Improvements in Information Extraction in Legal Text by Active Learning

Cristian Cardellino, Laura Alonso Alemany, Serena Villata, Elena Cabrio

Improvements in Information Extraction in Legal Text by Active Learning

Cristian CARDELLINO [a,1], Laura ALONSO ALEMANY [a], Serena VILLATA [b] and Elena CABRIO [c]

[a] *University of Cordoba*
[b] *INRIA Sophia Antipolis*
[c] *University of Nice-Sophia Antipolis*

Abstract. Managing licensing information and data rights is becoming a crucial issue in the Linked (Open) Data scenario. An open problem in this scenario is how to associate machine-readable licenses specifications to the data, so that automated approaches to treat such information can be fruitfully exploited to avoid data misuse. This means that we need a way to automatically extract from a natural language document specifying a certain license a machine-readable description of the terms of use and reuse identified in such license.

Ontology-based Information Extraction is crucial to translate natural language documents into Linked Data. This connection supports consumers in navigating documents and semantically related data. However, the performances of automated information extraction systems are far from being perfect, and rely heavily on human intervention, either to create heuristics, to annotate examples for inferring models, or to interpret or validate patterns emerging from data.

In this paper, we apply different Active Learning strategies to Information Extraction (IE) from licenses in English, with highly repetitive text, few annotated or unannotated examples available, and very fine precision needed. We show that the most popular approach to active learning, i.e., uncertainty sampling for instance selection, does not provide a good performance in this setting. We show that we can obtain a similar effect to that of density-based methods using uncertainty sampling, by just reversing the ranking criterion, and choosing the *most certain* instead of the *most uncertain* instances.

Introduction

The need to automatically translate legal texts describing licenses, contracts, technical documents into machine-readable ones is becoming more and more important in order to allow for the automated processing, verification, etc of such texts. However, as we are in the legal field, such machine-readable formulation of the licenses requires a high degree of reliability.

[1]Corresponding Author: ccardellino@famaf.unc.edu.ar

The goal of IE for textual licenses is to identify a *prohibition*, a *permission* or an *obligation* expressed by a license. When these fragments are identified, they are converted into an RDF machine-readable specification of the license itself using the ODRL vocabulary[2]. It defines the classes to which each text fragment needs to be translated by the system, and specifies different kinds of Policies (i.e., Agreement, Offer, Privacy, Request, Set and Ticket). We adopt `Set`, a policy expression that consists in entities from the complete model. Permissions, prohibitions and duties (i.e., the requirements specified in CC REL) are specified in terms of an `action`. For instance, we may have the action of attributing an `asset` (anything which can be subject to a policy), i.e., `odrl: action odrl: attribute`.

Given the high precision required for this kind of information extraction, a supervised approach seems the method of choice. Active learning techniques [15] aim to get powerful insights on the inner workings of automated classifiers and resort to human experts to analyze examples that will most improve their performance. Active Learning is used to address this problem in the following way. Starting from a small manually annotated dataset, a model is learnt. Then, the model is applied to an unannotated dataset, and instances in this dataset are ranked according to the certainty of the model to label them, ranking highest those with most certainty or with most uncertainty. The highest ranking instances are presented to the oracle, who annotates them, associating each instance to one or more of the classes defined by the ODRL ontology or the class "null" if none of the available classes apply for the instance. Finally, the system is trained again with the annotated corpus, now enhanced with the newly annotated examples.

However, the simplest, most popular active learning approach, namely uncertainty-based instance sampling, does not perform well in these conditions: a skewed distribution, with scarcely populated minority classes and a catch-all majority class that is difficult to singularize. Density-based methods tend to work better in such contexts. However, such methods are complex and difficult to put into practice by the average practitioner. We show that a very simple approach, selecting for annotation those instances where the classifier is most certain (*reversed uncertainty sampling*) does provide a clear improvement over the passive learning and uncertainty sampling approaches. Indeed, it is known that uncertainty sampling does not work well with skewed distributions or with few examples, in those cases, density estimation methods work best. We show that using *reversed uncertainty sampling* in this particular context yields results in the lines of density estimation methods.

