more of the following: workers misunderstanding the task, missing context from the sentences, frames with overlapping or abstract definitions. The results show a clear link between inter-annotator disagreement and ambiguity, either in the sentence, frame, or the task itself. We argue that collapsing such cases to a single, discrete truth value (i.e. correct or incorrect) is inappropriate, creating brittle, incomplete datasets, and therefore arbitrary targets for machine learning. We further argued that ranking examples by a score is informative, and that the crowd offers alternate interpretations that are often sensible.

6.2 Limitations & Future Directions

The research presented in this thesis has several possible directions for future work. In addition to the limitations specific to the material in each chapter, we identify three overarching issues to be explored in the future work: (1) expanding the experimental work on capturing ground truth ambiguity beyond relation extraction and frame disambiguation, (2) optimizing for the cost of data collection, and (3) building natural language processing models that learn to recognize ambiguity.

6.2.1 *Disagreement beyond Relations and Frames*

To paraphrase Judea Pearl [103], proving completeness of a theory is notoriously difficult, and should be avoided if one wants to finish a PhD on time. This work does not claim completeness – while we were able to successfully study the impact of disagreement on relation extraction and frame disambiguation, there are *many more tasks and domains in natural language processing* that could be added to this analysis. Already, the CrowdTruth methodology for disagreement-preserving crowdsourcing has been applied to a variety of other tasks outside the scope of this thesis, such as named entity recognition [62], topical relevance of paragraphs [67], and textual description of videos [66]. Additionally, Section 1.2.4 discussed other ambiguity-prone tasks where our methodology for capturing and interpreting disagreement could be explored: anaphora resolution [106], ontology alignment and evaluation [27], part-of-speech tagging [105], and establishing grammatical correctness of text [79].

Another interesting future direction would be to explore the *compositionality of ambiguity*, as it applies to more complex natural language processing tasks, such as text summarization, machine translation, and question answering. We expect that ambiguity at the low-level of the text (e.g. ambiguous relations) will propagate and influence the ambiguity of the entire text. However, the compositional nature of language could potentially complicate the way this propagation occurs – for instance, it is conceivable to have a text where every entity and relation is unambiguous, but when considering the whole text, ambiguity is present. This is frequently the case in legal texts [50], which employ well-defined concepts but are usually open to multiple interpretations. A disagreement-based analysis of such texts could be used to identify the exact step in the language composition where ambiguity appears.

Finally, while the CrowdTruth method of aggregating crowdsourcing results has shown promising results, it should be *compared with more baselines that go beyond majority vote*. In Section 1.2.2, we have presented several alternative crowdsourcing aggregation metrics [16, 68, 75, 125, 128, 129], out of which the most promising appear to be the Bayesian methods [101] that model worker reliability in combination with task difficulty. Future work should explore how the CrowdTruth approach compares to these methods in terms of quality of the ground truth they produce. Furthermore, it would be important to investigate whether Bayesian methods are able to identify ambiguity in the input data and annotations like CrowdTruth is doing.

6.2.2 *The Cost of Disagreement*

While in this thesis we have discussed how crowdsourcing is cheaper than domain experts, an analysis on how to optimize the cost of acquiring crowd annotations is still needed. To collect the different perspectives of the crowd, the CrowdTruth methodology uses a comparatively high number of annotators per task – each task in this thesis used at least 10 workers per unit. Traditional crowdsourcing approaches tend to use less annotators, but this is not usually because of intentionally avoiding multiple perspectives, rather that the cost of employing many annotators is prohibitive.

In Chapter 3, we discussed the value in using a high number of workers per task, and also how the nature of the task (i.e. being more or less open and subjective) influences the optimal number of workers. Building on these results, an important future direction is to build a methodology for finding the optimal crowd payment and number of workers for a task, while also collecting the full spectrum of crowd opinions that can be expressed. As proposed by Lin and Weld [84], a possible solution could be to implement an incremental method to collect annotations – start with a smaller number of annotators for each input unit, then collect more judgments only if the smaller set of workers disagree.

Future work should compare CrowdTruth with other methods that optimize for cost of collection, like the general-purpose one proposed by Mizusawa et al. [94]. More specifically for the task of relation extraction, Liu et al. [85] proposed the Gated Crowd method to identify and train the highest skilled workers such that using only one worker per sentence is enough to bring significant improvement for the training of a relation extraction classifier. A comparison between Gated Crowd and CrowdTruth for relation extraction could be used as a starting point for a combined methodology, one that is able to separate between examples that need relabeling from a single highly skilled worker, and examples which are ambiguous and thus need multiple perspectives.

6.2.3 *Learning Ambiguity*

In Chapter 4, we have shown how the performance of relation extraction models can be improved using disagreement-preserving crowd data, and in Chapter 5, we discussed the link between inter-worker disagreement and ambiguity. The logical next step would be to incorporate ambiguity into the natural language processing models, and learn to predict it. Loss functions in models can be modified to work with continuous scores that express confidence in a label. However, ambiguity is a slightly different feature of the text, one that refers to multiple possible interpretations, and not to poor quality labels. Therefore, it should be possible to use ambiguity and label confidence scores in combination – e.g. by having high confidence that an annotation is ambiguous.

Models that learn to predict ambiguity are difficult to implement, because ambiguity is usually an outlier in the data, and is thus difficult to generalize from. Lebanoff and Liu [80] have done promising work in this direction, by learning to predict vague words and sentences in privacy policies. Future work should explore how to generalize this method to the tasks discussed in this thesis (relation extraction and frame disambiguation), as well as other ambiguity-prone natural language processing tasks.

BIBLIOGRAPHY

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. "TensorFlow: A System for Large-Scale Machine Learning." In: *OSDI*. Vol. 16. 2016, pp. 265–283.

[2] Gabor Angeli, Julie Tibshirani, Jean Wu, and Christopher D Manning. "Combining distant and partial supervision for relation extraction." In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1556–1567.

[3] Alan R Aronson. "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program." In: *Proceedings of the AMIA Symposium*. American Medical Informatics Association. 2001, p. 17.

[4] Lora Aroyo and Chris Welty. "Harnessing disagreement for event semantics." In: *Proceedings of the 2nd International Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2012), 11th International Semantic Web Conference*. 2012, p. 31.

[5] Lora Aroyo and Chris Welty. "Crowd Truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard." In: WebSci '13 (2013).

[6] Lora Aroyo and Chris Welty. "Measuring crowd truth for medical relation extraction." In: *AAAI 2013 Fall Symposium on Semantics for Big Data*. 2013.

[7] Lora Aroyo and Chris Welty. "The Three Sides of CrowdTruth." In: *Journal of Human Computation* 1 (1 2014), pp. 31–34. DOI: 10.15346/hc.v1i1.3.

[8] Lora Aroyo and Chris Welty. "Truth is a lie: Crowd Truth and the seven myths of human annotation." In: *AI Magazine* 36.1 (2015), pp. 15–24.

[9] Lora Aroyo, Anca Dumitrache, Praveen Paritosh, Alex Quinn, and Chris Welty. "Subjectivity, Ambiguity and Disagreement in Crowdsourcing Workshop (SAD2018)." In: *AI Magazine – HCOMP 2018 reports (to appear)* (2018).

[10] Ron Artstein and Massimo Poesio. "Inter-coder agreement for computational linguistics." In: *Computational Linguistics* 34.4 (2008), pp. 555–596.

[11] Collin F Baker. "FrameNet, current collaborations and future goals." In: *Language Resources and Evaluation* 46.2 (2012), pp. 269–286.

[12] Collin F Baker, Charles J Fillmore, and John B Lowe. "The Berkeley FrameNet project." In: *Proceedings of the 17th international conference on Computational linguistics-Volume 1*. Association for Computational Linguistics. 1998, pp. 86–90.

[13] Henri Bal, Dick Epema, Cees de Laat, Rob van Nieuwpoort, John Romein, Frank Seinstra, Cees Snoek, and Harry Wijshoff. "A medium-scale distributed system for computer science research: Infrastructure for the long term." In: *Computer* 49.5 (2016), pp. 54–63.

[14] Petra Saskia Bayerl and Karsten Ingmar Paul. "What Determines Inter-coder Agreement in Manual Annotations? A Meta-analytic Investigation." In: *Comput. Linguist.* 37.4 (Dec. 2011), pp. 699–725. ISSN: 0891-2017. DOI: 10.1162/COLI_a_00074. URL: http://dx.doi.org/10.1162/COLI_a_00074.

[15] Olivier Bodenreider. "The unified medical language system (UMLS): integrating biomedical terminology." In: *Nucleic acids research* 32.suppl 1 (2004), pp. D267–D270.

[16] Alessandro Bozzon, Marco Brambilla, Stefano Ceri, and Andrea Mauri. "Reactive crowdsourcing." In: *Proceedings of the 22nd international conference on World Wide Web*. WWW '13. International World Wide Web Conferences Steering Committee, 2013, pp. 153–164. ISBN: 978-1-4503-2035-1.

[17] Jonathan Bragg, Daniel S Weld, et al. "Crowdsourcing multi-label classification for taxonomy creation." In: *First AAAI conference on human computation and crowdsourcing*. 2013.

[18] John D Burger, Emily Doughty, Sam Bayer, David Tresner-Kirsch, Ben Wellner, John Aberdeen, Kyungjoon Lee, Maricel G Kann, and Lynette Hirschman. "Validating candidate gene-mutation relations in MEDLINE abstracts via crowdsourcing." In: *Data Integration in the Life Sciences*. Springer. 2012, pp. 83–91.

[19] Chris Callison-Burch and Mark Dredze. "Creating speech and language data with Amazon's Mechanical Turk." In: *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics. 2010, pp. 1–12.

[20] Jean Carletta. "Assessing Agreement on Classification Tasks: The Kappa Statistic." In: *Comput. Linguist.* 22.2 (June 1996), pp. 249–254. ISSN: 0891-2017. URL: http://dl.acm.org/citation.cfm?id=230386.230390.

