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 how to measure the quality of crowd annotations 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 ground truth corpus. This traditional approach to gathering annotation, based on restrictive annotation guidelines, often results in the unjust penalization of qualified workers with a different perspective from the general consensus, as well as over-generalized observations, and a loss of ambiguity inherent to language, which could render the annotated data unsuitable for use in training natural language processing systems.

The CrowdTruth methodology has been proposed to perform crowdsourcing 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 annotated data. The methodology represents the crowdsourcing system as a triangle 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 for the three crowdsourcing components, by modeling the way that the components interact with each other – e.g. in ambiguous sentences, we expect to have more disagreement between workers, therefore workers on those sentences should not be considered less trustworthy.

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). These experiments show the role of inter-annotator disagreement in establishing data quality, beyond simply for 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 that obtained from domain experts. We explore this argument 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 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.

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, therefore corpora collected in this way tend to be small in size. Chapter 4 describes how such a relatively small crowdsourced corpus can be used to correct a large corpus of training data for relation classification, 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 workers disagreed between themselves or with the expert annotators show that inter-annotator disagreement is an indicator of ambiguity in both frames and sentences. 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.