Factors Influencing Robustness and Effectiveness of Conditional Random Fields in Active Learning Frameworks

Mahnoosh Kholghi[1,2], Laurianne Sitbon[1], Guido Zuccon[1], Anthony Nguyen[2]
[1]Science and Engineering Faculty, Queensland University of Technology, Brisbane 4000, Queensland, Australia
[2]The Australian e-Health Research Centre, CSIRO, Brisbane 4029, Queensland, Australia
{m1.kholghi, laurianne.sitbon, g.zuccon}@qut.edu.au, anthony.nguyen@csiro.au

Abstract

Active learning approaches reduce the annotation cost required by traditional supervised approaches to reach the same effectiveness by actively selecting informative instances during the learning phase. However, effectiveness and robustness of the learnt models are influenced by a number of factors. In this paper we investigate the factors that affect the effectiveness, more specifically in terms of stability and robustness, of active learning models built using conditional random fields (CRFs) for information extraction applications. Stability, defined as a small variation of performance when small variation of the training data or a small variation of the parameters occur, is a major issue for machine learning models, but even more so in the active learning framework which aims to minimise the amount of training data required. The factors we investigate are a) the choice of incremental vs. standard active learning, b) the feature set used as a representation of the text (i.e., morphological features, syntactic features, or semantic features) and c) Gaussian prior variance as one of the important CRFs parameters. Our empirical findings show that incremental learning and the Gaussian prior variance lead to more stable and robust models across iterations. Our study also demonstrates that orthographical, morphological and contextual features as a group of basic features play an important role in learning effective models across all iterations.

Keywords: active learning, robustness, effectiveness, conditional random fields, Gaussian prior variance, concept extraction.

1 Introduction

Concept extraction is a significant initial step in any information extraction system and includes recognising meaningful entities and assigning them to predefined classes (e.g., person, organization, location; in the medical domain: problem, test, treatment) (Nadkarni et al., 2011). In this paper we use datasets and tasks in the clinical domain, where the concepts to be extracted are clinical concepts and the documents are clinical records.

The three main approaches to extract target concepts and entities from free text resources are dictionaries, rules and machine learning. Target entities usually appear as multi-token sequences in the text; these often cannot be captured directly using only lexical resources (Gurulingappa, 2012; Meystre et al., 2008; Roberts, 2012). For example in this sentence from the i2b2/VA 2010 dataset (Uzuner et al., 2011):

"*She had a workup by her neurologist and an MRI revealed a C5-6 disc herniation with cord compression and a T2 signal change at that level.*"

"*a C5-6 disc herniation*" is a multi-token concept of type "*problem*".

Manually creating resources or rules for dictionary and rule-based approaches is not only expensive and time-consuming, but also is a complex, error prone task. Additionally, these approaches are not adaptable and scalable to other domains and languages (Gurulingappa, 2012; Meystre, et al., 2008; Nadkarni, et al., 2011; Roberts, 2012).

Machine learning-based approaches have been extensively leveraged to extract concepts in several information extraction tasks (Jiang, 2012; Piskorski & Yangarber, 2013).

Since they have first been proposed in 2001, Conditional random fields (Lafferty et al., 2001) and in particular linear-chain CRFs, have shown the most promising results among other supervised machine learning algorithms to extract entities and concepts from text, in particular in the clinical domain (Suominen et al., 2013; Uzuner et al., 2008; Uzuner et al., 2010; Uzuner, et al., 2011). This motivates the use of linear chain CRFs in our active learning-based framework. CRFs approaches are supervised and therefore require fairly large amounts of high quality annotated data to build powerful statistical models. Creating the required annotated data to train a supervised model is laborious and expensive due to the necessary manual effort and the domain expert involvement.

Active learning (AL) was introduced to reduce the annotation costs across all supervised machine learning approaches (Settles, 2012) by selectively labelling informative instances and is a well-motivated and promising solution to address the problem of creating costly annotated datasets for training classifiers.

