Active Learning for Assisted Corpus Construction: A Case Study in Knowledge Discovery from Biomedical Text

Alejandro Piad-Morffis[1], **Yoan Gutiérrez**[2,3], **Yudivián Almeida-Cruz**[1], and **Rafael Muñoz**[2,3]

[1]School of Math and Computer Science, University of Havana
{apiad,yudy}@matcom.uh.cu
[2]Deparment of Software and Computing Systems, University of Alicante
[3]University Institute for Computing Research, University of Alicante
{ygutierrez,rafael}@dlsi.ua.es

Abstract

This paper presents an active learning approach that aims to reduce the human effort required during the annotation of natural language corpora composed of entities and semantic relations. Our approach assists human annotators by intelligently selecting the most informative sentences to annotate and then pre-annotating them with a few highly accurate entities and semantic relations. We define an uncertainty-based query strategy with a weighted density factor, using similarity metrics based on sentence embeddings. As a case study, we evaluate our approach via simulation in a biomedical corpus and estimate the potential reduction in total annotation time. Experimental results suggest that the query strategy reduces by between 35% and 40% the number of sentences that must be manually annotated to develop systems able to reach a target F_1 score, while the pre-annotation strategy produces an additional 24% reduction in the total annotation time. Overall, our preliminary experiments suggest that as much as 60% of the annotation time could be saved while producing corpora that have the same usefulness for training machine learning algorithms. An open-source computational tool that implements the aforementioned strategies is presented and published online for the research community.

1 Introduction

Machine learning, and specifically supervised learning, is one of the most effective tools for automating complex cognitive tasks, such as recognizing objects in images or understanding natural language text. One of the main bottlenecks of supervised learning is the need for high-quality datasets of labeled samples on which statistical models can be trained. These datasets are usually built by human experts in a lengthy and costly manual process. Active learning (Cohn, 2010) is an alternative paradigm to conventional supervised learning that has been proposed to reduce the costs involved in manual annotation .

The key idea underlying active learning is that a learning algorithm can perform better with less training examples if it is allowed to actively select which examples to learn from (Settles, 2009). In the supervised learning context, this paradigm changes the role of the human expert. In conventional supervised learning contexts, the human expert guides the learning process by providing a large dataset of labeled examples. However, in active learning the active role is shifted to the algorithm and the human expert becomes an oracle, participating in a labeling-training-query loop. In the active paradigm, a model is incrementally built by training on a partial collection of samples and then selecting one or more unlabeled samples to query the human oracle for labels and increase the training set. This approach introduces the new problem of how to best select the query samples so as to maximize the model's performance while minimizing the effort of the human participant.

The simplest active learning scenario consists of the classification of independent elements x_i drawn from a pool of unlabeled samples. Examples range from image classification (Gal et al., 2017) to sentiment mining (Kranjc et al., 2015), in which the minimal level of sampling (e.g., an image or text document) corresponds to the minimal level of decision. i.e, a single label is assigned to each x_i. More complex scenarios arise when the decision level is more fine-grained than the sampling level. In the domain of text mining, an interesting scenario is the task of entity and relation extraction from natural language text (Zhang et al., 2012). In this scenario the sampling level is a sentence, but the minimal level of decision involves each token or pair of tokens in the sentence, and furthermore, these decisions are in general not

independent within the same sentence. In this case, it is not trivial to estimate how informative an unlabeled sample will be, since each sample has several sources of uncertainty.

This research focuses on the tasks of entity and relation extraction, proposing an active learning strategy to reduce the overall time of annotation for human experts, by actively selecting the most informative sentences to annotate. We also consider the problem of providing some instances of entities and relations pre-annotated to further reduce the annotation time for the human, while minimizing the number of erroneous suggestions. In contrast with the usual formulation, we focus on the problem of obtaining the corpus per-se, and the task of training the underlying machine learning models is considered as means to an end rather than as the primary objective.

The contributions of this research can be summarized as follows:

- We present an active learning strategy for the problem of entity and relation extraction from natural language text that greatly reduces the annotation time for human experts by actively selecting the most informative sentences and providing pre-annotated suggestions when possible.
- We propose an informativeness measure for entity and relation extraction that factors in the uncertainty of annotations in a sentence counter-balanced by its similarity to the labeled set.
- We evaluate our proposal in an experimental corpus in the biomedical domain, and study the impact of the query strategy and the benefit of providing pre-annotated suggestions, in terms of reducing the overall time of annotation.
- As a tangible result, we provide the source code for a prototype annotation tool that implements the aforementioned strategies and is compatible with the BRAT annotation tool[1].