[21] Tommaso Caselli, Rachele Sprugnoli, and Oana Inel. "Temporal Information Annotation: Crowd vs. Experts." In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Ed. by Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. European Language Resources Association (ELRA), 2016. ISBN: 978-2-9517408-9-1.

[22] Silvana Castano, Alfio Ferrara, and Stefano Montanelli. "Human-in-the-Loop Web Resource Classification." In: *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer. 2016, pp. 229–244.

[23] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. "Revolt: Collaborative Crowdsourcing for Labeling Machine Learning Datasets." In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI '17. ACM, 2017. DOI: `10.1145/3025453.3026044`. URL: `http://doi.acm.org/10.1145/3025453.3026044`.

[24] Nancy Chang, Russell Lee-Goldman, and Michael Tseng. "Linguistic Wisdom from the Crowd." In: *Third AAAI Conference on Human Computation and Crowdsourcing*. 2016.

[25] Nancy Chang, Praveen Paritosh, David Huynh, and Collin Baker. "Scaling semantic frame annotation." In: *Proceedings of The 9th Linguistic Annotation Workshop*. 2015, pp. 1–10.

[26] Wendy W Chapman, Prakash M Nadkarni, Lynette Hirschman, Leonard W D'Avolio, Guergana K Savova, and Ozlem Uzuner. "Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions." In: *Journal of the American Medical Informatics Association* 18.5 (2011), pp. 540–543.

[27] Michelle Cheatham and Pascal Hitzler. "Conference v2.0: An Uncertain Version of the OAEI Conference Benchmark." In: *The Semantic Web – ISWC 2014: 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part II*. Ed. by Peter Mika, Tania Tudorache, Abraham Bernstein, Chris Welty, Craig Knoblock, Denny Vrandečić, Paul Groth, Natasha Noy, Krzysztof Janowicz, and Carole Goble. Springer International Publishing, 2014, pp. 33–48. ISBN: 978-3-319-11915-1. DOI: `10.1007/978-3-319-11915-1_3`. URL: `https://doi.org/10.1007/978-3-319-11915-1_3`.

[28] David L Chen and William B Dolan. "Building a persistent workforce on mechanical turk for multilingual data collection." In: *Proceedings of The 3rd Human Computation Workshop (HCOMP 2011)*. 2011.

[29] Jinxiu Chen, Donghong Ji, Chew Lim Tan, and Zhengyu Niu. "Relation Extraction Using Label Propagation Based Semi-supervised Learning." In: *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*. ACL-44. Association for Computational Linguistics, 2006, pp. 129–136. DOI: 10.3115/1220175.1220192. URL: https://doi.org/10.3115/1220175.1220192.

[30] Patricia W. Cheng. "From Covariation to Causation: A Causal Power Theory." In: *Psychological Review* 104.2 (1997), pp. 367–405.

[31] Veronika Cheplygina and Josien PW Pluim. "Crowd disagreement of medical images is informative." In: *arXiv preprint arXiv:1806.08174* (2018).

[32] Lydia B. Chilton, Greg Little, Darren Edge, Daniel S. Weld, and James A. Landay. "Cascade: crowdsourcing taxonomy creation." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '13. ACM, 2013, pp. 1999–2008. ISBN: 978-1-4503-1899-0. DOI: 10.1145/2470654.2466265.

[33] J Cohen. "Kappa: Coefficient of concordance." In: *Educ. Psych. Measurement* 20.37 (1960).

[34] Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. "ZenCrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking." In: *Proceedings of the 21st international conference on World Wide Web*. ACM. 2012, pp. 469–478.

[35] Thomas Demeester, Tim Rocktäschel, and Sebastian Riedel. "Regularizing Relation Representations by First-order Implications." In: *AKBC@ NAACL-HLT*. 2016, pp. 75–80.

[36] Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. "Mechanical Cheat: Spamming Schemes and Adversarial Techniques on Crowdsourcing Platforms." In: *Crowd-Search*. 2012, pp. 26–30.

[37] Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. "Mechanical Cheat: Spamming Schemes and Adversarial Techniques on Crowdsourcing Platforms." In: *Crowd-Search*. 2012, pp. 26–30.

[38] Anca Dumitrache. "Crowdsourcing disagreement for collecting semantic annotation." In: *European Semantic Web Conference*. Springer. 2015, pp. 701–710.

[39] Anca Dumitrache, Lora Aroyo, and Chris Welty. “Achieving Expert-Level Annotation Quality with CrowdTruth: The Case of Medical Relation Extraction.” In: *Proceedings of International Workshop on Biomedical Data Mining, Modeling, and Semantic Integration: A Promising Approach to Solving Unmet Medical Needs (BDM2I 2015)*. (Oct. 11, 2015). Ed. by Dezhao Song, Adam Fermier, Cui Tao, and Frank Schilder. CEUR Workshop Proceedings 1428. 2015. URL: `http://ceur-ws.org/Vol-1428/BDM2I_2015_paper_3.pdf`.

[40] Anca Dumitrache, Lora Aroyo, and Chris Welty. “Achieving Expert-Level Annotation Quality with CrowdTruth: the Case of Medical Relation Extraction.” In: *Proceedings of Biomedical Data Mining, Modeling, and Semantic Integration (BDM2I) Workshop, International Semantic Web Conference (ISWC) 2015*. 2015.

[41] Anca Dumitrache, Lora Aroyo, and Chris Welty. *Medical Relation Extraction Gold Standard with CrowdTruth*. Apr. 2016. DOI: `10.5281/zenodo.50676`. URL: `https://doi.org/10.5281/zenodo.50676`.

[42] Anca Dumitrache, Lora Aroyo, and Chris Welty. “False Positive and Cross-relation Signals in Distant Supervision Data.” In: *Proceedings of the 6th Workshop on Automated Knowledge Base Construction (AKBC)*. 2017.

[43] Anca Dumitrache, Lora Aroyo, and Chris Welty. “Capturing Ambiguity in Crowdsourcing Frame Disambiguation.” In: *Proceedings of the Sixth AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2018, Zürich, Switzerland, July 5-8, 2018.* Ed. by Yiling Chen and Gabriella Kazai. AAAI Press, 2018, pp. 12–20. ISBN: 978-1-57735-799-5. URL: `https://aaai.org/ocs/index.php/HCOMP/HCOMP18/paper/view/17923`.

[44] Anca Dumitrache, Lora Aroyo, and Chris Welty. *CrowdTruth Corpus for Open Domain Relation Extraction from Sentences*. Oct. 2018. DOI: `10.5281/zenodo.1472330`. URL: `https://doi.org/10.5281/zenodo.1472330`.

[45] Anca Dumitrache, Lora Aroyo, and Chris Welty. “Crowdsourcing Ground Truth for Medical Relation Extraction.” In: *ACM Transacations on Interactive Intelligent Systems (TiiS)* 8.2 (July 2018), 11:1–11:20. ISSN: 2160-6455. DOI: `10.1145/3152889`. URL: `http://doi.acm.org/10.1145/3152889`.

[46] Anca Dumitrache, Lora Aroyo, and Chris Welty. “Crowdsourcing Semantic Label Propagation in Relation Classification.” In: *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*. 2018, pp. 16–21.

[47] Anca Dumitrache, Lora Aroyo, and Chris Welty. *FrameNet Semantic Frame Disambiguation with CrowdTruth*. Oct. 2018. DOI: 10.5281/zenodo.1472345. URL: https://doi.org/10.5281/zenodo.1472345.

[48] Anca Dumitrache, Oana Inel, Lora Aroyo, Benjamin Timmermans, and Chris Welty. "CrowdTruth 2.0: Quality Metrics for Crowdsourcing with Disagreement." In: *arXiv preprint arXiv:1808.06080* (2018).

[49] Anca Dumitrache, Oana Inel, Benjamin Timmermans, Carlos Ortiz, Robert-Jan Sips, and Lora Aroyo. "Empirical Methodology for Crowdsourcing Ground Truth." In: *Semantic Web Journal (in publication)* (2018).

[50] Lauren B Edelman. "Legal ambiguity and symbolic structures: Organizational mediation of civil rights law." In: *American journal of Sociology* 97.6 (1992), pp. 1531–1576.

[51] Paul Felt, Kevin Black, Eric K Ringger, Kevin D Seppi, and Robbie Haertel. "Early Gains Matter: A Case for Preferring Generative over Discriminative Crowdsourcing Models." In: *HLT-NAACL*. 2015, pp. 882–891.

[52] Xiaocheng Feng, Jiang Guo, Bing Qin, Ting Liu, and Yongjie Liu. "Effective Deep Memory Networks for Distant Supervised Relation Extraction." In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*. Ed. by Carles Sierra. ijcai.org, 2017, pp. 4002–4008. ISBN: 978-0-9992411-0-3. DOI: 10.24963/ijcai.2017/559. URL: https://doi.org/10.24963/ijcai.2017/559.

[53] Tim Finin, Will Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. "Annotating Named Entities in Twitter Data with Crowdsourcing." In: *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. CSLDAMT '10. Association for Computational Linguistics, 2010, pp. 80–88. URL: http://dl.acm.org/citation.cfm?id=1866696.1866709.

[54] Nicholas FitzGerald, Oscar Täckström, Kuzman Ganchev, and Dipanjan Das. "Semantic role labeling with neural network factors." In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2015, pp. 960–970.

[55] Arthur Flexer and Thomas Grill. "The Problem of Limited Interrater Agreement in Modelling Music Similarity." In: *Journal of New Music Research* 45.3 (2016). PMID: 28190932, pp. 239–251. DOI: 10.1080/09298215.2016.1200631. eprint: https://doi.org/10.1080/09298215.2016.1200631. URL: https://doi.org/10.1080/09298215.2016.1200631.