Active learning algorithms are advantageous in machine learning tasks where plenty of unlabelled samples are available and easy to access, but labelled data is scarce and expensive to be prepared. Active learning models are built in an iterative process, unlike other supervised machine learning models. As shown in Figure 1, a first model is built using a supervised algorithm on an initial labelled set, which represents a small portion of the whole annotated data (less than 1%). Then, in an iterative process, "informative" instances are selected using a query strategy, removed from the unlabelled set and added to the training set to build a new model using the supervised algorithm. The process continues until a stopping point which depends on the task (e.g., reaching at least the same effectiveness as supervised approach). By building a model on informative instances rather than the other instances, the active learning approach guarantees that the highest effectiveness can be yielded by the model.

There are some important elements in the active learning process:

(1) When active learning is performed in real situations, a human annotator labels each selected informative instances just after they have been selected. In this paper, instead, we simulate this activity and use the gold standard annotation to label the selected instances.

(2) In each iteration of *standard* active learning, a model is built from scratch, i.e., without considering the model from the previous iteration. However, at each iteration, it is also possible to build a model by updating the model from the previous iteration in an *incremental* active learning setting.

The main challenge of active learning approaches is to identify informative instances, and therefore it becomes essential to determine which selection criterion (also called query strategy) is the most suitable for a given task. A number of query strategies have been proposed, e.g., uncertainty sampling (Lewis & Catlett, 1994), query-by-committee (Seung et al., 1992), and information density (Settles & Craven, 2008). Uncertainty sampling (Lewis & Catlett, 1994) is currently the most widely used query strategy across active learning tasks and thus in this paper we will consider only uncertainty sampling to select data instances to label.

The goal of active learning is to maximize the effectiveness of the supervised machine learning model by minimizing the annotation effort. All supervised learning algorithms rely on a number of parameters that are typically tuned on a portion of the existing large set of annotated data (e.g., using cross validation). However, in an active learning framework there is less flexibility to build such tuned set of parameters, therefore the selected supervised algorithm needs to be as robust as possible to small changes in the training set and in its parameters, so that it can be used reliably in the active learning process. In particular, previous work (Nguyen & Patrick, 2014) has observed that some AL framework generated large variation in effectiveness of the models built during successive iterations, rendering the choice of a stopping point difficult. It is therefore essential to identify the parameters of the supervised learning model (here CRFs), the feature set used to train the model and the parameters of the AL framework (here standard vs. incremental) in order to establish what values are the most likely to lead to reliable and stable models. Settles and Craven (2008) have studied the effect of different AL query strategies on the effectiveness of concept extraction from text. They have demonstrated that uncertainty sampling methods (least confidence (Lewis & Catlett, 1994), margin (Scheffer et al., 2001), and entropy (Shannon, 1948)) are computationally the most efficient and, among the tested uncertainty sampling methods, least confidence and sequence entropy achieved better effectiveness compared to others. However, the factors that affect the stability and robustness of the AL models have not yet been investigated. Additionally, there has been no study to measure the impact of the Gaussian prior variance, one parameter of the CRFs model, on the robustness of the classifier.

In this paper we address the following questions: to what extent the AL models are reliable and robust? What factors affect the stability and robustness of the AL models? How different feature sets and parameter values of CRFs influence the robustness of the AL models? How incremental learning can help to build more reliable and robust models within the AL framework?

We answer these questions by conducting an intensive experimental evaluation on data from the i2b2/VA 2010 NLP challenge (Uzuner, et al., 2011) and the ShARe/CLEF 2013 eHealth Evaluation Lab (task 1) (Suominen, et al., 2013). The goal of these challenges is to extract concepts related to medical problems, tests and treatments and disorder mentions, respectively in i2b2/VA 2010 and ShARe/CLEF 2013. We rely on training data from these datasets to train active learning models and leverage the test data to evaluate the robustness of the built models.

The remainder of the paper is organized as follows: Section 2 introduces the feature set, the supervised CRFs approach and the active learning. Section 3 describes our experimental and evaluation settings. Results are reported in Section 4 and discussed in Section 5; Section 6 concludes the paper outlining directions of future investigation.

2 Active Learning Framework

In this section, we explain the features used to describe the data for classification. Then we briefly introduce CRFs and its parameter (Gaussian prior variance). Finally, we explain the query strategy and the incremental active learning settings.

2.1 Features for Conditional Random Fields

Figure 2 shows the feature categories that we use to inform the supervised learning algorithms in both supervised and active learning approaches.