The remaining sections of this paper are organized as follows. Section 2 reviews the most relevant research related with active learning in general and specifically for entity and relation extraction. Section 3 presents the formal definition for our active learning approach. Section 4 describes a computational prototype tool that implements this strategy. Section 5 evaluates our proposal in the context of a corpus of entities and relations in the biomedical domain. Section 6 presents a discussion of the most relevant insights that our research provides. Finally, Section 7 presents the main conclusions of our research.

[1]For the purpose of preserving anonymity the tool is provided as supplementary source code at this point, but will be eventually release with an Open Source license.

2 Related Works

This section reviews some of the most relevant research related with active learning in general, and specifically focused on entity detection and relation extraction. One of the most important design decisions in active learning is how to intelligently select the novel unlabeled samples in the most efficient way. The underlying assumption is that we want to train a model to the highest possible performance (measured in precision, F_1, etc.) while minimizing the human cost (measured in time, number of samples manually labeled, or any other suitable metric). This requirement is often framed as the selection of the *most informative* unlabeled samples, and formalized in terms of a query strategy (Settles, 2009). The most common query strategies for general-purpose active learning can be grouped into the following categories:

(i) Uncertainty sampling: The most informative samples are considered those with the highest degree of uncertainty, given some measure of uncertainty for each sample (Lewis and Catlett, 1994).

(ii) Query by committee: The most informative samples are considered those with the highest disagreement among a committee of either different models or different hypotheses from the same underlying model (Seung et al., 1992).

(iii) Expected model change: The most informative samples are considered those that produce the highest change in the model's hypothesis if they were included in the training set (Settles et al., 2008).

(iv) Variance and error reduction: The most informative samples are those which produce the highest reduction in the model's generalization error or, as a proxy, its variance (Roy and McCallum, 2001).

Expected model change (iii) and variance/error reduction (iv) strategies are heavily dependent on the specific learning model used. In contrast, uncertainty sampling (i) and query by committee (ii) are applicable in general with a high degree of model agnosticism. Furthermore, relevant subsets of both strategies can be formalized under a single framework if we define the uncertainty as a measure of the entropy of the model's predicted output. In this framework, query-by-committee can be implemented via weighted voting, thereby assigning empirical probabilities to the possible outputs.

Weighted density is a complimentary strategy in which the most informative samples are weighted by how representative they are of the input space, for example, by measuring their similarity to the remaining samples (Settles and Craven, 2008). This approach attempts to counter-balance a noticeable tendency to select outliers as the most informative samples —a problem associated with other query strategies— since outliers are often the samples that create the highest amount of uncertainty, disagreement or hypothesis change.

Recent advances in natural language processing have produced an increased interest in active learning to alleviate the requirement for large annotated corpora (Olsson, 2009; Tchoua et al., 2019). Settles and Craven (2008) compare several strategies for active learning in sequence labeling scenarios, concluding that query strategies based on measures of sequence entropy combined with weighted sampling outperform other variants. Meduri et al. (2020) propose a comprehensive benchmark to evaluate different active learning strategies for entity matching. In the task of named entity recognition, CRF models have been used to select query samples (Claveau and Kijak, 2017; Lin et al., 2019). The task of relation extraction also benefits from active learning approaches, both in general-purpose settings (Fu and Grishman, 2013) and in domain-specific settings (Zhang et al., 2012). However, despite the growing body of research, it is still a challenge to apply active learning in joint entity recognition and relation extraction, especially in scenarios with low resources (Gao et al., 2019).

3 Active Learning Strategy for Entity-Relation Annotation

This section presents our active learning strategy for human-in-the-loop annotation of corpora based on entity recognition and relation extraction. A high-level overview of the process is illustrated in Figure 1.

Our active learning strategy is designed for an arbitrary corpora of independent natural language sentences, each of which must be annotated by a human expert at token level. We consider a predefined set $\mathcal{E}$ of entity labels, each of which can span one or more tokens, continuous or discontinuous. Additionally, there is a predefined set $\mathcal{R}$ of binary relation types between entities, where the possible relations between each pair of entities can depend on the entity type, i.e., not all relation types are defined for all entity labels. There is no sub-token annotation, and not all tokens need to be annotated. This abstract annotation schema can represent a broad range of different tasks, from domain-specific relation extraction (e.g., gene-protein interaction) to general-purpose semantic representation (e.g., AMR parsing).