Experimental results show that if we choose to annotate first those instances where the classifier shows more uncertainty, the performance of the system does not improve quickly, and, in some cases, it improves more slowly than if instances are added at random.

The rest of the paper is organized as follows: in Section 1 we discuss the active learning approach and related work. Section 2 explains how we apply active learning techniques to IE to this problem, and experimental results comparing different approaches are discussed in Section 3.

[2] `http://www.w3.org/community/odrl/`

1. Relevant work

Active learning [15] is a more "intelligent" approach to machine learning, whose objective is to optimize the manual labelling of examples or features. This optimization is obtained by choosing examples to be manually labelled, by following some given metric or indicator to maximize the performance of a machine learning algorithm, instead of choosing them randomly from a sample. This capability is specially valuable in the context of knowledge-intensive Information Extraction, where very obtaining examples is costly and therefore optimizing examples becomes crucial.

The process works as follows: the algorithm inspects a set of unlabeled examples, and ranks them by how much they could improve the algorithm's performance if they were labelled. Then, a human annotator (the so-called "oracle") annotates the highest ranking examples, which are then added to the starting set of training examples from which the algorithm infers its classification model, and the loop begins again. In some active learning approaches, the oracle may annotate features describing instances, and not (only) instances themselves. This latter approach provides even faster learning in some cases [6,13,10,11].

Concerning the labelling of instances, different strategies have been applied to determine the most useful instances to be labelled by a human judge, including expected model change, expected error reduction or density-weighted methods [14]. The most intuitive and popular strategy is *uncertainty sampling* [9], which chooses those instances or features where the algorithm is most uncertain. This strategy has been successfully applied to Information Extraction tasks [3,12]. Uncertainty can be calculated by different methods depending on the learning algorithm. Specially popular are methods exploiting the margins of Support Vector Machines, as in [16]. The simplest methods exploit directly the certainty that the classifier provides for each instance that is classified automatically. This is the information that we are exploiting.

However, we did not only use uncertainty sampling, but also the exact opposite. We explored both prioritizing items where the classifier is least certain and where it is least most certain. We followed the intuition that, when a model is very small, based on very few data, it can be improved faster by providing evidence that consolidates the core of the model. This is achieved by choosing items with highest certainty, which would help to define more accurately the generative centers of the model, and can help to redirect wrong assumptions that a model with very few data can easily make. This intuition, which lies at the heart of well-known semi-supervised learning techniques like self-training (or *bootstrapping*), has also been noted by approaches combining density estimation methods when very few examples are available, and uncertainty sampling when the training dataset has grown [5,17].

Other approaches have been applied to fight the problem of learning with few examples, by finding the optimal seed examples to build a training set [4,7]. However, these approaches are complex and difficult to implement, thus lie beyond the capacities of the regular NLP practitioner. In contrast, the approach presented here is conceptually simple and easy to implement, as it is a wrapper method over your best-know classifier.

We started from the NLL2RDF framework proposed by Cabrio et al. [1], a system generating RDF licenses from natural language documents specifying such licenses. However, the performances of the framework were not satisfiable, considering the high degree of reliability required for such machine-readable formulation of the licenses, i.e., if the original license states that action A is forbidden and this prohibition is not reported in the RDF version of the license then this could lead to misuses of the data that is associated to that machine-readable license. For this reason, we need to revise the architecture of the NLL2RDF framework to improve the performances and the coverage, leading in this way to an improvement of the reliability of the system itself. An example of how NLL2RDF works is as follows:

Example 1.1. The aim of the NLL2RDF system consists in starting from the natural language formulation of a license, provided by the user, and then translating it in an RDF representation using the ODRL vocabulary. Let us consider the following text extracted from the Open Government License[3]:

> The Licensor grants you a worldwide, royalty-free, perpetual, non-exclusive licence to use the Information subject to the conditions below. This licence does not affect your freedom under fair dealing or fair use or any other copyright or database right exceptions and limitations. You are free to:
>
> - copy, publish, distribute and transmit the Information;
> - adapt the Information;
> - exploit the Information commercially and non-commercially for example, by combining it with other Information, or by including it in your own product or application.
>
> You must, where you do any of the above:
>
> - acknowledge the source of the Information by including any attribution statement specified by the Information Provider(s) and, where possible, provide a link to this licence;

The resulting RDF license generated automatically by NLL2RDF is as follows:

```
:ukOGL2.0 a odrl:Set ;
          odrl:duty [
            a odrl:Duty ;
            odrl:action odrl:attribute, odrl:attachPolicy
          ] ;
          odrl:permission [
            a odrl:Permission ;
            odrl:action odrl:derive, odrl:distribute, odrl:reproduce
          ] .
```

We developed [2] an active learning tool inspired on Dualist [13] to improve the accuracy of NLL2RDF. In the first approach adopted to develop the NLL2RDF system [1], a Support Vector Machine classifier was used. Texts were characterized by the unigrams, bigrams and trigrams of lemmas, obtaining an

[3] http://www.nationalarchives.gov.uk/doc/open-government-licence/version/2/

f-measure that ranged from 0.3 to 0.78 depending on the class, with 0.5 average. Later on we included bigrams and trigrams of words that co-occur in a window of three to five words. This last feature is aimed to capture slight variations in form that convey essentially the same meaning, as in the following example:

Example 1.2.

> Each party agrees to comply strictly with all such laws and regulations and acknowledges their responsibility to obtain such licenses to export, re-export, or import as may be required.
>
> Each party shall comply fully with all such laws and regulations and acknowledges its responsibility to obtain such licenses to export, re-export or import as may be required.
>
> You agree to comply strictly with all such laws and regulations and acknowledge that you have the responsibility to obtain such licenses to export, re-export, or import as may be required.

where the skipgram "`acknowledge - responsibility`" is common to all utterances, capturing their similarity in meaning.

As in Dualist, we provide a graphical user interface for the human oracle to answer the queries of the active learning algorithm. The base machine learning algorithm is also a Multinomial Naïve Bayes, but our method for ranking instances is uncertainty/certainty sampling based on the confidence of the classifier. Features can also be labelled, using Information Gain to select them, but sequentially with respect to instances, not simultaneously as in Dualist. As an addition, our approach allows for multiclass labeling, that is, an instance can be labelled with more than one class. Our active learning framework source together with the dataset is available at `https://github.com/crscardellino/nll2rdf-active-learner`.

In the present paper, we go further in the experimental evaluation presented in Cardellino et al. [2] showing that the most popular approach to active learning, i.e., uncertainty sampling for instance selection, does not provide a good performance in this setting. Density-based methods are the usual alternative to uncertainty sampling, in contexts with very few labelled instances. In this paper, we show that we can obtain a similar effect to that of density-based methods using uncertainty sampling, by just reversing the ranking criterion, and choosing most certain instances instead of most uncertain ones.

2. Learning strategies

As a baseline to assess the improvement provided by the active learning approach to the problem, we assess the performance of two passive learning approaches, Support Vector Machines and Multinomial Naive Bayes. As can be seen in Table 1, the best performing out-of-the-box approach was a Support Vector Machine classifier, which reached an average accuracy of 76%, against 63% for Multinomial Naïve Bayes. We then applied feature selection as a preprocessing step. We calculated the Information Gain of each feature with respect to the classes, and

	plain	FS	one vs. all	one vs. all & (class-specific) FS
Support Vector Machine	76	76	71	(73) 73
Multinomial Naïve Bayes	63	72	60	(**83**) 78

Table 1. Accuracy of two passive learning classifiers with different configurations.

kept only the 50 features with most Information Gain, as long as they all had an Information Gain over 0.001, those with Information Gain below that threshold were discarded. This improved the performance of Multinomial Naïve Bayes but not of Support Vector Machines. We then applied one vs. all classification, where a different classifier is trained to distinguish each individual class from all the rest. This, combined with a separate Feature Selection (FS) preprocess for each of the classifiers yields a significant improvement in performance for Multinomial Naïve Bayes, reaching an accuracy of 83%, while the performance of Support Vector Machines decreases. Given these results, we decided to use a base classifier for active learning Multinomial Naïve Bayes with one-vs.-all classification and Information Gain feature selection for each classifier.