The considered feature sets include rules implemented by regular expressions to identify acronyms, punctuations, capital letters and any combination of digits and letters; suffix and prefix characters with different length (up to 4); character 2-grams, 3-grams, and 4-grams; and a window of three previous and following words.

Engineered features are extracted with the Stanford Part-Of-Speech (POS) tagger (Toutanova et al., 2003) to produce POS tags as standard engineered features. Semantic features comprising of SNOMED CT and UMLS semantic groups as advanced engineered features are obtained using the Medtex system (a medical NLP toolkit) (Nguyen et al., 2009). Here we map semantic group features to "*1*" and "*0*" (present and absent, respectively) for each token. To this aim, we first differentiate our target semantic types from all UMLS and SNOMED CT semantic groups. Target semantic types are specified based on the target concept types required to be extracted (*problem*, *test*, and *treatment* in the i2b2/VA 2010 dataset and *disorder* in ShARe/CLEF 2013). For example, the following UMLS semantic groups represent the disorder concepts: *Congenital Abnormality, Acquired Abnormality, Injury or Poisoning, Pathologic function, Disease or Syndrome, and Mental or Behavioural Dysfunction, Cell or Molecular Dysfunction, Experimental Model of Disease, Anatomical Abnormality, Neoplastic Process, Sign and Symptoms* (Pradhan et al., 2013). We then assign "*1*" to target semantic types and "*0*" to non-target semantic types.

2.2 CRFs and Gaussian Prior Variance

The concept extraction problem requires to assign a sequence of labels $\vec{y} = (y_1, \dots, y_n)$ to a sequence of input tokens $\vec{x} = (x_1, \dots, x_n)$.

Conditional random fields is a probabilistic method for extracting and labelling sequential data. CRFs naturally encode dependencies between different entities of a sequence and typically outperform other supervised learning algorithms (e.g., support vector machines (Joachims, 1998)) in sequence labelling tasks (Li et al., 2008). In this paper we use a first-order linear-chain CRFs as supervised learning algorithm within the active learning framework.

Conditional random fields models measure the conditional probability of the outputs ($\vec{y}$) based on the given inputs ($\vec{x}$) with a set of parameters θ:

$$P_\theta(\vec{y}|\vec{x}) = \frac{1}{Z_\theta(\vec{x})} exp\left(\sum_{i=1}^{n}\sum_{j=1}^{m} \lambda_j f_j(y_{i-1}, y_i, x_i)\right) \quad (1)$$

where $Z_\theta(\vec{x})$ is the normalization factor, $f_j(.)$ are feature functions, and $\theta = (\lambda_1, \dots, \lambda_m)$ represent the parameters to weight the corresponding features. Each $f_j(.)$ is the transition feature function between label state $i-1$ and i on the sequence $\vec{x}$ at position i.

The model parameters θ are estimated by penalized maximum log-likelihood L on some training data T (Tomanek & Hahn, 2009):

$$L(T) = \sum_{(\vec{x},\vec{y})\in T} log\, p(\vec{y}|\vec{x}) - \sum_{i=1}^{m} \frac{\lambda_i^2}{2\sigma^2} \quad (2)$$

Regularization is used to penalize weight vectors with large norm. The regularization parameter ($\frac{1}{2\sigma^2}$) specifies the intensity of the penalty. If θ is modelled using a Gaussian prior, the regularization can be seen as a maximum a posteriori estimation of θ (Lafferty, et al., 2001).

Gaussian prior variance is an important parameter in CRFs, because it prevents over-fitting thus allowing to build reliable and robust models. In particular, the Gaussian prior variance specifies the variance of the feature weights: when the Gaussian prior variance is large, the feature weights deviate more from zero. If the Gaussian prior variance is set to infinite, then the feature weights can assume any real value. The latter case occurs when the values of the feature weights of the learnt model are not constrained by a limit; this results in over-fitting. A generalizable model should then have small feature weights values.

In our experiments, we investigate the effect of the Gaussian prior variance on the robustness and the effectiveness of the learnt AL models.