The active learning strategy proposed in this research works iteratively in batches of K sentences (e.g., $K = 10$). At any point there will be a labeled pool $\mathbf{L}$ with $|\mathbf{L}| = n \times K$ sentences that have been manually annotated by a human annotator, and a large unlabeled pool $\mathbf{U}$ of raw sentences. Initially, the human annotator selects K representative sentences and performs a full manual annotation (step 0). Afterwards, two machine learning models are iteratively trained on the manually labeled sentences (step 1) and a metric of informativeness, $I(s)$, is computed for each sentence $s \in \mathbf{U}$ (step2). The top K sentences in terms of $I(\cdot)$ are selected (step 3) and the model produces a prediction of entity and relation labels for each one (step 4). Each prediction has an associated metric of uncertainty, $H(\cdot)$, estimated by the models. Based on this uncertainty and pre-defined thresholds u_e and u_r for entities and relations respectively, all the entities e_i (relations r_j) with an estimated uncertainty $H(e_i) > u_e$ ($H(r_j) > u_r$) are discarded. Finally the selected and partially annotated sentences are presented to the human annotator, who must correct the incorrect annotations and add the missing ones (step 5). The corrected sentences are incorporated to the labeled pool for the next iteration (step 6).

The following components for the active learning strategy need to be specified: a machine learning model M_E that predicts entity labels; a machine learning model M_R that predicts relations; and, suitable definitions for $I(\cdot)$ and $H(\cdot)$ based

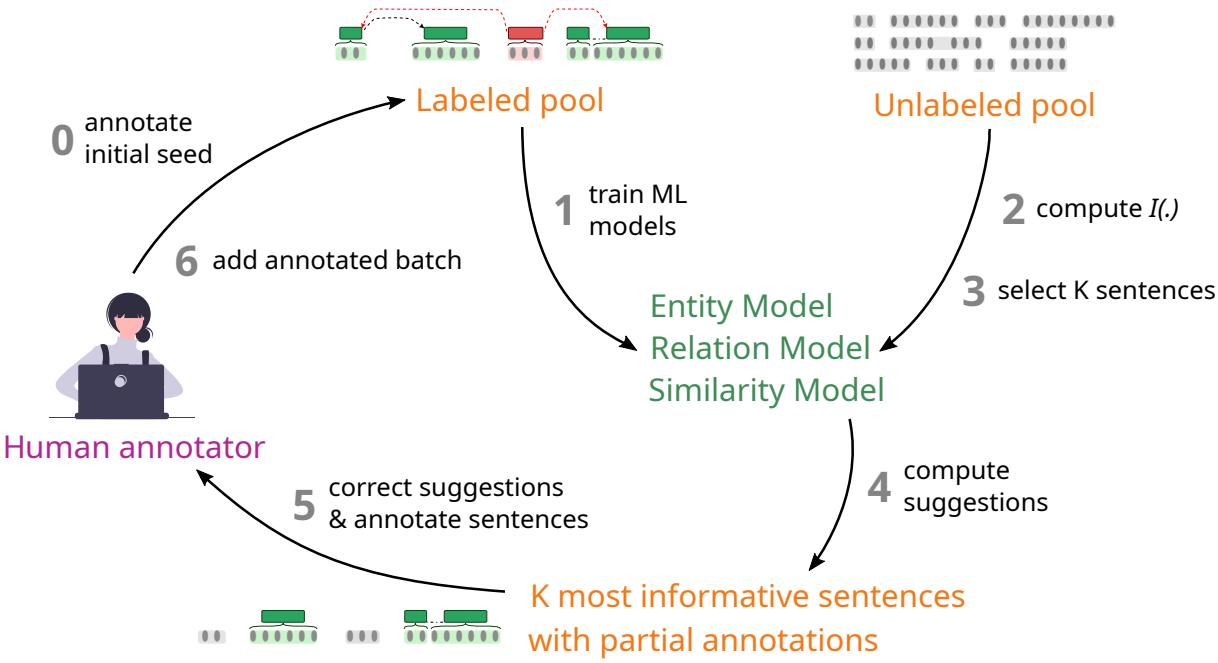

Figure 1: High-level illustration of the active learning strategy presented in this research for human-in-the-loop annotation of entity-relation corpora.

on these machine learning models. We will not define specific machine learning models at this point, since different models can be suitable for different corpora. For our strategy to work, the machine learning models M_E and M_R are only required to provide a probability distribution over the possible labels rather than a single prediction. This is a soft restriction that many machine learning models comply with.

To measure the informativeness $I(s_i)$ of each sentence $s_i \in \mathbf{U}$, we define a metric based on uncertainty sampling with weighted density, inspired by Settles and Craven (2008).