[56] Frederic Font, Joan Serrà, and Xavier Serra. "Audio clip classification using social tags and the effect of tag expansion." In: *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio*. Audio Engineering Society. 2014.

[57] Marco Fossati, Claudio Giuliano, and Sara Tonelli. "Outsourcing FrameNet to the crowd." In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vol. 2. 2013, pp. 742–747.

[58] Aldo Gangemi, Mehwish Alam, Luigi Asprino, Valentina Presutti, and Diego Reforgiato Recupero. "Framester: a wide coverage linguistic linked data hub." In: *European Knowledge Acquisition Workshop*. Springer. 2016, pp. 239–254.

[59] Karl Moritz Hermann, Dipanjan Das, Jason Weston, and Kuzman Ganchev. "Semantic frame identification with distributed word representations." In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vol. 1. 2014, pp. 1448–1458.

[60] Jisup Hong and Collin F. Baker. "How Good is the Crowd at "Real" WSD?" In: *Proceedings of the 5th Linguistic Annotation Workshop*. LAW V '11. Association for Computational Linguistics, 2011, pp. 30–37. ISBN: 978-1-932432-93-0. URL: http://dl.acm.org/citation.cfm?id=2018966.2018970.

[61] Dirk Hovy, Barbara Plank, and Anders Søgaard. "Experiments with crowdsourced re-annotation of a POS tagging data set." In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 2014, pp. 377–382.

[62] Oana Inel and Lora Aroyo. "Harnessing diversity in crowds and machines for better ner performance." In: *European Semantic Web Conference*. Springer. 2017, pp. 289–304.

[63] Oana Inel, Tommaso Caselli, and Lora Aroyo. "Crowdsourcing Salient Information from News and Tweets." In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Ed. by Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. European Language Resources Association (ELRA), 2016. ISBN: 978-2-9517408-9-1.

[64] Oana Inel, Lora Aroyo, Chris Welty, and Robert-Jan Sips. "Exploiting Crowdsourcing Disagreement with Various Domain-Independent Quality Measures." In: *Proceedings of the 3rd International Workshop on Detection, Representation, and Exploitation*

of Events in the Semantic Web (DeRiVE 2013), 12th International Semantic Web Conference. 2013.

[65] Oana Inel, Khalid Khamkham, Tatiana Cristea, Anca Dumitrache, Arne Rutjes, Jelle van der Ploeg, Lukasz Romaszko, Lora Aroyo, and Robert-Jan Sips. "CrowdTruth: Machine-Human Computation Framework for Harnessing Disagreement in Gathering Annotated Data." In: *The Semantic Web – ISWC 2014: 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part II*. Ed. by Peter Mika, Tania Tudorache, Abraham Bernstein, Chris Welty, Craig Knoblock, Denny Vrandečić, Paul Groth, Natasha Noy, Krzysztof Janowicz, and Carole Goble. Springer International Publishing, 2014, pp. 486–504. ISBN: 978-3-319-11915-1. DOI: 10.1007/978-3-319-11915-1_31. URL: https://doi.org/10.1007/978-3-319-11915-1_31.

[66] Oana Inel, Sabrina Sauer, Lora Aroyo, Alessandro Bozzon, and Matteo Venanzi. "A study of narrative creation by means of crowds and niches." In: *CEUR Workshop Proceedings*. Vol. 2173. CEUR Workshop Proceedings. 2018.

[67] Oana Inel, Giannis Haralabopoulos, Dan Li, Christophe Van Gysel, Zoltán Szlávik, Elena Simperl, Evangelos Kanoulas, and Lora Aroyo. "Studying Topical Relevance with Evidence-based Crowdsourcing." In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM. 2018, pp. 1253–1262.

[68] Panagiotis G. Ipeirotis, Foster Provost, and Jing Wang. "Quality management on Amazon Mechanical Turk." In: *Proceedings of the ACM SIGKDD Workshop on Human Computation*. HCOMP '10. ACM, 2010, pp. 64–67. ISBN: 978-1-4503-0222-7. DOI: 10.1145/1837885.1837906.

[69] Guoliang Ji, Kang Liu, Shizhu He, Jun Zhao, et al. "Distant Supervision for Relation Extraction with Sentence-Level Attention and Entity Descriptions." In: *AAAI*. 2017, pp. 3060–3066.

[70] Xiaotian Jiang, Quan Wang, Peng Li, and Bin Wang. "Relation Extraction with Multi-instance Multi-label Convolutional Neural Networks." In: *COLING*. 2016.

[71] David Jurgens. "Embracing Ambiguity: A Comparison of Annotation Methodologies for Crowdsourcing Word Sense Labels." In: *HLT-NAACL*. 2013, pp. 556–562.

[72] Maurice G Kendall. "A new measure of rank correlation." In: *Biometrika* 30.1/2 (1938), pp. 81–93.

[73] Halil Kilicoglu, Graciela Rosemblat, Marcelo Fiszman, and Thomas C Rindflesch. "Constructing a semantic predication gold standard from the biomedical literature." In: *BMC bioinformatics* 12.1 (2011), p. 486.

[74] Aniket Kittur, Ed H. Chi, and Bongwon Suh. "Crowdsourcing User Studies with Mechanical Turk." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '08. ACM, 2008, pp. 453–456. ISBN: 978-1-60558-011-1. DOI: 10.1145/1357054.1357127. URL: http://doi.acm.org/10.1145/1357054.1357127.

[75] Aniket Kittur, Ed H. Chi, and Bongwon Suh. "Crowdsourcing user studies with Mechanical Turk." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '08. ACM, 2008, pp. 453–456. ISBN: 978-1-60558-011-1. DOI: 10.1145/1357054.1357127.

[76] Krippendorff Klaus. *Content analysis: An introduction to its methodology*. 2013.

[77] James Q Knowlton. "On the definition of "picture"." In: *AV Communication Review* 14.2 (1966), pp. 157–183.

[78] Sarath Kumar Kondreddi, Peter Triantafillou, and Gerhard Weikum. "Combining information extraction and human computing for crowdsourced knowledge acquisition." In: *30th International Conference on Data Engineering*. IEEE. 2014, pp. 988–999.

[79] Jey Han Lau, Alexander Clark, and Shalom Lappin. "Measuring gradience in speakers' grammaticality judgements." In: *Proceedings of the 36th Annual Conference of the Cognitive Science Society*. 2014, pp. 821–826.

[80] Logan Lebanoff and Fei Liu. "Automatic detection of vague words and sentences in privacy policies." In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2018.

[81] Jongwuk Lee, Hyunsouk Cho, Jin-Woo Park, Young-rok Cha, Seung-won Hwang, Zaiqing Nie, and Ji-Rong Wen. "Hybrid entity clustering using crowds and data." English. In: *The VLDB Journal* 22.5 (2013), pp. 711–726. ISSN: 1066-8888. DOI: 10.1007/s00778-013-0328-8.

[82] Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. "Zero-Shot Relation Extraction via Reading Comprehension." In: *CoNLL 2017* (2017), p. 333.

[83] Tong Shu Li, Benjamin M Good, and Andrew I Su. "Exposing ambiguities in a relation-extraction gold standard with crowdsourcing." In: *arXiv preprint arXiv:1505.06256* (2015).

[84] Christopher H Lin, Daniel S Weld, et al. "To Re (label), or Not To Re (label)." In: *Second AAAI Conference on Human Computation and Crowdsourcing*. 2014.

[85] Angli Liu, Stephen Soderland, Jonathan Bragg, Christopher H Lin, Xiao Ling, and Daniel S Weld. "Effective crowd annotation for relation extraction." In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2016, pp. 897–906.

[86] Andrew Mao, Ece Kamar, Yiling Chen, Eric Horvitz, Megan E Schwamb, Chris J Lintott, and Arfon M Smith. "Volunteering versus work for pay: Incentives and tradeoffs in crowdsourcing." In: *First AAAI conference on human computation and crowdsourcing*. 2013.

[87] Diego Marcheggiani and Fabrizio Sebastiani. "On the Effects of Low-Quality Training Data on Information Extraction from Clinical Reports." In: *J. Data and Information Quality* 9.1 (Sept. 2017), 1:1–1:25. ISSN: 1936-1955. DOI: 10.1145/3106235. URL: http://doi.acm.org/10.1145/3106235.

[88] Tyler McDonnell, Matthew Lease, Mucahid Kutlu, and Tamer Elsayed. "Why is that relevant? Collecting annotator rationales for relevance judgments." In: *Fourth AAAI Conference on Human Computation and Crowdsourcing*. 2016.

[89] Quinn McNemar. "Note on the sampling error of the difference between correlated proportions or percentages." In: *Psychometrika* 12.2 (1947), pp. 153–157.

[90] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. "Distributed representations of words and phrases and their compositionality." In: *Advances in neural information processing systems*. 2013, pp. 3111–3119.

[91] George A Miller. "WordNet: a lexical database for English." In: *Communications of the ACM* 38.11 (1995), pp. 39–41.

[92] Lora Aroyo Emiel van Miltenburg Benjamin Timmermans. "The VU Sound Corpus: Adding More Fine-grained Annotations to the Freesound Database." In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Ed. by Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. European Language Resources Association (ELRA), 2016. ISBN: 978-2-9517408-9-1.

[93] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. "Distant supervision for relation extraction without labeled data." In: *Joint Conference of the 47th Annual Meeting of the ACL and*

the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2. Association for Computational Linguistics. 2009, pp. 1003–1011.

[94] Ken Mizusawa, Keishi Tajima, Masaki Matsubara, Toshiyuki Amagasa, and Atsuyuki Morishima. "Efficient Pipeline Processing of Crowdsourcing Workflows." In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. CIKM '18. ACM, 2018, pp. 1559–1562. ISBN: 978-1-4503-6014-2. DOI: 10.1145/3269206.3269292. URL: http://doi.acm.org/10.1145/3269206.3269292.