2.3 Incremental Active Learning and Query Strategy

As shown in Figure 1, in a standard active learning framework, where a pool of unlabelled instances is available, first a supervised model (Θ) is built on an initial small, randomly selected labelled set. Then a batch of informative instances (B) is selected using the query strategy $\varphi^\Theta(u_i)$. The query strategy estimates the informativeness of an unlabelled instance $u_i \in \mathcal{U}$ based on the model Θ. The selected batch of instances is removed from the unlabelled set and added to the labelled set to train a new model "from scratch", i.e., without considering the parameters from the previous model. This process continues until a stopping criterion is satisfied.

In an incremental setting, all parameter values, including the feature weights, are kept to be updated in a new iteration. This significantly reduces the training time, because a model does not have to be trained from scratch at each iteration and the parameters are already initialized in the previous step (Figure 3).

In our experiment, we investigate the difference in stability and robustness of the standard and incremental active learning approaches.

2.3.1 Query Strategy

At each iteration of the active learning loop, we use uncertainty sampling to query the unlabelled instances and select the most informative instances. Informativeness is estimated according to how uncertain the model is about the label of the unlabelled instance (i.e., the classification uncertainty of the model). Instances with the highest uncertainty are selected for labelling and inclusion in the labelled set used for training in the following iteration.

We use Least Confidence (LC) as it is known as one of the most effective uncertainty sampling methods. LC uses the confidence of the latest model Θ with parameters θ in predicting the label $\vec{y}$ of a sequence $\vec{x}$ (Culotta & McCallum, 2005):

$$\varphi_{LC}^{\theta}(\vec{x}) = 1 - P_\theta(\vec{y}^*|\vec{x}) \quad (3)$$

The confidence of the CRFs model is estimated using the posterior probability described in Equation (1) and $\vec{y}^*$ is the most likely label sequence obtained using the Viterbi algorithm:

$$\vec{y}^* = arg\, max_{\vec{y}\in Y^n}\, exp\left(\sum_{i=1}^{n}\sum_{j=1}^{m} \lambda_j f_j(y_{i-1}, y_i, x_i)\right) \quad (4)$$

The algorithms describing the active learning framework for both the standard and the incremental settings are shown in Figure 3 using the least confidence query strategy.

3 Experimental Framework

The experimental framework we propose aims to provide a study of the factors that impact stability and robustness, as well as to investigate the effectiveness of the learnt active learning models. Our experimental framework consists of three consecutive steps, as shown in Figure 4:

1. Investigate the impact of different feature combinations on robustness and effectiveness of the learnt models, when considering the default parameter values for CRFs and the standard AL.
2. Investigate the impact of incremental AL vs. standard AL on robustness and effectiveness of the learnt models, leveraging the best feature combination and the default CRFs parameters.
3. Investigate the impact of the CRFs parameters on robustness and effectiveness of the learnt models when considering the feature set and AL setting that showed the highest effectiveness at steps 1 and 2.

3.1 Dataset

Our experiments leverage data and task definitions from the i2b2/VA 2010 NLP task and the ShARe/CLEF 2013 eHealth Evaluation Lab (task 1). We use the same split of train and test sets defined in the original datasets.

The i2b2/VA 2010 NLP task (Uzuner, et al., 2011) requires to extract medical problems, tests and treatments from clinical reports. The reports used in this task are a combination of discharge summaries and progress notes supplied by three different health providers. The training and testing sets include 349 and 477 reports, respectively. These reports are organized as a collection of phrases and sentences (each report file containing a phrase or sentence per line). After dividing the dataset into phrases and sentences, we obtain 30,673 and 45,025 sequences in the training and test set, respectively.

The ShARe/CLEF 2013 eHealth Evaluation Lab (task 1) (Suominen, et al., 2013) requires to extract and identify disorder mentions from clinical free-text notes. The dataset for this task consists of 200 training and 100 test documents, including discharge summaries, electrocardiogram, echocardiogram, and radiology reports from a U.S. intensive care unit. As for the i2b2/VA 2010 dataset, we divided each report from this dataset into sequences based on line breaks. Overall, this produced 2,742 and 2,325 sequences in the training and test set, respectively.

3.2 Evaluation Methodology

We use the MALLET toolkit (McCallum, 2002) to train CRFs classifiers. For AL, the initial labelled set is formed by randomly selecting 1% of the training data. The batch size is set to 200 sequences for i2b2/VA 2010 and 30 for ShARe/CLEF 2013 across all experiments, leading to a total of 153 and 91 batches[1], respectively.