First, given the set of n entity annotations $E_i \in \mathcal{E}^n = \{e_1^i, \ldots, e_n^i\}$ and m relation annotations $R_i \in \mathcal{R}^m = \{r_1^i, \ldots, r_m^i\}$ predicted for a sentence s_i, we define the uncertainty of each entity e_k^i (or relation r_k^i) as the entropy of the probability distribution for all possible labels for that entity or relation. Formally:

$$H(e_k^i) = -\sum_{l_j \in \mathcal{E}} P(e_k^i = l_j | s_i; \theta) \log_2 P(e_k^i = l_j | s_i; \theta)$$

$$H(r_k^i) = -\sum_{l_j \in \mathcal{R}} P(r_k^i = l_j | s_i; \theta) \log_2 P(r_k^i = l_j | s_i; \theta)$$

Where θ represents the parameters of the machine learning model used to estimate these probabilities.

We can define the mean uncertainty associated to the predicted entities and relations, respectively, as follows:

$$\hat{H}(E_i) = \frac{1}{n} \sum_{e_k^i \in E_i} H(e_k^i) \qquad \hat{H}(R_i) = \frac{1}{m} \sum_{r_k^i \in R_i} H(r_k^i)$$

Second, we define an information density metric $ID(s_i)$ to estimate how representative each sentence s_i is with respect to the input space. In a similar formulation to Settles and Craven (2008), $ID(s_i)$ is defined as the average similarity of the sentence s_i to the cluster of K labeled sentences:

$$ID(s_i) = \frac{1}{K} \sum_{s_j \in \mathbf{L}_i^*} sim(s_i, s_j)$$

Where $\mathbf{L}_i^*$ is the subset of K labeled sentences that maximize the similarity metric with respect to s_i. Any sensible similarity metric can be used. In this research we propose the use of *Doc2Vec* embeddings (Le and Mikolov, 2014) pre-trained on the unlabeled set $\mathbf{U}$ to estimate sentence similarity.

Finally, the overall informativeness of an unlabeled sentence s_i is estimated based on the uncertainty measures $H(\cdot)$ of each component, weighted by the information density of the sentence:

$$I(s_i) = \left[\hat{H}(E_i) + \hat{H}(R_i)\right] \times ID(s_i)^\beta$$

Where β is a scaling factor to balance exploitation versus exploration, i.e., decreasing the uncertainty of the model versus selecting more varied sentences to reduce model bias.

If we consider the annotation of a sentence as a stochastic process, where each entity or relation annotation is a random event, then $\hat{H}(\cdot)$ is a finite approximation of the entropy rate of the annotation process. This provides an intuitive interpretation for the informativeness measure $I(\cdot)$ in the domain of information theory. The most informative sentences are those whose entropy rate is maximum (weighted by density). Maximum entropy rate has been successfully applied to feature selection in other machine learning scenarios (Einicke et al., 2017).

4 Computational Prototype

The strategy presented in section 3 is implemented as a web application that can be integrated with the BRAT annotation tool (Stenetorp et al., 2012). This application is independent of BRAT and relies only on the file system to iteratively select batches of sentences and apply suggestions. The web interface is simple to use, allowing the user to ask for a new batch, and decide whether to accept, modify or discard the annotation suggestions (see Figure 2). This tool is compatible with any entity and relation annotation schema that can be represented in BRAT Standof ANN format (Yepes et al., 2013).

As explained in Section 3, two different machine learning models M_E and M_R must be implemented to evaluate the informativeness metric $I(\cdot)$. These models must provide probability estimates for each label, and should be efficient enough to be trained in the same time it takes a human annotator to annotate a single batch, such that the new batch is always ready. For the previous reasons, we selected two simple machine learning models based on standard formulations for the problems of entity recognition and relation extraction respectively.

For entity model M_E, we select a conditional random field (CRF) classification model with syntactic and semantic features extracted with the `spacy` library. The extracted features include coarse and fine-grained part-of-speech tagging, lemmatization, a standard NER labeling, as well as indicator variables for several syntactic patterns (e.g., numbers, dates, punctuation, emails, URLs, etc.). By this means, the entity recognition problem is framed as a sequence tagging problem using the BILOUV encoding and Viterbi decoding. Special hand-crafted rules are designed to account for multi-word entities with discontinuous word spans. The uncertainty of each entity $H(e_k^i)$ is estimated by the normalized marginal probabilities of the CRF model on the token sequence, averaging the probabilities of the tokens that correspond to the same entity. Despite its simplicity, this model achieves an F_1 score of 0.78 in the entity extraction subtask of the eHealth-KD Challenge 2020, which is competitive with state-of-the-art techniques in past benchmarks (Piad-Morffis et al., 2019b).