[95] Jonathan M Mortensen, Mark A Musen, and Natalya F Noy. "Crowdsourcing the verification of relationships in biomedical ontologies." In: *AMIA Annual Symposium Proceedings*. American Medical Informatics Association. 2013, p. 1020.

[96] Thien Huu Nguyen and Ralph Grishman. "Relation extraction: Perspective from convolutional neural networks." In: *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*. 2015, pp. 39–48.

[97] Stefanie Nowak and Stefan Rüger. "How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation." In: *Proceedings of the international conference on Multimedia information retrieval*. ACM. 2010, pp. 557–566.

[98] Natalya F Noy, Jonathan Mortensen, Mark A Musen, and Paul R Alexander. "Mechanical turk as an ontology engineer?: using microtasks as a component of an ontology-engineering workflow." In: *Proceedings of the 5th Annual ACM Web Science Conference*. ACM. 2013, pp. 262–271.

[99] Charles Kay Ogden and I.A. Richards. *The meaning of meaning*. Trubner & Co, 1923.

[100] Jasper Oosterman, Archana Nottamkandath, Chris Dijkshoorn, Alessandro Bozzon, Geert-Jan Houben, and Lora Aroyo. "Crowdsourcing knowledge-intensive tasks in cultural heritage." In: *Proceedings of the 2014 ACM conference on Web science*. ACM. 2014, pp. 267–268.

[101] Silviu Paun, Bob Carpenter, JD Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. "Comparing Bayesian Models of Annotation." In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2018.

[102] Ellie Pavlick, Travis Wolfe, Pushpendre Rastogi, Chris Callison-Burch, Mark Dredze, and Benjamin Van Durme. "FrameNet+: Fast paraphrastic tripling of FrameNet." In: *Proceedings of the*

53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Vol. 2. 2015, pp. 408–413.

[103] Judea Pearl and Dana Mackenzie. *The Book of Why*. Basic Books, 2018.

[104] Maria Pershina, Bonan Min, Wei Xu, and Ralph Grishman. "Infusion of labeled data into distant supervision for relation extraction." In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vol. 2. 2014, pp. 732–738.

[105] Barbara Plank, Dirk Hovy, and Anders Søgaard. "Linguistically debatable or just plain wrong?" In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 2014, pp. 507–511.

[106] Massimo Poesio and Ron Artstein. "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*. Association for Computational Linguistics. 2005, pp. 76–83.

[107] Dražen Prelec, H Sebastian Seung, and John McCoy. "A solution to the single-question crowd wisdom problem." In: *Nature* 541.7638 (2017), pp. 532–535.

[108] James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. "The TimeBank corpus." In: 2003 (2003), p. 40.

[109] Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M Marlin. "Relation extraction with matrix factorization and universal schemas." In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2013, pp. 74–84.

[110] Marta Sabou, Kalina Bontcheva, and Arno Scharl. "Crowdsourcing Research Opportunities: Lessons from Natural Language Processing." In: *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies*. i-KNOW '12. ACM, 2012, 17:1–17:8. ISBN: 978-1-4503-1242-4. DOI: 10.1145/2362456.2362479. URL: http://doi.acm.org/10.1145/2362456.2362479.

[111] Cícero Nogueira dos Santos, Bing Xiang, and Bowen Zhou. "Classifying Relations by Ranking with Convolutional Neural Networks." In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International*

Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers. The Association for Computer Linguistics, 2015, pp. 626–634. ISBN: 978-1-941643-72-3. URL: http://aclweb.org/anthology/P/P15/P15-1061.pdf.

[112] Christina Sarasua, Elena Simperl, Natasha Noy, Abraham Bernstein, and Jan Marco Leimeister. "Crowdsourcing and the Semantic Web: A research manifesto." In: *Human Computation (HCOMP)* 2.1 (2015), pp. 3–17.

[113] Mike Schaekermann, Edith Law, Alex C. Williams, and William Callaghan. "Resolvable vs. Irresolvable Ambiguity: A New Hybrid Framework for Dealing with Uncertain Ground Truth." In: *1st Workshop on Human-Centered Machine Learning at SIGCHI 2016*. 2016.

[114] Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. "Resolvable vs. Irresolvable Disagreement: A Study on Worker Deliberation in Crowd Work." In: *Proceedings of the ACM on Human-Computer Interaction* 2.CSCW (2018), p. 154.

[115] Yaron Singer and Manas Mittal. "Pricing mechanisms for crowdsourcing markets." In: *Proceedings of the 22nd international conference on World Wide Web*. WWW '13. International World Wide Web Conferences Steering Committee, 2013, pp. 1157–1166. ISBN: 978-1-4503-2035-1. URL: http://dl.acm.org/citation.cfm?id=2488388.2488489.

[116] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. "Cheap and Fast—but is It Good?: Evaluating Non-expert Annotations for Natural Language Tasks." In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '08. Association for Computational Linguistics, 2008, pp. 254–263. URL: http://dl.acm.org/citation.cfm?id=1613715.1613751.

[117] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. "Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks." In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '08. Association for Computational Linguistics, 2008, pp. 254–263.

[118] Guillermo Soberón, Lora Aroyo, Chris Welty, Oana Inel, Hui Lin, and Manfred Overmeen. "Measuring Crowd Truth: Disagreement Metrics Combined with Worker Behavior Filters." In: *1st International Workshop on Crowdsourcing the Semantic Web, 12th International Semantic Web Conference*. 2013.

[119] Lucas Sterckx, Thomas Demeester, Johannes Deleu, and Chris Develder. "Knowledge base population using semantic label propagation." In: *Knowledge-Based Systems* 108.C (2016), pp. 79–91.

[120] Md Arafat Sultan, Steven Bethard, and Tamara Sumner. "DLS @ CU: Sentence Similarity from Word Alignment and Semantic Vector Composition." In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. 2015, pp. 148–153.

[121] Swabha Swayamdipta, Sam Thomson, Chris Dyer, and Noah A. Smith. "Frame-Semantic Parsing with Softmax-Margin Segmental RNNs and a Syntactic Scaffold." In: *arXiv preprint arXiv:1706.09528* (2017).

[122] Erik M Van Mulligen, Annie Fourrier-Reglat, David Gurwitz, Mariam Molokhia, Ainhoa Nieto, Gianluca Trifiro, Jan A Kors, and Laura I Furlong. "The EU-ADR corpus: annotated drugs, diseases, targets, and their relationships." In: *Journal of biomedical informatics* 45.5 (2012), pp. 879–884.

[123] Luis Von Ahn. "Human computation." In: *Design Automation Conference, 2009. DAC'09. 46th ACM/IEEE*. IEEE. 2009, pp. 418–419.

[124] Chang Wang and James Fan. "Medical Relation Extraction with Manifold Models." In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*. The Association for Computer Linguistics, 2014, pp. 828–838. ISBN: 978-1-937284-72-5. URL: `http://aclweb.org/anthology/P/P14/P14-1078.pdf`.

[125] Peter Welinder, Steve Branson, Pietro Perona, and Serge J Belongie. "The multidimensional wisdom of crowds." In: *Advances in neural information processing systems*. 2010, pp. 2424–2432.

[126] Chris Welty, James Fan, David Gondek, and Andrew Schlaikjer. "Large Scale Relation Detection." In: *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*. FAM-LbR '10. Association for Computational Linguistics, 2010, pp. 24–33. URL: `http://dl.acm.org/citation.cfm?id=1866775.1866779`.

[127] Chris Welty, Ken Barker, Lora Aroyo, and Shilpa Arora. "Query driven hypothesis generation for answering queries over nlp graphs." In: *The Semantic Web–ISWC 2012*. Springer, 2012, pp. 228–242.

[128] Keenon Werling, Arun Tejasvi Chaganty, Percy S Liang, and Christopher D Manning. "On-the-job learning with bayesian decision theory." In: *Advances in Neural Information Processing Systems*. 2015, pp. 3465–3473.

[129] Jacob Whitehill, Ting fan Wu, Jacob Bergsma, Javier R. Movellan, and Paul L. Ruvolo. "Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise." In: *Advances in Neural Information Processing Systems 22*. Ed. by Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta. Curran Associates, Inc., 2009, pp. 2035–2043. URL: `http://papers.nips.cc/paper/3644-whose-vote-should-count-more-optimal-integration-of-labels-from-labelers-of-unknown-expertise.pdf`.

[130] Zhu Xiaojin and Ghahramani Zoubin. "Learning from labeled and unlabeled data with label propagation." In: *Technical Report CMU-CALD-02–107, Carnegie Mellon University* (2002).

[131] Haijun Zhai, Todd Lingren, Louise Deleger, Qi Li, Megan Kaiser, Laura Stoutenborough, and Imre Solti. "Web 2.0-based crowdsourcing for high-quality gold standard development in clinical natural language processing." In: *Journal of medical Internet research* 15.4 (2013).

[132] Ce Zhang, Feng Niu, Christopher Ré, and Jude Shavlik. "Big data versus the crowd: Looking for relationships in all the right places." In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics. 2012, pp. 825–834.

[133] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. "Attention-based bidirectional long short-term memory networks for relation classification." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vol. 2. 2016, pp. 207–212.

SUMMARY

As knowledge available on the Web expands, natural language processing methods have become invaluable for facilitating data navigation. Tasks such as knowledge base completion and disambiguation are solved with machine learning models for natural language processing that require a lot of data. Human-annotated gold standard, or ground truth, is used for training, testing, and evaluation of these machine learning components.

In recent years, crowdsourcing has become a viable method used to collect ground truth data. But what makes annotations high quality is still a matter of discussion. When collecting multiple annotations for the same task, it is likely that inter-worker disagreement will be present. In typical annotation setups it is assumed that one correct answer exists for every question, and that disagreement must be eliminated from the corpus. This traditional approach to gathering annotation, based on restrictive annotation guidelines, can often results in over-generalized observations, as well as a loss of ambiguity inherent to language, thus becoming unsuitable for use in training natural language processing systems.