Concept extraction effectiveness is measured by Precision, Recall, and F1-measure. Evaluation metrics are computed on test data using the multi-segmentation evaluator implemented in the MALLET toolkit, which considers segments that span across multiple tokens.

The robustness and stability of AL models are analysed by examining the learning curves of the AL approaches across batches. For each batch, learning curves plot the F1-measure achieved by the AL classifier trained with the data contained in the labelled set up to the considered batch.

To further analyse the robustness of AL models, we also perform 10-fold cross validation experiments on the training data. In these experiments the training set is split in ten random sets; for a given fold, nine are used as labelled train data and one as test data. The effectiveness of active learning in each batch is averaged across ten test such folds.

4 Results

Section 4.1 reports the impact of different feature sets on the robustness of the learnt active learning models. Section 4.2 examines how incremental active learning affects the robustness of the learnt models. Finally, Section 4.3 analyses the impact of the CRFs parameter (Gaussian prior variance) on robustness and effectiveness of the active learning models.

4.1 Effect of Feature Sets

We define the following feature sets to evaluate the effect of different features on supervised and active learning approaches:

- **O** : Only token itself as a feature;
- **A** : Observed features;
- **B** : Standard engineered features (POS tags);
- **C**: Advanced engineered features (SNOMED CT and UMLS semantic groups).

We first consider each **O**, **A**, **B**, and **C** as a separate feature set within the supervised and active learning approaches. Figure 5 demonstrates the learning curve for the active learning approach against the supervised learning effectiveness with different feature groups (**O**, **A**, **B**, and **C)** for both the i2b2/VA 2010 and ShARe/CLEF 2013 datasets. These experiments consider the standard active learning approach with the default parameter values for the MALLET CRFs. Table 1 reports the highest effectiveness achieved by the active learning approach across the batches using different groups of features (**O**, **A**, **B**, and **C**) for both datasets.

The results shown in Figure 5 and Table 1 show that token features, POS tags and semantic features, when used alone, provide similar effectiveness on the ShARe/CLEF 2013 dataset, while POS tags and semantic features are generally superior than tokens alone in the i2b2/VA 2010 dataset. However, all these feature sets provide inferior effectiveness when compared to the observed feature set (**A**). These results suggest that basic linguistic features (**A**) including orthographical, morphological and lexical, and contextual features, are generally more effective than other feature sets (**O**, **B**, and **C**) in both supervised and active learning settings.

[1] The choice of batch size was done with respect to the number of sentences in each dataset. While this is a parameter which may ultimately influence effectiveness and stability of the learnt model, we do not explore it in the paper and leave it for future work.

From Table 1 we can further observe that the best active learning setting adds improvements for each feature set with respect to the supervised approach, e.g. **O** in i2b2. Also, the best effectiveness of both supervised and active learning approaches is achieved when leveraging observed features (**A**) in learning process.

Figure 5 shows that AL effectiveness varies greatly across batches. Specifically, when the highest effectiveness is achieved, standard AL does not seem to guarantee that effectiveness to be maintained if more batches are included in the training data, i.e. there are substantial fluctuations (and thus instability) in active learning curve. A sudden decrease in effectiveness between a batch B_i and the subsequent batch B_j suggests that the model learnt on data from up to batch B_i is over-fitted to that labelled set. Hence, the learnt model is not reliable for selecting the informative instances that form the next batch B_j from the unlabelled data. On the other hand, the active learning curve produced by representing data using the observed feature set **A** is smoother than the learning curves observed for other feature sets. This observation shows that feature set **A** leads not only to better effectiveness (Table 1), but also to more robust active learning models that exhibit stability across batches (Figure 5 (I-b and II-b)).

Next, we explore which combination of features provides the highest effectiveness in the active learning settings.

Table 2 and Figure 6 show that combining all considered feature sets (**O**, **A**, **B**, and **C**) provides higher effectiveness than using the individual feature sets alone.

In addition, combining feature sets improves the stability across the active learning batches in both datasets. However, the shapes of the active learning curves suggest that there are other factors, along with the feature sets, that contribute to the robustness and stability of the learnt models.