In the case of the relation model M_R, this subtask is more complex and simple baselines perform significantly worse than state-of-the-art models. However, since complex models cannot be trained in the required time, we decided to maintain a simple baseline. The problem of extracting all relations in a sentence is modeled as a set of independent classification problems between all pairs of entities in the sentence. Each pair is represented by the same characteristics used in the entity recognition subtask, applied to both entities under analysis, plus a bag-of-words encoding of the tokens that appear in the smallest dependency subtree that contains both entities. The uncertainty $H(r_j^i)$ of each pair is computed from the marginal probabilities provided by a logistic regression model trained on each pair representation. For the information density metric $ID(\cdot)$ an implementation of *Doc2Vec* from the `gensim` library is used.

5 Experiments and Results

To validate the effectiveness of the active learning strategy proposed in this research we selected a recent manually annotated corpus of Spanish sentences in the biomedical domain, i.e., the eHealth-KD 2020 corpus (Piad-Morffis et al., 2020). This selection was motivated by the relative complexity of the annotation schema proposed in this corpus, which contains different entity and relation types, multi-word tokens, overlapping annotations and other characteristics that make it a challenging annotation process even for human experts (Piad-Morffis et al., 2019a). The corpus contains a total of 1300 sentences in Spanish, manually annotated with a general-purpose entity-relation schema. Of these, a set of 1000 sentences is used as the unlabeled pool $\mathbf{U}$, and the remaining 300 are used for testing the final performance of all machine learning models trained in the experimentation. The corpus has been split following the authors' recommendations. Figure 2 shows an illustrative example of the annotation schema applied to 3 exemplary English sentences, in the context of the prototype

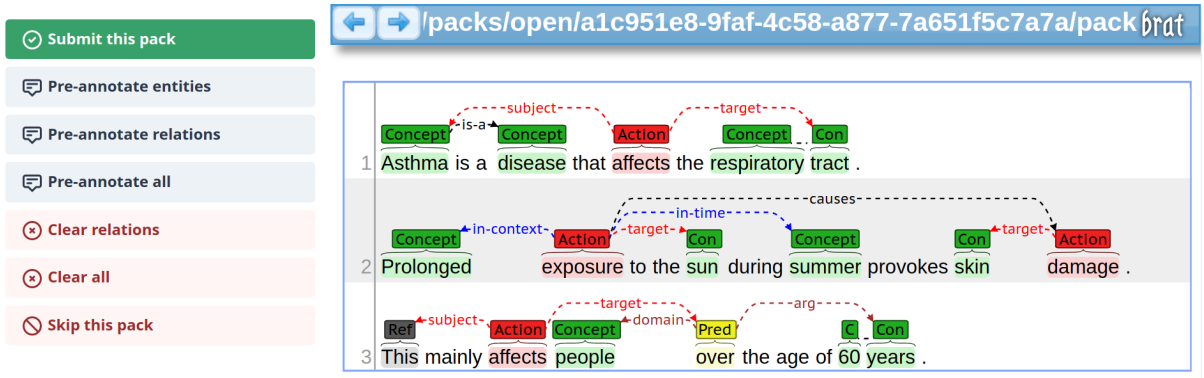

Figure 2: Screenshot of the web application prototype for semi-automatic corpus annotations integrated with the BRAT annotation tool. The right panel shows an illustrative selection of annotated sentences in the schema of the eHealth-KD 2020 corpus (Piad-Morffis et al., 2020), see Section 5.

application developed in this research.

We simulated the assisted annotation process to evaluate the effect to annotating the corpus using active learning strategies versus annotating the corpus in the original order without suggestions (baseline). As the process of annotating a corpus is expensive, it was simulated using the gold annotations in the training collection. The improvement can be estimated by comparing how many sentences need annotating to reach a specific performance of the machine learning algorithms (measured in terms of F_1 in the testing collection).

To study the relative impact of the different components of our query strategy, we evaluated three different variants. They consisted of using the full query strategy proposed in Section 3 with $\beta = 1$, as well as considering only entity uncertainty $\hat{H}(E_i)$ and relation uncertainty $\hat{H}(R_i)$ respectively. Figure 3 shows how the F_1 metric improved with each batch of sentences, for the first 500 sentences. The target F_1 is the final score obtained by training the models M_E and M_R on the full 1000 sentences of the corpus. In general, the curves that correspond to the active learning strategy (i.e., assisted variants) approach the target F_1 significantly faster than the unassisted baseline.

To illustrate the degree of time-reduction achieved, Figure 4 shows the minimum number of sentences that must be annotated to reach different relative target F_1 scores. For example, after annotating the first 400 sentences it is possible to achieve a 95% of the ultimate F_1 score when using all the corpus. However, to reach the target score, the first 880 out of a 1000 sentences must be annotated if the corpus is annotated in the original order (baseline). By contrast, using our active learning strategy only between 530 to 580 sentences must be annotated to reach the same target F_1, thereby saving between a 35% and a 40% of human annotation time.