The CrowdTruth methodology has been proposed to perform crowdsourcing while preserving inter-annotator disagreement. This thesis explores how the CrowdTruth methodology can be used to collect ground truth data for the training and evaluation of natural language processing models. We present work done across several tasks (relation extraction, semantic frame disambiguation) and domains (medical, open), showing the role of inter-annotator disagreement beyond simply identifying low quality workers.

Chapter 2 argues that disagreement does not need to be eliminated from ground truth data in order to achieve data quality comparable to domain experts. We explore this question for the use case of medical relation extraction from sentences. In the medical domain it is typically assumed that expert annotators are required to get the best quality ground truth. This work shows that, by capturing the inter-annotator disagreement with the CrowdTruth method, medical relation classifiers trained on crowd annotations perform the same as those trained on expert annotations. Furthermore, classifiers trained on crowd annotations perform better than those trained with automatically-labeled data. Using the crowd also reduces the cost (monetary and in time required to find annotators) for collecting the data.

Chapter 3 continues the investigation into the quality of the disagreement - preserving crowd data, by comparing the quality of crowd data aggregated with CrowdTruth metrics and majority vote, a consen-

sus - enforcing metric, over a diverse set of crowdsourcing tasks. We show that by applying the CrowdTruth methodology, we collect richer data that allows us to reason about ambiguity of content. Furthermore, an increased number of crowd workers leads to growth and stabilization in the quality of annotations, going against the usual practice of employing a small number of annotators.

After establishing the quality of the disagreement-preserving crowd data, in Chapter 4 we discuss how CrowdTruth data can be used to improve the performance of a model for relation classification for sentences. We build on work from Chapter 2, where we have shown that training models on on crowd annotations gives better results than training with data automatically-labeled with distant supervision. However, crowd data is expensive to collect. Chapter 4 describes how to correct a large corpus of training data for relation classification by using only a relatively small crowdsourced corpus, with two different methods: (1) by manually propagating the false positive and cross-relation signals identified with the help of the crowd, and (2) by adapting the semantic label propagation method to work with CrowdTruth data.

Finally, in Chapter 5, we explore how inter-annotator disagreement can be used as an indicator for language ambiguity for the task of disambiguating semantic frames (i.e. high-level concepts that represent the meanings of words). Similarly to Chapter 2, we show that the crowd achieves comparative quality with domain experts. A qualitative evaluation of cases when crowd and expert disagree shows that inter-annotator disagreement is an indicator of ambiguity in both frames and sentences. We demonstrate that the cases in which the crowd workers could not agree exhibit ambiguity, either in the sentence, frame, or the task itself, arguing that collapsing such cases to a single, discrete truth value (i.e. correct or incorrect) is inappropriate, creating arbitrary targets for machine learning.

SAMENVATTING

???

SUMAR

dezacorduri în anotarea datelor pentru prelucrarea limbajului natural

SIKS DISSERTATION SERIES

2011

2011-1 Botond Cseke (RUN) *Variational Algorithms for Bayesian Inference in Latent Gaussian Models*

2011-2 Nick Tinnemeier (UU) *Organizing Agent Organizations. Syntax and Operational Semantics of an Organization-Oriented Programming Language*

2011-3 Jan Martijn van der Werf (TUE) *Compositional Design and Verification of Component-Based Information Systems*

2011-4 Hado van Hasselt (UU) *Insights in Reinforcement Learning; Formal analysis and empirical evaluation of temporal-difference learning algorithms*

2011-5 Base van der Raadt (VU) *Enterprise Architecture Coming of Age - Increasing the Performance of an Emerging Discipline.*

2011-6 Yiwen Wang (TUE) *Semantically-Enhanced Recommendations in Cultural Heritage*

2011-7 Yujia Cao (UT) *Multimodal Information Presentation for High Load Human Computer Interaction*

2011-8 Nieske Vergunst (UU) *BDI-based Generation of Robust Task-Oriented Dialogues*

2011-9 Tim de Jong (OU) *Contextualised Mobile Media for Learning*

2011-10 Bart Bogaert (UvT) *Cloud Content Contention*

2011-11 Dhaval Vyas (UT) *Designing for Awareness: An Experience-focused HCI Perspective*

2011-12 Carmen Bratosin (TUE) *Grid Architecture for Distributed Process Mining*

2011-13 Xiaoyu Mao (UvT) *Airport under Control. Multiagent Scheduling for Airport Ground Handling*

2011-14 Milan Lovric (EUR) *Behavioral Finance and Agent-Based Artificial Markets*

2011-15 Marijn Koolen (UvA) *The Meaning of Structure: the Value of Link Evidence for Information Retrieval*

2011-16 Maarten Schadd (UM) *Selective Search in Games of Different Complexity*

2011-17 Jiyin He (UVA) *Exploring Topic Structure: Coherence, Diversity and Relatedness*

2011-18 Mark Ponsen (UM) *Strategic Decision-Making in complex games*

2011-19 Ellen Rusman (OU) *The Mind ' s Eye on Personal Profiles*

2011-20 Qing Gu (VU) *Guiding service-oriented software engineering - A view-based approach*

2011-21 Linda Terlouw (TUD) *Modularization and Specification of Service-Oriented Systems*

2011-22 Junte Zhang (UVA) *System Evaluation of Archival Description and Access*

2011-23 Wouter Weerkamp (UVA) *Finding People and their Utterances in Social Media*

2011-24 Herwin van Welbergen (UT) *Behavior Generation for Interpersonal Coordination with Virtual Humans On Specifying, Scheduling and Realizing Multimodal Virtual Human Behavior*

2011-25 Syed Waqar ul Qounain Jaffry (VU) *Analysis and Validation of Models for Trust Dynamics*

2011-26 Matthijs Aart Pontier (VU) *Virtual Agents for Human Communication - Emotion Regulation and Involvement-Distance Trade-Offs in Embodied Conversational Agents and Robots*

2011-27 Aniel Bhulai (VU) *Dynamic website optimization through autonomous management of design patterns*

2011-28 Rianne Kaptein (UVA) *Effective Focused Retrieval by Exploiting Query Context and Document Structure*

2011-29 Faisal Kamiran (TUE) *Discrimination-aware Classification*

2011-30 Egon van den Broek (UT) *Affective Signal Processing (ASP): Unraveling the mystery of emotions*

2011-31 Ludo Waltman (EUR) *Computational and Game-Theoretic Approaches for Modeling Bounded Rationality*

2011-32 Nees-Jan van Eck (EUR) *Methodological Advances in Bibliometric Mapping of Science*

2011-33 Tom van der Weide (UU) *Arguing to Motivate Decisions*

2011-34 Paolo Turrini (UU) *Strategic Reasoning in Interdependence: Logical and Game-theoretical Investigations*

2011-35 Maaike Harbers (UU) *Explaining Agent Behavior in Virtual Training*

2011-36 Erik van der Spek (UU) *Experiments in serious game design: a cognitive approach*

2011-37 Adriana Burlutiu (RUN) *Machine Learning for Pairwise Data, Applications for Preference Learning and Supervised Network Inference*

2011-38 Nyree Lemmens (UM) *Bee-inspired Distributed Optimization*

2011-39 Joost Westra (UU) *Organizing Adaptation using Agents in Serious Games*

2011-40 Viktor Clerc (VU) *Architectural Knowledge Management in Global Software Development*

2011-41 Luan Ibraimi (UT) *Cryptographically Enforced Distributed Data Access Control*

2011-42 Michal Sindlar (UU) *Explaining Behavior through Mental State Attribution*

2011-43 Henk van der Schuur (UU) *Process Improvement through Software Operation Knowledge*

2011-44 Boris Reuderink (UT) *Robust Brain-Computer Interfaces*

2011-45 Herman Stehouwer (UvT) *Statistical Language Models for Alternative Sequence Selection*

2011-46 Beibei Hu (TUD) *Towards Contextualized Information Delivery: A Rule-based Architecture for the Domain of Mobile Police Work*

2011-47 Azizi Bin Ab Aziz (VU) *Exploring Computational Models for Intelligent Support of Persons with Depression*

2011-48 Mark Ter Maat (UT) *Response Selection and Turn-taking for a Sensitive Artificial Listening Agent*

2011-49 Andreea Niculescu (UT) *Conversational interfaces for task-oriented spoken dialogues: design aspects influencing interaction quality*

2012

2012-1 Terry Kakeeto (UvT) *Relationship Marketing for SMEs in Uganda*

2012-2 Muhammad Umair(VU) *Adaptivity, emotion, and Rationality in Human and Ambient Agent Models*

2012-3 Adam Vanya (VU) *Supporting Architecture Evolution by Mining Software Repositories*

2012-4 Jurriaan Souer (UU) *Development of Content Management System-based Web Applications*

2012-5 Marijn Plomp (UU) *Maturing Interorganisational Information Systems*

2012-6 Wolfgang Reinhardt (OU) *Awareness Support for Knowledge Workers in Research Networks*

2012-7 Rianne van Lambalgen (VU) *When the Going Gets Tough: Exploring Agent-based Models of Human Performance under Demanding Conditions*

2012-8 Gerben de Vries (UVA) *Kernel Methods for Vessel Trajectories*

2012-9 Ricardo Neisse (UT) *Trust and Privacy Management Support for Context-Aware Service Platforms*

2012-10 David Smits (TUE) *Towards a Generic Distributed Adaptive Hypermedia Environment*

2012-11 J.C.B. Rantham Prabhakara (TUE) *Process Mining in the Large: Preprocessing, Discovery, and Diagnostics*

2012-12 Kees van der Sluijs (TUE) *Model Driven Design and Data Integration in Semantic Web Information Systems*