As a base classifier for active learning we used Multinomial Naïve Bayes in a one-vs-all setting with separated feature selection for each classifier. The system keeps for each one-vs.-all classifier only the top 50 features with highest Information Gain with the class or only those features with more than 0.001 Information Gain with the class, whichever condition produces the biggest set.

We compare two different approaches to select unannotated instances to be labelled first: those where the classifier is most uncertain (*uncertainty sampling*) and those where the classifier is most certain (*reversed uncertainty sampling*).

3. Experimental setting

In this section, we first describe the dataset we use to evaluate our active learning approach in Section 3.1 and the evaluation metrics (Section 3.2) we adopted to analyze the results, that are finally discussed in Section 3.3.

3.1. Dataset

To compare the performance of different learning methods we used a manually annotated dataset of licenses. The corpus consists of the original labelled set of 37 licenses, totalling 41,340 words, and an unlabeled set of 396 licenses, totalling 482,259 words. It is composed of software licenses, source code licenses, data licenses, and content licenses; they are public as well as private domain licenses, totalling 162 labelled instances, with a mean of 12.46 instances per class.

The class with the most instances is *permission-to-distribute*, with a total of 33 instances, while there are three classes with just one instance: *permission-to-read, prohibition-to-derive and requirement-to-attach-source*. We discarded classes with less than 5 labelled instances because we could not carry out and evaluate a useful simulation with so few instances.

The training and evaluation corpus have been tagged previously and each instance was assigned to a single class. It must be noted that the majority of

sentences in the corpus do not belong to any of the classes established by the ODRL vocabulary. In the classification setting, these examples belong to the class "null", which is actually composed of several heterogeneous classes with very different semantics, with the only common factor that their semantics are not captured by the ODRL vocabulary. An example of the class "null" follows:

Example 3.1. `Users are entirely responsible, to the exclusion of the Author and any other persons, for compliance with (1) regulations set by owners or administrators of employed equipment, (2) licensing terms of any other software, and (3) local regulations regarding use, including those regarding import, export, and use of encryption software.`

For the manual dataset annotation we tokenized the sentences using Stanford Parser [8], and we then added the annotation of the relation (e.g., *permission*, *prohibition*, etc.). The Stanford Parser is also used to parse the instances of the unannotated corpus. From the unannotated corpus, sentences are taken as instances to be annotated by the automated classifier or the oracle.

3.2. Evaluation methods and metrics

The evaluation task is done with an automated simulation of the active learning loop on the annotated corpus. In this simulation, from the 156 original instances on the corpus, we started with an initial random set of 20 instances (roughly 12% of the annotated corpus). From this initial set the first model was learned. After each iteration, the model was evaluated using 10-fold cross-validation on the corpus used for training the current model.

With this initial model, we proceed to use the rest of the annotated instances as the unannotated corpus. With the data from the first model we carry out the selection of the queries from this "unannotated corpus" for manual annotation. In our simulation, manual annotation is substituted by providing, for each example, the label that is associated to it in the corpus.

Once again the annotated corpus is used in a second iteration for creation and evaluation of a new model. The process is repeated until all the "unannotated" instances are assigned their label. The number of newly annotated instances per iteration in our experiments is: 1, 3, 5 and 10.

The goal of this simulation is to show the steep of the curves in each one of the query selection methods in comparison to each other, with the steepest slope as the best query selection strategy.

3.3. Analysis of results

In Figure 1 we can see the learning curves of our active learning approach, obtained as described in the previous Section. We can see that the "most certain" strategy performs consistently better than the passive and most uncertain strategies, improving performance with fewer instances. The other two perform comparably if the number of instances added at each iteration is high, and the "most uncertain" approach performs even worse than the passive approach (random) if

instances are added one at a time for each iteration. These results confirm our hypothesis that, for models inferred from very few training examples, maximizing the entropy of examples is not useful, while providing more evidence to define the core of the classes provides an improvement in performance.