Another interesting finding is to estimate the extent to which the suggested annotations further reduce the total annotation time. A human annotator using the tool will need to accept some of the suggested annotations, correct the ones that are wrong and annotate the ones that are missing. Each of these actions has a different cost in time. For quantifying the improvement in overall time that the suggestions produce, we assigned a relative cost (in terms of abstract time units) to each of the following types of annotations:

Missing annotations: annotations that the model did not suggest and the human annotator must produce, have a cost of 1 time unit.

Spurious annotations: annotations that the model suggested and are wrong, which must be eliminated by the human annotator, have a cost of 2 time units.

Correct annotations: since the human annotator must at least recognize the annotation is correct, the cost is 0.25 time units.

Partial annotations: annotations that are partially correct either because the spans are partially covering or the label is wrong, have a cost of 0.5 time units.

This cost structure assumes that the problem of correcting wrong annotations is more complex than simply producing the correct annotations, while acknowledging that even agreeing with correct annotations has a non-zero cost. For an active learning

strategy to be helpful it must provide enough correct annotations to outweigh the cost of correcting the wrong annotations; hence, it should prioritize precision over recall.

Figure 5 shows the relative effect (in terms of reducing the overall annotation time) of enabling annotation suggestions for different combinations of the entity and relation thresholds u_e and u_r. It can be observed that on average, the entity suggestions produce a positive effect (green color) across a wide range of thresholds, while the relation suggestions tend to produce a negative effect as more suggestions are allowed. This is a direct consequence of the M_R relation model's performance, which achieves at most an F_1 score of 0.27, while the entity model M_E achieves up to a 0.78 score. The optimal time reduction is achieved for an entity threshold $u_e = 2.4$ and a relation threshold $u_r = 0$, producing an estimated 24% reduction in the total annotation time. Interestingly, these parameters result in an overall performance for the machine learning model of $F_1 = 0.54$, with a precision of 0.78 and a recall of 0.41. As expected, given the asymmetrical cost structure, it is preferable to prioritize precision rather than recall for the annotation suggestions.

An overall reduction in annotation time for this experimental simulation can be estimated by combining the improvements provided by annotation suggestions and the sentence ordering. Assuming both effects are independent, the best case scenario for this corpus suggests the following. Using the active learning approach a human annotator would have needed to annotate only 530 sentences out of 1000, each of them with an estimate time cost of 76% compared to the full annotation. This results in an overall reduction of as much as 60% of the total annotation time, producing a smaller corpus on which machine learning models can still be trained, delivering the same performance as those trained on the original corpus.

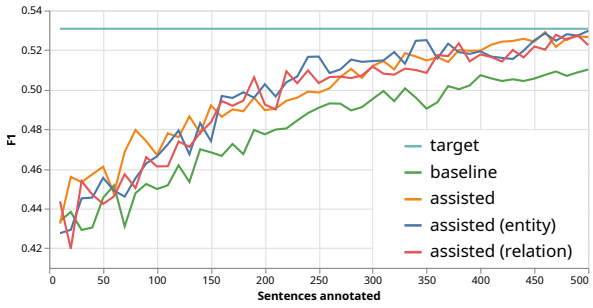

Figure 3: Iterative improvement of the machine learning model performance in terms of F_1 with and without using the active learning strategy for sentence suggestion.

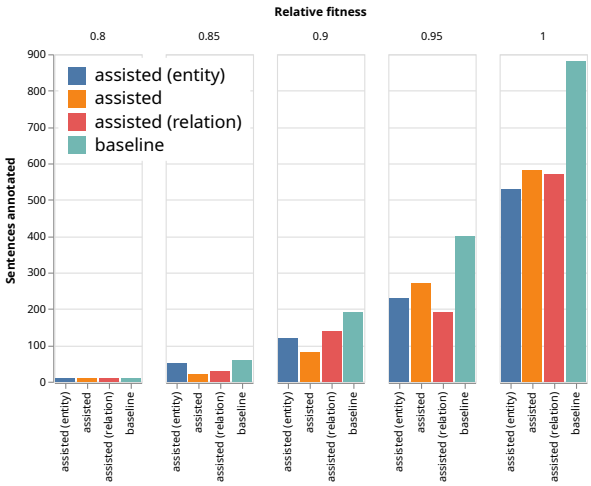

Figure 4: Minimum number of sentences necessary to reach a specific performance relative to the target F_1 with each active learning strategy for sentence suggestion.