2012-13 Suleman Shahid (UvT) *Fun and Face: Exploring non-verbal expressions of emotion during playful interactions*

2012-14 Evgeny Knutov (TUE) *Generic Adaptation Framework for Unifying Adaptive Web-based Systems*

2012-15 Natalie van der Wal (VU) *Social Agents. Agent-Based Modelling of Integrated Internal and Social Dynamics of Cognitive and Affective Processes.*

2012-16 Fiemke Both (VU) *Helping people by understanding them - Ambient Agents supporting task execution and depression treatment*

2012-17 Amal Elgammal (UvT) *Towards a Comprehensive Framework for Business Process Compliance*

2012-18 Eltjo Poort (VU) *Improving Solution Architecting Practices*

2012-19 Helen Schonenberg (TUE) *What's Next? Operational Support for Business Process Execution*

2012-20 Ali Bahramisharif (RUN) *Covert Visual Spatial Attention, a Robust Paradigm for Brain-Computer Interfacing*

2012-21 Roberto Cornacchia (TUD) *Querying Sparse Matrices for Information Retrieval*

2012-22 Thijs Vis (UvT) *Intelligence, politie en veiligheidsdienst: verenigbare grootheden?*

2012-23 Christian Muehl (UT) *Toward Affective Brain-Computer Interfaces: Exploring the Neurophysiology of Affect during Human Media Interaction*

2012-24 Laurens van der Werff (UT) *Evaluation of Noisy Transcripts for Spoken Document Retrieval*

2012-25 Silja Eckartz (UT) *Managing the Business Case Development in Inter-Organizational IT Projects: A Methodology and its Application*

2012-26 Emile de Maat (UVA) *Making Sense of Legal Text*

2012-27 Hayrettin Gurkok (UT) *Mind the Sheep! User Experience Evaluation & Brain-Computer Interface Games*

2012-28 Nancy Pascall (UvT) *Engendering Technology Empowering Women*

2012-29 Almer Tigelaar (UT) *Peer-to-Peer Information Retrieval*

2012-30 Alina Pommeranz (TUD) *Designing Human-Centered Systems for Reflective Decision Making*

2012-31 Emily Bagarukayo (RUN) *A Learning by Construction Approach for Higher Order Cognitive Skills Improvement, Building Capacity and Infrastructure*

2012-32 Wietske Visser (TUD) *Qualitative multi-criteria preference representation and reasoning*

2012-33 Rory Sie (OUN) *Coalitions in Cooperation Networks (COCOON)*

2012-34 Pavol Jancura (RUN) *Evolutionary analysis in PPI networks and applications*

2012-35 Evert Haasdijk (VU) *Never Too Old To Learn – On-line Evolution of Controllers in Swarm- and Modular Robotics*

2012-36 Denis Ssebugwawo (RUN) *Analysis and Evaluation of Collaborative Modeling Processes*

2012-37 Agnes Nakakawa (RUN) *A Collaboration Process for Enterprise Architecture Creation*

2012-38 Selmar Smit (VU) *Parameter Tuning and Scientific Testing in Evolutionary Algorithms*

2012-39 Hassan Fatemi (UT) *Risk-aware design of value and coordination networks*

2012-40 Agus Gunawan (UvT) *Information Access for SMEs in Indonesia*

2012-41 Sebastian Kelle (OU) *Game Design Patterns for Learning*

2012-42 Dominique Verpoorten (OU) *Reflection Amplifiers in self-regulated Learning*

2012-43 Anna Tordai (VU) *On Combining Alignment Techniques*

2012-44 Benedikt Kratz (UvT) *A Model and Language for Business-aware Transactions*

2012-45 Simon Carter (UVA) *Exploration and Exploitation of Multilingual Data for Statistical Machine Translation*

2012-46 Manos Tsagkias (UVA) *Mining Social Media: Tracking Content and Predicting Behavior*

2012-47 Jorn Bakker (TUE) *Handling Abrupt Changes in Evolving Time-series Data*

2012-48 Michael Kaisers (UM) *Learning against Learning - Evolutionary dynamics of reinforcement learning algorithms in strategic interactions*

2012-49 Steven van Kervel (TUD) *Ontologogy driven Enterprise Information Systems Engineering*

2012-50 Jeroen de Jong (TUD) *Heuristics in Dynamic Sceduling; a practical framework with a case study in elevator dispatching*

2013

2013-1 Viorel Milea (EUR) *News Analytics for Financial Decision Support*

2013-2 Erietta Liarou (CWI) *MonetDB/DataCell: Leveraging the Column-store Database Technology for Efficient and Scalable Stream Processing*

2013-3 Szymon Klarman (VU) *Reasoning with Contexts in Description Logics*

2013-4 Chetan Yadati(TUD) *Coordinating autonomous planning and scheduling*

2013-5 Dulce Pumareja (UT) *Groupware Requirements Evolutions Patterns*

2013-6 Romulo Goncalves(CWI) *The Data Cyclotron: Juggling Data and Queries for a Data Warehouse Audience*

2013-7 Giel van Lankveld (UvT) *Quantifying Individual Player Differences*

2013-8 Robbert-Jan Merk(VU) *Making enemies: cognitive modeling for opponent agents in fighter pilot simulators*

2013-9 Fabio Gori (RUN) *Metagenomic Data Analysis: Computational Methods and Applications*

2013-10 Jeewanie Jayasinghe Arachchige(UvT) *A Unified Modeling Framework for Service Design.*

2013-11 Evangelos Pournaras(TUD) *Multi-level Reconfigurable Self-organization in Overlay Services*

2013-12 Marian Razavian(VU) *Knowledge-driven Migration to Services*

2013-13 Mohammad Safiri(UT) *Service Tailoring: User-centric creation of integrated IT-based homecare services to support independent living of elderly*

2013-14 Jafar Tanha (UVA) *Ensemble Approaches to Semi-Supervised Learning Learning*

2013-15 Daniel Hennes (UM) *Multiagent Learning - Dynamic Games and Applications*

2013-16 Eric Kok (UU) *Exploring the practical benefits of argumentation in multi-agent deliberation*

2013-17 Koen Kok (VU) *The PowerMatcher: Smart Coordination for the Smart Electricity Grid*

2013-18 Jeroen Janssens (UvT) *Outlier Selection and One-Class Classification*

2013-19 Renze Steenhuizen (TUD) *Coordinated Multi-Agent Planning and Scheduling*

2013-20 Katja Hofmann (UvA) *Fast and Reliable Online Learning to Rank for Information Retrieval*

2013-21 Sander Wubben (UvT) *Text-to-text generation by monolingual machine translation*

2013-22 Tom Claassen (RUN) *Causal Discovery and Logic*

2013-23 Patricio de Alencar Silva(UvT) *Value Activity Monitoring*

2013-24 Haitham Bou Ammar (UM) *Automated Transfer in Reinforcement Learning*

2013-25 Agnieszka Anna Latoszek-Berendsen (UM) *Intention-based Decision Support. A new way of representing and implementing clinical guidelines in a Decision Support System*

2013-26 Alireza Zarghami (UT) *Architectural Support for Dynamic Homecare Service Provisioning*

2013-27 Mohammad Huq (UT) *Inference-based Framework Managing Data Provenance*

2013-28 Frans van der Sluis (UT) *When Complexity becomes Interesting: An Inquiry into the Information eXperience*

2013-29 Iwan de Kok (UT) *Listening Heads*

2013-30 Joyce Nakatumba (TUE) *Resource-Aware Business Process Management: Analysis and Support*

2013-31 Dinh Khoa Nguyen (UvT) *Blueprint Model and Language for Engineering Cloud Applications*

2013-32 Kamakshi Rajagopal (OUN) *Networking For Learning; The role of Networking in a Lifelong Learner's Professional Development*

2013-33 Qi Gao (TUD) *User Modeling and Personalization in the Microblogging Sphere*

2013-34 Kien Tjin-Kam-Jet (UT) *Distributed Deep Web Search*

2013-35 Abdallah El Ali (UvA) *Minimal Mobile Human Computer Interaction*

2013-36 Than Lam Hoang (TUE) *Pattern Mining in Data Streams*

2013-37 Dirk Borner (OUN) *Ambient Learning Displays*

2013-38 Eelco den Heijer (VU) *Autonomous Evolutionary Art*

2013-39 Joop de Jong (TUD) *A Method for Enterprise Ontology based Design of Enterprise Information Systems*

2013-40 Pim Nijssen (UM) *Monte-Carlo Tree Search for Multi-Player Games*

2013-41 Jochem Liem (UVA) *Supporting the Conceptual Modelling of Dynamic Systems: A Knowledge Engineering Perspective on Qualitative Reasoning*

2013-42 Leon Planken (TUD) *Algorithms for Simple Temporal Reasoning*

2013-43 Marc Bron (UVA) *Exploration and Contextualization through Interaction and Concepts*

2014

2014-1 Nicola Barile (UU) *Studies in Learning Monotone Models from Data*

2014-2 Fiona Tuliyano (RUN) *Combining System Dynamics with a Domain Modeling Method*

2014-3 Sergio Raul Duarte Torres (UT) *Information Retrieval for Children: Search Behavior and Solutions*

2014-4 Hanna Jochmann-Mannak (UT) *Websites for children: search strategies and interface design - Three studies on children's search performance and evaluation*

2014-5 Jurriaan van Reijsen (UU) *Knowledge Perspectives on Advancing Dynamic Capability*

2014-6 Damian Tamburri (VU) *Supporting Networked Software Development*

2014-7 Arya Adriansyah (TUE) *Aligning Observed and Modeled Behavior*

2014-8 Samur Araujo (TUD) *Data Integration over Distributed and Heterogeneous Data Endpoints*

2014-9 Philip Jackson (UvT) *Toward Human-Level Artificial Intelligence: Representation and Computation of Meaning in Natural Language*