Figure 7 reports the results obtained when using 10-fold cross validation on the training data for both datasets and when all feature sets are used. In this experiment, active learning effectiveness values are averaged across the testing folds.

The active learning curves reported in Figure 7 are similar to those in Figure 6. Thus, the cross-validation experiments confirm what suggested by the train-test experiments: when all feature sets are combined, the models built using AL are more robust than those built using only one feature set.

4.2 Effect of Incremental Learning

In this section, we aim to study the effect of standard learning vs. incremental learning on stability and effectiveness of the learnt models within the active learning framework. We leverage the combination of all features and we use the default value of the MALLET CRFs parameters. Incremental active learning is applied throughout the training set, i.e., the values of CRFs parameters are updated in each iteration of active learning. We call this new setting the Incremental Active Learning for Concept Extraction (InALCE) and compare it with the standard active learning framework (ALCE).

Figure 8 reports the F1-measure achieved by ALCE and InALCE compared to the F1-measure obtained by the supervised classifier. Incremental active learning achieves higher effectiveness compared to standard active learning with less training data (i.e. requiring less batches): we suggest that this is because in the incremental active learning approach, the parameters of the learnt CRFs are maintained and updated in the subsequent iteration. While, in standard active learning the CRFs model is built from scratch at each iteration. Incremental learning is less prone to sudden changes in the training data, in particular negative changes. Therefore, the learnt models using incremental active learning in the InALCE framework are more robust rather than the models built using standard active learning in ALCE, as suggested by the smoother learning curve of InALCE when compared to those generated by ALCE. It subsequently leads to more accurate selection of informative instances in each iteration of InALCE and stability across the batches.

4.3 Effect of Gaussian Prior Variance in CRFs setting

As described in Section 2.2, the Gaussian prior variance is an important CRFs parameter as its value influences the robustness of the learnt model. In this section, we investigate the impact of this parameter on the robustness of the AL models built within the incremental AL approach using all features sets (Section 4.1), as this setting provided the highest effectiveness.

The default value for the Gaussian prior variance in MALLET is 10. Smaller values for the Gaussian prior variance limit the deviation of the feature weights from zero: this often avoids over-fitting. However, if the Gaussian prior variance is zero, then it will force all weights to be zero. To explore the impact of this parameter on the stability of the learnt models, we perform an empirical evaluation of different Gaussian prior variance values between 1 and 10 (with a step of 2). The empirical results suggest that a Gaussian prior variance value of 1 leads to the highest effectiveness for both supervised and active learning. This parameter value also exhibit the smoother AL learning curve with respect to the other tested values, resulting in more robust AL models.

Figure 9 reports the effectiveness for both supervised and active learning in un-tuned (Gaussian prior variance set to the default value of 10) and tuned settings (Gaussian prior variance set to 1). As shown in the figure, incremental active learning with tuned parameter provides the highest effectiveness and the most robustness across batches.

5 Discussion

The results reported in Section 4 show that feature sets, incremental learning and CRFs parameters (specifically, the Gaussian prior variance) play an important role in the stability, robustness and effectiveness of the active learning models learnt across batches.

In this paper we showed that the observed feature set (**A**), which includes orthographical, lexical and morphological, and contextual features, significantly increases the effectiveness of both supervised and active learning classifiers. While POS tags and semantic feature lead to poor effectiveness and unreliable models when used individually, they are useful to augment the data representation and the highest effectiveness is achieved when combining all feature sets.

Our analysis also demonstrated that incremental active learning not only reduces the amount of training data required (also compared to standard AL), but also leads to more robust and more effective models compared to the standard setting.

Finally, we have shown that the Gaussian prior variance used in CRFs influences both the effectiveness and the stability of the active learning models. The empirical results have demonstrated that tuning this parameter increases the effectiveness of both supervised and active learning models, but it has a minor effect on the stability compared to the influence of feature sets and the incremental setting.