6 Discussion and Future Work

The machine learning model used for entity recognition M_E achieves a result comparable with the state-of-the-art in this corpus while being simple enough to be trained during annotation. Not only does the model produce a significant reduction in the number of sentences that need to be annotated but it also is capable of pre-annotating entities that are often correct, even when factoring in the significantly higher cost of correcting the wrong suggestions. By contrast, the relation extraction model M_R performed significantly worse than current state-of-the-art in this corpus. However, even if the pre-annoted relations suggested by this model are on average not beneficial, it is interesting to note that using only the uncertainty of relations $\hat{H}(R_i)$ as a query strategy still produces a significant time reduction (see Figure 4, *assisted (relations)*). Unfortunately, all good performing models for this problem are composed of complex deep learning architectures that cannot be trained sufficiently fast enough to be used during the annotation process in a commodity hardware.

Indeed, using simple models is necessary in active learning scenarios where algorithms must be trained interactively, but even without considering this issue, there are additional factors to con-

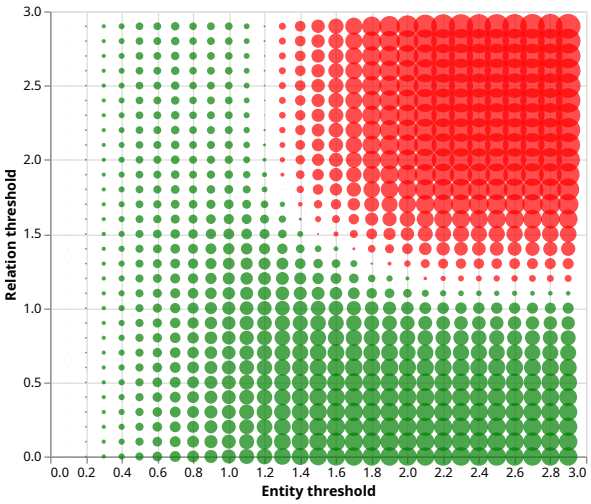

Figure 5: Effect of using entity and relation suggestions for different values of the entity uncertainty threshold u_e and relation uncertainty threshold u_r. Green values indicate a reduction in the annotation cost in terms of relative time units.

sider related to model complexity. We argue that an interesting trade-off exists between the capacity of a model and its usefulness for active learning. Very simple models (underfit) will have a high uncertainty in all samples, while very complex models (overfit) will overestimate their certainty. In both cases, the informativeness $I(\cdot)$ for all sentences will be very similar, and there is no sensible way to choose the most informative ones. This suggests that there may be an optimal middle ground where the model learns enough to provide useful suggestions while still maintaining a healthy level of uncertainty. The fact that even weak baselines (like the relation model M_R) are still a useful source of information when actively selecting unlabeled sentences is one surprising conclusion of our research. This seems to suggest that even in very complex scenarios where state-of-the-art models are impossible to train interactively, using weaker surrogate models can still provide a significant benefit for human-in-the-loop learning.

Regarding the generalization of our approach, the fact that the corpus is in the Spanish language is irrelevant for our experimental results since the machine learning models used are language-agnostic and no language-specific heuristics were applied. Hence, these results should generalize to other languages and annotation schemas albeit with different baseline F_1 scores according to the complexity of the underlying learning problem. Nevertheless, we are interest in evaluating our approach using languages other than English since the creation of linguistic resources is one of the main difficulties of NLP research, especially for other languages. An ongoing research priority is to validate this strategy on other corpora with different annotation schemas.

In future work, we will explore how to explicitly control the complexity of a model during the active learning process by controlling the model's capacity. Two strategies that can be analyzed are the use of ensemble methods and deep learning architectures with early stopping. In both cases, the intuitive idea is to iteratively refine a machine learning model up to the point where a sufficiently good performance is achieved but before the model overfits on the small labelled set of sentences, such that uncertainty measures are still relevant. Another interesting scenario in which to apply this approach is when many annotators exist for the same text. In this case, the models can learn contradictory hypotheses due to differences between annotators. Interestingly, in the case of active learning, this is a positive phenomena, since inter-annotator disagreement is a good measure of sentence difficulty. Active learning models trained on a pool of sentences with multiple, possibly contradictory annotations, will naturally tend to select sentences that are more likely to cause disagreement between annotators. In this context, it can be interesting to explore query-by-ensemble methods where each model in the ensemble is trained on a different annotator's pool to maximize model variance.