2014-10 Ivan Salvador Razo Zapata (VU) *Service Value Networks*

2014-11 Janneke van der Zwaan (TUD) *An Empathic Virtual Buddy for Social Support*

2014-12 Willem van Willigen (VU) *Look Ma, No Hands: Aspects of Autonomous Vehicle Control*

2014-13 Arlette van Wissen (VU) *Agent-Based Support for Behavior Change: Models and Applications in Health and Safety Domains*

2014-14 Yangyang Shi (TUD) *Language Models With Meta-information*

2014-15 Natalya Mogles (VU) *Agent-Based Analysis and Support of Human Functioning in Complex Socio-Technical Systems: Applications in Safety and Healthcare*

2014-16 Krystyna Milian (VU) *Supporting trial recruitment and design by automatically interpreting eligibility criteria*

2014-17 Kathrin Dentler (VU) *Computing healthcare quality indicators automatically: Secondary Use of Patient Data and Semantic Interoperability*

2014-18 Mattijs Ghijsen (UVA) *Methods and Models for the Design and Study of Dynamic Agent Organizations*

2014-19 Vinicius Ramos (TUE) *Adaptive Hypermedia Courses: Qualitative and Quantitative Evaluation and Tool Support*

2014-20 Mena Habib (UT) *Named Entity Extraction and Disambiguation for Informal Text: The Missing Link*

2014-21 Kassidy Clark (TUD) *Negotiation and Monitoring in Open Environments*

2014-22 Marieke Peeters (UU) *Personalized Educational Games - Developing agent-supported scenario-based training*

2014-23 Eleftherios Sidirourgos (UVA/CWI) *Space Efficient Indexes for the Big Data Era*

2014-24 Davide Ceolin (VU) *Trusting Semi-structured Web Data*

2014-25 Martijn Lappenschaar (RUN) *New network models for the analysis of disease interaction*

2014-26 Tim Baarslag (TUD) *What to Bid and When to Stop*

2014-27 Rui Jorge Almeida (EUR) *Conditional Density Models Integrating Fuzzy and Probabilistic Representations of Uncertainty*

2014-28 Anna Chmielowiec (VU) *Decentralized k-Clique Matching*

2014-29 Jaap Kabbedijk (UU) *Variability in Multi-Tenant Enterprise Software*

2014-30 Peter de Cock (UvT) *Anticipating Criminal Behaviour*

2014-31 Leo van Moergestel (UU) *Agent Technology in Agile Multiparallel Manufacturing and Product Support*

2014-32 Naser Ayat (UvA) *On Entity Resolution in Probabilistic Data*

2014-33 Tesfa Tegegne (RUN) *Service Discovery in eHealth*

2014-34 Christina Manteli (VU) *The Effect of Governance in Global Software Development: Analyzing Transactive Memory Systems.*

2014-35 Joost van Ooijen (UU) *Cognitive Agents in Virtual Worlds: A Middleware Design Approach*

2014-36 Joos Buijs (TUE) *Flexible Evolutionary Algorithms for Mining Structured Process Models*

2014-37 Maral Dadvar (UT) *Experts and Machines United Against Cyberbullying*

2014-38 Danny Plass-Oude Bos (UT) *Making brain-computer interfaces better: improving usability through post-processing.*

2014-39 Jasmina Maric (UvT) *Web Communities, Immigration, and Social Capital*

2014-40 Walter Omona (RUN) *A Framework for Knowledge Management Using ICT in Higher Education*

2014-41 Frederic Hogenboom (EUR) *Automated Detection of Financial Events in News Text*

2014-42 Carsten Eijckhof (CWI/TUD) *Contextual Multidimensional Relevance Models*

2014-43 Kevin Vlaanderen (UU) *Supporting Process Improvement using Method Increments*

2014-44 Paulien Meesters (UvT) *Intelligent Blauw. Intelligence-gestuurde politiezorg in gebiedsgebonden eenheden.*

2014-45 Birgit Schmitz (OUN) *Mobile Games for Learning: A Pattern-Based Approach*

2014-46 Ke Tao (TUD) *Social Web Data Analytics: Relevance, Redundancy, Diversity*

2014-47 Shangsong Liang (UVA) *Fusion and Diversification in Information Retrieval*

2015

2015-1 Niels Netten (UvA) *Machine Learning for Relevance of Information in Crisis Response*

2015-2 Faiza Bukhsh (UvT) *Smart auditing: Innovative Compliance Checking in Customs Controls*

2015-3 Twan van Laarhoven (RUN) *Machine learning for network data*

2015-4 Howard Spoelstra (OUN) *Collaborations in Open Learning Environments*

2015-5 Christoph Bosch (UT) *Cryptographically Enforced Search Pattern Hiding*

2015-6 Farideh Heidari (TUD) *Business Process Quality Computation - Computing Non-Functional Requirements to Improve Business Processes*

2015-7 Maria-Hendrike Peetz (UvA) *Time-Aware Online Reputation Analysis*

2015-8 Jie Jiang (TUD) *Organizational Compliance: An agent-based model for designing and evaluating organizational interactions*

2015-9 Randy Klaassen (UT) *HCI Perspectives on Behavior Change Support Systems*

2015-10 Henry Hermans (OUN) *OpenU: design of an integrated system to support lifelong learning*

2015-11 Yongming Luo (TUE) *Designing algorithms for big graph datasets: A study of computing bisimulation and joins*

2015-12 Julie M. Birkholz (VU) *Modi Operandi of Social Network Dynamics: The Effect of Context on Scientific Collaboration Networks*

2015-13 Giuseppe Procaccianti (VU) *Energy-Efficient Software*

2015-14 Bart van Straalen (UT) *A cognitive approach to modeling bad news conversations*

2015-15 Klaas Andries de Graaf (VU) *Ontology-based Software Architecture Documentation*

2015-16 Changyun Wei (UT) *Cognitive Coordination for Cooperative Multi-Robot Teamwork*

2015-17 Andre van Cleeff (UT) *Physical and Digital Security Mechanisms: Properties, Combinations and Trade-offs*

2015-18 Holger Pirk (CWI) *Waste Not, Want Not! - Managing Relational Data in Asymmetric Memories*

2015-19 Bernardo Tabuenca (OUN) *Ubiquitous Technology for Lifelong Learners*

2015-20 Lois Vanhee (UU) *Using Culture and Values to Support Flexible Coordination*

2015-21 Sibren Fetter (OUN) *Using Peer-Support to Expand and Stabilize Online Learning*

2015-22 Zhemin Zhu (UT) *Co-occurrence Rate Networks*

2015-23 Luit Gazendam (VU) *Cataloguer Support in Cultural Heritage*

2015-24 Richard Berendsen (UVA) *Finding People, Papers, and Posts: Vertical Search Algorithms and Evaluation*

2015-25 Steven Woudenberg (UU) *Bayesian Tools for Early Disease Detection*

2015-26 Alexander Hogenboom (EUR) *Sentiment Analysis of Text Guided by Semantics and Structure*

2015-27 Sandor Heman (CWI) *Updating compressed colomn stores*

2015-28 Janet Bagorogoza (TiU) *Knowledge Management and High Performance; The Uganda Financial Institutions Model for HPO*

2015-29 Hendrik Baier (UM) *Monte-Carlo Tree Search Enhancements for One-Player and Two-Player Domains*

2015-30 Kiavash Bahreini (OU) *Real-time Multimodal Emotion Recognition in E-Learning*

2015-31 Yakup Koe (TUD) *On the robustness of Power Grids*

2015-32 Jerome Gard (UL) *Corporate Venture Management in SMEs*

2015-33 Frederik Schadd (TUD) *Ontology Mapping with Auxiliary Resources*

2015-34 Victor de Graaf (UT) *Gesocial Recommender Systems*

2015-35 Jungxao Xu (TUD) *Affective Body Language of Humanoid Robots: Perception and Effects in Human Robot Interaction*

2016

2016-1 Syed Saiden Abbas (RUN) *Recognition of Shapes by Humans and Machines*

2016-2 Michiel Christiaan Meulendijk (UU) *Optimizing medication reviews through decision support: prescribing a better pill to swallow*

2016-3 Maya Sappelli (RUN) *Knowledge Work in Context: User Centered Knowledge Worker Support*

2016-4 Laurens Rietveld (VU) *Publishing and Consuming Linked Data*

2016-5 Evgeny Sherkhonov (UVA) *Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers*

2016-6 Michel Wilson (TUD) *Robust scheduling in an uncertain environment*

2016-7 Jeroen de Man (VU) *Measuring and modeling negative emotions for virtual training*

2016-8 Matje van de Camp (TiU) *A Link to the Past: Constructing Historical Social Networks from Unstructured Data*

2016-9 Archana Nottamkandath (VU) *Trusting Crowdsourced Information on Cultural Artefacts*

2016-10 George Karafotias (VUA) *Parameter Control for Evolutionary Algorithms*

2016-11 Anne Schuth (UVA) *Search Engines that Learn from Their Users*

2016-12 Max Knobbout (UU) *Logics for Modelling and Verifying Normative Multi-Agent Systems*

2016-13 Nana Baah Gyan (VU) *The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach*

2016-14 Ravi Khadka (UU) *Revisiting Legacy Software System Modernization*

2016-15 Steffen Michels (RUN) *Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments*

2016-16 Guangliang Li (UVA) *Socially Intelligent Autonomous Agents that Learn from Human Reward*