6 Conclusion and Future work

In this paper, we have established that the robustness and the effectiveness of the active learning models for medical concept extraction depend on: feature set, incremental learning setting, and tuning of the supervised classifier parameters. This was demonstrated by conducting a large empirical evaluation on two medical datasets, the i2b2/VA 2010 and the ShARe/CLEF 2013 (task1). The evaluation showed that basic linguistic and lexical features increase the stability and robustness of the learnt models compared to domain specific semantic features. We also studied the effect of incremental learning and the Gaussian prior variance (CRFs parameter), observing that they increase both the effectiveness and the stability of the learnt models on both datasets.

This work represents the first step in analysing the stability and robustness of the learnt active learning models: further work is required to examine the influence of the considered factors for other types of concept extraction tasks.

7 References

Culotta, A., & McCallum, A. (2005). Reducing labeling effort for structured prediction tasks. *Proceedings of the National Conference on Artificial Intelligence (AAAI)* (pp. 746–751): AAAI Press.

Gurulingappa, H. (2012). *Mining the medical and patent literature to support healthcare and pharmacovigilance* (Ph.D. dissertation). University of Bonn, Bonn, Germany.

Jiang, J. (2012). Information extraction from text. In *Mining Text Data* (pp. 11-41): Springer.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *Proceedings of the 10th European Conference on Machine Learning (ECML-98)* (pp. 137-142): Springer-Verlag.

Lafferty, J. D., McCallum, A., & Pereira, F. C. N. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)* (pp. 282-289). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Lewis, D. D., & Catlett, J. (1994). Heterogenous Uncertainty Sampling for Supervised Learning. *Proceedings of the 18th International Conference on Machine Learning* (pp. 148-156): Morgan Kaufmann.

Li, D., Kipper-Schuler, K., & Savova, G. (2008). Conditional random fields and support vector machines for disorder named entity recognition in clinical texts. *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing* (pp. 94-95). Stroudsburg, PA, USA: Association for Computational Linguistics.

McCallum, A. K. (2002). MALLET: A Machine Learning for Language Toolkit. Retrieved from http://mallet.cs.umass.edu

Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C., & Hurdle, J. F. (2008). Extracting information from textual documents in the electronic health record: a review of recent research. *IMIA Year Book of Medical Informatics, 47*(Suppl 1), 128-144.

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association, 18*(5), 544-551.

Nguyen, A. N., Lawley, M. J., Hansen, D. P., & Colquist, S. (2009). A simple pipeline application for identifying and negating SNOMED clinical terminology in free text. *Proceedings of the Health Informatics Conference (HIC)* (pp. 188-193): Health Informatics Society of Australia (HISA).

Nguyen, D. H. M., & Patrick, J. D. (2014). Supervised machine learning and active learning in classification of radiology reports. *Journal of the American Medical Informatics Association*.

Piskorski, J., & Yangarber, R. (2013). Information extraction: Past, present and future. In T. Poibeau, H. Saggion, J. Piskorski & R. Yangarber (Eds.), *Multi-source, Multilingual Information Extraction and Summarization* (pp. 23-49): Springer Berlin Heidelberg.

Pradhan, S., Elhadad, N., South, B., Martinez, D., Christensen, L., Vogel, A., Suominen, H., Chapman, W., & Savova, G. (2013). Task 1: ShARe/CLEF eHealth Evaluation Lab 2013. In *Online Working Notes of CLEF, CLEF 2013*.

Roberts, A. (2012). *Clinical information extraction: lowering the barrier* (Ph.D. dissertation). University of Sheffield, Sheffield, United Kingdom.

Scheffer, T., Decomain, C., & Wrobel, S. (2001). Active hidden markov models for information extraction. *Proceedings of the International Conference on Advances in Intelligent Data Analysis (CAIDA)* (pp. 309-318): Springer-Verlag.

Settles, B. (2012). *Active Learning* (Vol. 6): Morgan & Claypool Publishers.

Settles, B., & Craven, M. (2008). An analysis of active learning strategies for sequence labeling tasks. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 1070-1079): Association for Computational Linguistics.

Seung, H. S., Opper, M., & Sompolinsky, H. (1992). Query by committee. *Proceedings of the fifth annual workshop on Computational learning theory* (pp. 287-294). 130417: ACM.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*, 379–423, 623–656.