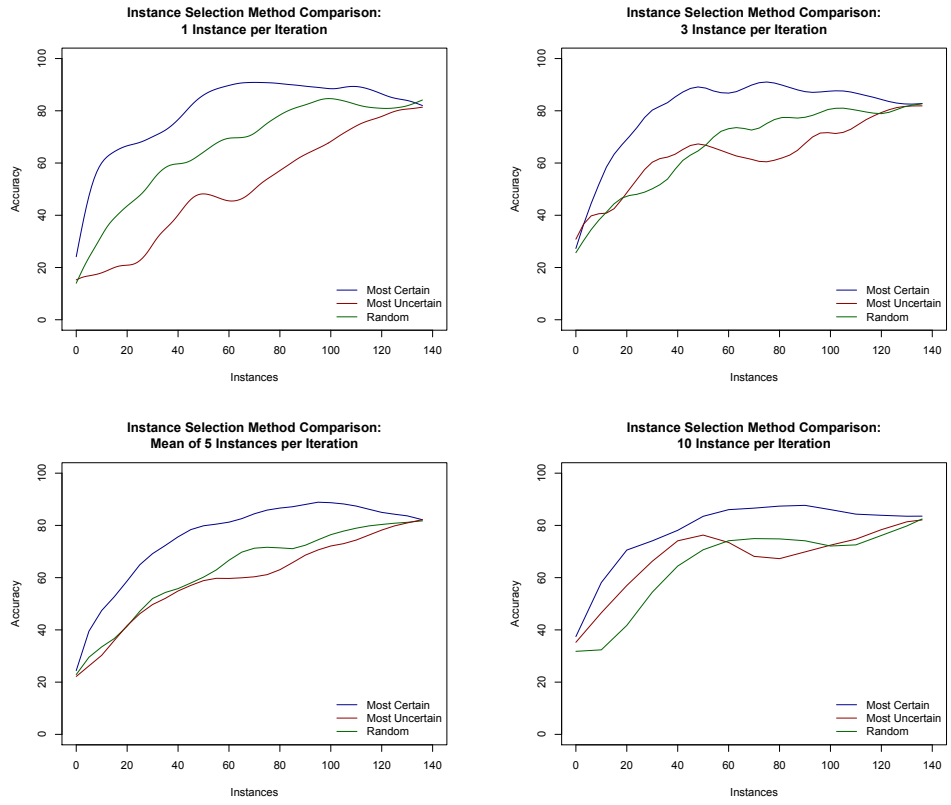

Figure 1. Learning curves of active learning approaches with different policies for instance selection. In the y axis we depict accuracy, in the x axis, the number of instances added to training, and the different lines represent different strategies to pick the instances to be added in each iteration to the training corpus: random, ranked by most certain or by most uncertain.

When examples are selected applying the "most uncertain" strategy are, they mostly belong to the "null" class, that is, they do not signal any of the classes relevant for the problem.

Providing examples for the class "null" is specially harmful for the resulting model for two main reasons. First, it grows the majority class, while small classes are kept with the same few examples, thus adding the problem of having an imbalanced dataset to the problem of having small classes with few instances. Second, the class "null" is composed by many heterogeneous classes that are not included in the ODRL vocabulary, so its characterization is difficult and may be misleading.

Besides this configuration of classes, which can be found in very different domains, the domain of IE in licenses and legal text in general may be specially

prone to an improvement of performance by labeling most certain examples first, because licenses and legal texts in general tend to be very formulaic, repeating the same wordings with very few variations, and small differences in form may signal differences in meaning, much more than in other domains, where differences in meaning are signaled by bigger differences in wordings.

Results for the best performances achieved by different passive learning approaches are summarized in Table 1. Those results were obtained using the whole dataset, corresponding to the rightmost extreme in the graphics of Figure 1.