7 Conclusions

In this article, we present an approach for reducing the time involved in manually annotating a corpus of natural language sentences that contains entities and relations. This approach uses active learning with uncertainty sampling and weighted density, and provides an estimated reduction of 60% of total annotation time in a simulated experiment with a real corpus. This improvement is derived from two independent factors: intelligently sorting which sentences to annotate and providing pre-annotated suggestions with a high-degree of certainty. The proposed strategies have been implemented into a computational tool that is applicable to a broad range of corpus annotation schemas and is available for the research community.

References

Vincent Claveau and Ewa Kijak. 2017. Strategies to Select Examples for Active Learning with Conditional Random Fields. In *CICLing*.

David Cohn. 2010. Active Learning. In *Encyclopedia of Machine Learning*.

Garry Allan Einicke, Haider A Sabti, David V Thiel, and Marta Fernandez. 2017. Maximum-entropy-rate selection of features for classifying changes in knee and ankle dynamics during running. *IEEE journal of biomedical and health informatics*, 22(4):1097–1103.

Lisheng Fu and Ralph Grishman. 2013. An efficient active learning framework for new relation types. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 692–698.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep Bayesian Active Learning with Image Data. In *ICML*.

Ning Gao, Nikos Karampatziakis, Rahul Potharaju, and Silviu Cucerzan. 2019. Active Entity Recognition in Low Resource Settings. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*.

Janez Kranjc, Jasmina Smailovic, Vid Podpecan, Miha Grcar, Martin Znidarsic, and Nada Lavrac. 2015. Active learning for sentiment analysis on data streams: Methodology and workflow implementation in the ClowdFlows platform. *Inf. Process. Manag.*, 51:187–203.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196.

David D. Lewis and Jason Catlett. 1994. Heterogeneous uncertainty sampling for supervised learning. In William W. Cohen and Haym Hirsh, editors, *Machine Learning Proceedings 1994*, pages 148 – 156. Morgan Kaufmann, San Francisco (CA).

Bill Yuchen Lin, Dongho Lee, Frank F. Xu, Ouyu Lan, and Xiang Ren. 2019. AlpacaTag: An Active Learning-based Crowd Annotation Framework for Sequence Tagging. In *ACL*.

Venkata Vamsikrishna Meduri, Lucian Popa, Prithviraj Sen, and Mohamed Sarwat. 2020. A Comprehensive Benchmark Framework for Active Learning Methods in Entity Matching. *ArXiv*, abs/2003.13114.

Fredrik Olsson. 2009. A literature survey of active machine learning in the context of natural language processing.

Alejandro Piad-Morffis, Juan P. Consuegra-Ayala, Yoan Gutiérrez, and Suilan Estevez-Velarde. 2020. eHealth-KD 2020.

Alejandro Piad-Morffis, Yoan Guitérrez, Suilan Estevez-Velarde, and Rafael Muñoz. 2019a. A General-Purpose Annotation Model for Knowledge Discovery: Case Study in Spanish Clinical Text. pages 79–88.

Alejandro Piad-Morffis, Yoan Gutiérrez, Juan Pablo Consuegra-Ayala, Suilan Estevez-Velarde, Yudivián Almeida-Cruz, Rubio Muñoz, and Andrés Montoyo. 2019b. Overview of the eHealth Knowledge Discovery Challenge at IberLEF 2019. In *IberLEF@SEPLN*.

Nicholas Roy and Andrew McCallum. 2001. Toward optimal active learning through monte carlo estimation of error reduction. *ICML, Williamstown*, pages 441–448.

Burr Settles. 2009. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences.

Burr Settles and Mark Craven. 2008. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1070–1079.

Burr Settles, Mark Craven, and Soumya Ray. 2008. Multiple-Instance Active Learning. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1289–1296. Curran Associates, Inc.

H. S. Seung, M. Opper, and H. Sompolinsky. 1992. Query by Committee. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, page 287–294, New York, NY, USA. Association for Computing Machinery.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107. Association for Computational Linguistics.

Roselyne Tchoua, Aswathy Ajith, Zhi Hong, Logan T. Ward, Kyle Chard, Debra Audus, Shrayesh Patel, Juan de Pablo, and Ian T Foster. 2019. Active Learning Yields Better Training Data for Scientific Named Entity Recognition. *2019 15th International Conference on eScience (eScience)*, pages 126–135.

Antonio Jimeno Yepes, Mariana Neves, and Karin Verspoor. 2013. Brat2BioC: conversion tool between brat and BioC. In *Proceedings of the Fourth BioCreative Challenge Evaluation Workshop*, volume 1, pages 46–53.

Hong-Tao Zhang, Min-Lie Huang, and Xiao-Yan Zhu. 2012. A unified active learning framework for biomedical relation extraction. *Journal of Computer Science and Technology*, 27(6):1302–1313.