2016-17 Berend Weel (VU) *Towards Embodied Evolution of Robot Organisms*

2016-18 Albert Merono Penuela (VU) *Refining Statistical Data on the Web*
2016-19 Julia Efremova (TUE) *Mining Social Structures from Genealogical Data*
2016-20 Daan Odijk (UVA) *Context & Semantics in News & Web Search*
2016-21 Alejandro Moreno Colleri (UT) *From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground*
2016-22 Grace Lewis (VU) *Software Architecture Strategies for Cyber-Foraging Systems*
2016-23 Fei Cai (UVA) *Query Auto Completion in Information Retrieval*
2016-24 Brend Wanders (UT) *Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach*
2016-25 Julia Kiseleva (TUE) *Using Contextual Information to Understand Searching and Browsing Behavior*
2016-26 Dilhan Thilakarathne (VU) *In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains*
2016-27 Wen Li (TUD) *Understanding Geo-spatial Information on Social Media*
2016-28 Mingxin Zhang (TUD) *Large-scale Agent-based Social Simulation - A study on epidemic prediction and control*
2016-29 Nicolas Honing (TUD) *Peak reduction in decentralised electricity systems -Markets and prices for flexible planning*
2016-30 Ruud Mattheij (UvT) *The Eyes Have It*
2016-31 Mohammad Khelghati (UT) *Deep web content monitoring*
2016-32 Eelco Vriezekolk (UT) *Assessing Telecommunication Service Availability Risks for Crisis Organisations*
2016-33 Peter Bloem (UVA) *Single Sample Statistics, exercises in learning from just one example*
2016-34 Dennis Schunselaar (TUE) *Configurable Process Trees: Elicitation, Analysis, and Enactment*
2016-35 Zhaochun Ren (UVA) *Monitoring Social Media: Summarization, Classification and Recommendation*
2016-36 Daphne Karreman (UT) *Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies*
2016-37 Giovanni Sileno (UvA) *Aligning Law and Action - a conceptual and computational inquiry*
2016-38 Andrea Minuto (UT) *Materials that Matter - Smart Materials meet Art & Interaction Design*
2016-39 Merijn Bruijnes (UT) *Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect*
2016-40 Christian Detweiler (TUD) *Accounting for Values in Design*
2016-41 Thomas King (TUD) *Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance*
2016-42 Spyros Martzoukos (UVA) *Combinatorial and Compositional Aspects of Bilingual Aligned Corpora*
2016-43 Saskia Koldijk (RUN) *Context-Aware Support for Stress Self-Management: From Theory to Practice*
2016-44 Thibault Sellam (UVA) *Automatic Assistants for Database Exploration*
2016-45 Bram van de Laar (UT) *Experiencing Brain-Computer Interface Control*
2016-46 Jorge Gallego Perez (UT) *Robots to Make you Happy*
2016-47 Christina Weber (UL) *Real-time foresight - Preparedness for dynamic innovation networks*
2016-48 Tanja Buttler (TUD) *Collecting Lessons Learned*
2016-49 Gleb Polevoy (TUD) *Participation and Interaction in Projects. A Game-Theoretic Analysis*
2016-50 Yan Wang (UVT) *The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains*

2017

2017-1 Jan-Jaap Oerlemans (UL) *Investigating Cybercrime*
2017-2 Sjoerd Timmer (UU) *Designing and Understanding Forensic Bayesian Networks using Argumentation*
2017-3 Daniel Harold Telgen (UU) *Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines*
2017-4 Mrunal Gawade (CWI) *Multi-Core Parallelism in a Column-Store*
2017-5 Mahdieh Shadi (UVA) *Collaboration Behavior*
2017-6 Damir Vandic (EUR) *Intelligent Information Systems for Web Product Search*
2017-7 Roel Bertens (UU) *Insight in Information: from Abstract to Anomaly*
2017-8 Rob Konijn (VU) *Detecting Interesting Differences: Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery*
2017-9 Dong Nguyen (UT) *Text as Social and Cultural Data: A Computational Perspective on Variation in Text*
2017-10 Robby van Delden (UT) *(Steering) Interactive Play Behavior*
2017-11 Florian Kunneman (RUN) *Modelling patterns of time and emotion in Twitter #anticipointment*
2017-12 Sander Leemans (TUE) *Robust Process Mining with Guarantees*
2017-13 Gijs Huisman (UT) *Social Touch Technology - Extending the reach of social touch through haptic technology*
2017-14 Shoshannah Tekofsky (UvT) *You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior*
2017-15 Peter Berck, Radboud University (RUN) *Memory-Based Text Correction*
2017-16 Aleksandr Chuklin (UVA) *Understanding and Modeling Users of Modern Search Engines*

2017-17 Daniel Dimov (UL) *Crowdsourced Online Dispute Resolution*

2017-18 Ridho Reinanda (UVA) *Entity Associations for Search*

2017-19 Jeroen Vuurens (TUD) *Proximity of Terms, Texts and Semantic Vectors in Information Retrieval*

2017-20 Mohammadbashir Sedighi (TUD) *Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility*

2017-21 Jeroen Linssen (UT) *Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)*

2017-22 Sara Magliacane (VU) *Logics for causal inference under uncertainty*

2017-23 David Graus (UVA) *Entities of Interest - Discovery in Digital Traces*

2017-24 Chang Wang (TUD) *Use of Affordances for Efficient Robot Learning*

2017-25 Veruska Zamborlini (VU) *Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search*

2017-26 Merel Jung (UT) *Socially intelligent robots that understand and respond to human touch*

2017-27 Michiel Joosse (UT) *Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors*

2017-28 John Klein (VU) *Architecture Practices for Complex Contexts*

2017-29 Adel Alhuraibi (UVT) *From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT*

2017-30 Wilma Latuny (UVT) *The Power of Facial Expressions*

2017-31 Ben Ruijl (UL) *Advances in computational methods for QFT calculations*

2017-32 Thaer Samar (RUN) *Access to and Retrievability of Content in Web Archives*

2017-33 Brigit van Loggem (OU) *Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity*

2017-34 Maren Scheffel (OUN) *The Evaluation Framework for Learning Analytics*

2017-35 Martine de Vos (VU) *Interpreting natural science spreadsheets*

2017-36 Yuanhao Guo (UL) *Shape Analysis for Phenotype Characterisation from High-throughput Imaging*

2017-37 Alejandro Montes Garcia (TUE) *WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy*

2017-38 Alex Kayal (TUD) *Normative Social Applications*

2017-39 Sara Ahmadi (RUN) *Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR*

2017-40 Altaf Hussain Abro (VUA) *Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems"*

2017-41 Adnan Manzoor (VUA) *Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle*

2017-42 Elena Sokolova (RUN) *Causal discovery from mixed and missing data with applications on ADHD datasets*

2017-43 Maaike de Boer (RUN) *Maaike de Boer (RUN)*

2017-44 Garm Lucassen (UU) *Understanding User Stories - Computational Linguistics in Agile Requirements Engineering*

2017-45 Bas Testerink (UU) *Decentralized Runtime Norm Enforcement*

2017-46 Jan Schneider (OU) *Sensor-based Learning Support*

2017-47 Yie Yang (TUD) *Crowd Knowledge Creation Acceleration*

2017-48 Angel Suarez (OU) *Collaborative inquiry-based learning*

2018

2018-1 Han van der Aa (VUA) *Comparing and Aligning Process Representations*

2018-2 Felix Mannhardt (TUE) *Multi-perspective Process Mining*

2018-3 Steven Bosems (UT) *Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction*

2018-4 Jordan Janeiro (TUD) *Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks*

2018-5 Hugo Huurdeman (UVA) *Supporting the Complex Dynamics of the Information Seeking Process*

2018-6 Dan Ionita (UT) *Model-Driven Information Security Risk Assessment of Socio-Technical Systems*

2018-7 Jieting Luo (UU) *A formal account of opportunism in multi-agent systems*

2018-8 Rick Smetsers (RUN) *Advances in Model Learning for Software Systems*

2018-9 Xu Xie (TUD) *Data Assimilation in Discrete Event Simulations*

2018-10 Julienka Mollee (VUA) *Moving forward: supporting physical activity behavior change through intelligent technology*

2018-11 Mahdi Sargolzaei (UVA) *Enabling Framework for Service-oriented Collaborative Networks*

2018-12 Xixi Lu (TUE) *Using behavioral context in process mining*

2018-13 Seyed Amin Tabatabaei (VUA) *Using behavioral context in process mining: Exploring the added value of computational models for increasing the use of renewable energy in the residential sector*

2018-14 Bart Joosten (UVT) *Detecting Social Signals with Spatiotemporal Gabor Filters*

2018-15 Naser Davarzani (UM) *Biomarker discovery in heart failure*

2018-16 Jaebok Kim (UT) *Automatic recognition of engagement and emotion in a group of children*

2018-17 Jianpeng Zhang (TUE) *On Graph Sample Clustering*

2018-18 Henriette Nakad (UL) *De Notaris en Private Rechtspraak*

2018-19 Minh Duc Pham (VUA) *Emergent relational schemas for RDF*

2018-20 Manxia Liu (RUN) *Time and Bayesian Networks*

2018-21 Aad Slootmaker (OUN) *EMERGO: a generic platform for authoring and playing scenario-based serious games*

2018-22 Eric Fernandes de Mello Araújo (VUA) *Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks*

2018-23 Kim Schouten (EUR) *Semantics-driven Aspect-Based Sentiment Analysis*

2018-24 Jered Vroon (UT) *Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots*

2018-25 Riste Gligorov (VUA) *Serious Games in Audio-Visual Collections*

2018-26 Roelof de Vries (UT) *Theory-Based And Tailor-Made: Motivational Messages for Behavior Change Technology*

2018-27 Maikel Leemans (TUE) *Hierarchical Process Mining for Scalable Software Analysis*

2018-28 Christian Willemse (UT) *Social Touch Technologies: How they feel and how they make you feel*

2019

2019-1 Rob van Eijk (UL) *Comparing and Aligning Process Representations*

2019-2 Emmanuelle Beauxis- Aussalet (CWI, UU) *Statistics and Visualizations for Assessing Class Size Uncertainty*

2019-3 Eduardo Gonzalez Lopez de Murillas (TUE) *Process Mining on Databases: Extracting Event Data from Real Life Data Sources*