4. Conclusions

In this paper, we have shown that, for the problem of inferring a classifier for legal text, where few labelled instances are available, active learning does provide a faster learning curve than traditional machine learning approaches, but only if the most certain examples are selected to be hand-tagged, in contrast with the most frequent approach in active learning, called *uncertainty sampling*, where human annotators are given to annotate examples that the classifier is most uncertain about.

We are planning to compare the results of reversed uncertainty sampling with those of density-based approaches. We also intend to carry out experiments in other NLP problems where a catch-all class is naturally found, like in word sense disambiguation where only some of the senses of a word are distinguished.

We will also explore with actual active learning, with human oracles involved in the loop instead of simulations. This will be especially enlightening for minority classes and to assess the impact of the catch-all class, which is a special case of the well-known skewed class distribution problem in active learning.

We plan to evaluate the performance of the system also using feature labeling by itself and in combination with instance labeling.

Moreover, we are planning to integrate other vocabularies for modelling Law in Artificial Intelligence, like those wihtin the LegalRuleML metamodel. We are also looking into integrating exceptions in a principled manner within an active learning approach to this problem.

References

[1] Elena Cabrio, Alessio Palmero Aprosio, and Serena Villata. 2014. These are your rights - A natural language processing approach to automated RDF licenses generation. In The Semantic Web: Trends and Challenges - 11th International Conference, ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014. Proceedings, pages 255–269.

[2] Cristian Cardellino, Serena Villata, Laura Alonso Alemany and Elena Cabrio. 2015. Information Extraction with Active Learning: A Case Study in Legal Text. In Computational Linguistics and Intelligent Text Processing - 16th International Conference, CICLing 2015, pages 483-494.

[3] Aron Culotta and Andrew McCallum. 2005. Reducing labeling effort for structured prediction tasks. In Proceedings of the 20th National Conference on Artificial Intelligence - Volume 2, AAAI'05, pages 746–751. AAAI Press.

[4] Dmitriy Dligach and Martha Palmer. 2011. Good seed makes a good crop: Accelerating active learning using language modeling. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2, HLT '11, pages 6–10, Stroudsburg, PA, USA. Association for Computational Linguistics.

[5] Pinar Donmez, Jaime G. Carbonell, and Paul N. Bennett. 2007. Dual strategy active learning. In Joost N. Kok, Jacek Koronacki, Ramon López de Mántaras, Stan Matwin, Dunja Mladenic, and Andrzej Skowron, editors, ECML, volume 4701 of Lecture Notes in Computer Science, pages 116–127. Springer.

[6] G. Druck, B. Settles, and A. McCallum. Active learning by labeling features. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 81–90. ACL, 2009.

[7] Michael Kearns. 1998. Efficient noise-tolerant learning from statistical queries. J. ACM, 45(6):983–1006, November.

[8] Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1, ACL-2003, pages 423-430, Stroudsburg, PA, USA. Association for Computational Linguistics.

[9] David D. Lewis and Jason Catlett. 1994. Heterogeneous uncertainty sampling for supervised learning. In In Proceedings of the Eleventh International Conference on Machine Learning, pages 148–156. Morgan Kaufmann.

[10] Jay Pujara, Ben London and Lise Getoor. 2011. Reducing label cost by combining feature labels and crowdsourcing. In ICML Workshop on Combining Learning Strategies to Reduce Label Cost.

[11] Christopher T. Symons and Itamar Arel. 2011. Multi-View Budgeted Learning under Label and Feature Constraints Using Label-Guided Graph-Based Regularization.

[12] B. Settles and M. Craven. 2008. An analysis of active learning strategies for sequence labeling tasks. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 1069–1078. ACL.

[13] B. Settles. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 1467–1478. ACL, 2011.

[14] B. Settles. 2009. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

[15] B. Settles. 2012. Active Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool.

[16] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. J. Mach. Learn. Res., 2:45–66, Mar. 2002.

[17] Jingbo Zhu, Huizhen Wang, Tianshun Yao, and Benjamin K Tsou. 2008. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1, pages 1137–1144. Association for Computational Linguistics.