Suominen, H., Salanterä, S., Velupillai, S., Chapman, W., Savova, G., Elhadad, N., Pradhan, S., South, B., Mowery, D., Jones, G. F., Leveling, J., Kelly, L., Goeuriot, L., Martinez, D., & Zuccon, G. (2013). Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. In P. Forner, H. Müller, R. Paredes, P. Rosso & B. Stein (Eds.), *Information Access Evaluation. Multilinguality, Multimodality, and Visualization* (Vol. 8138, pp. 212-231): Springer Berlin Heidelberg.

Tomanek, K., & Hahn, U. (2009). Semi-supervised active learning for sequence labeling. *Proceedings of the 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* (pp. 1039-1047): Association for Computational Linguistics.

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology* (Vol. 1, pp. 173-180): Association for Computational Linguistics.

Uzuner, Ö., Goldstein, I., Luo, Y., & Kohane, I. (2008). Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association, 15*(1), 14-24.

Uzuner, Ö., Solti, I., & Cadag, E. (2010). Extracting medication information from clinical text. *Journal of the American Medical Informatics Association, 17*(5), 514-518.

Uzuner, Ö., South, B. R., Shen, S., & DuVall, S. L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association, 18*(5), 552-556.

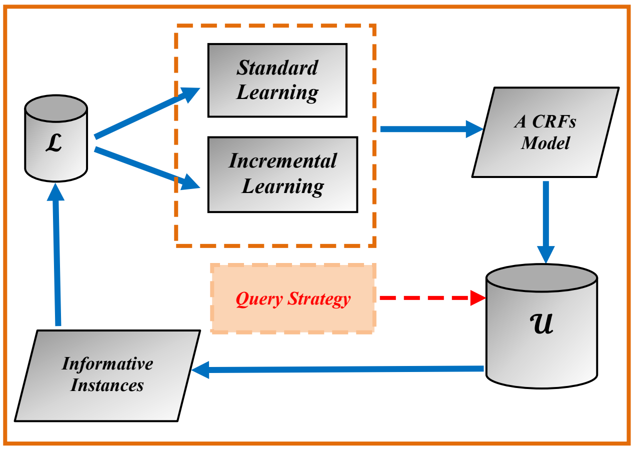


Figure 1: Standard vs. incremental active learning.

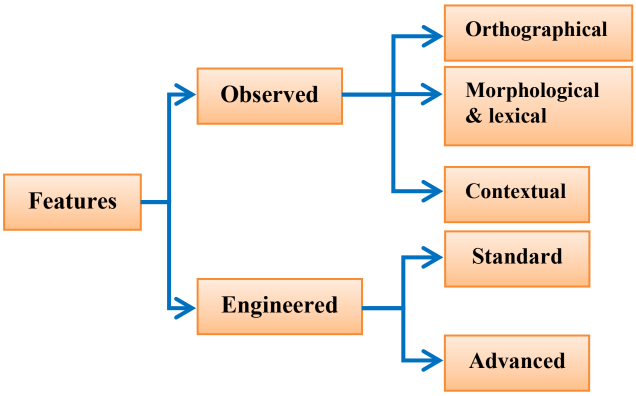


Figure 2: Feature groups.

```
Input:
Set of labeled instances $\mathcal{L}$, set of unlabeled instances $\mathcal{U}$, least confidence query strategy $\varphi_{LC}^{\Theta}(u_i)$, number of instances to be selected in each iteration (batch size) $B$.

Algorithm:
While stopping criterion not met
  1- train $\Theta$ model on $\mathcal{L}$
     (a) standard learning: train model $\Theta$ on $\mathcal{L}$ from scratch
        OR
     (b) incremental learning: update the model $\Theta$ from previous iteration with new $\mathcal{L}$
  2- for all $u_i \in \mathcal{U}$ do:
       $\varphi_{LC}^{\Theta}(u_i) = 1 - P_\theta(\vec{y}^*|\overrightarrow{u_i})$ (Equation 1 and 4)
     End
  3- for $b = 1$ to $B$ do:
       $u_b = \arg max_{u_i \in \mathcal{U}}\ \varphi_{LC}^{\Theta}(u_i)$
       $\mathcal{L} = \mathcal{L} \cup (u_b, label(u_b))$
       $\mathcal{U} = \mathcal{U} - u_b$
     End
End
```

Figure 3: The AL framework based on least confidence and incremental vs. standard learning.

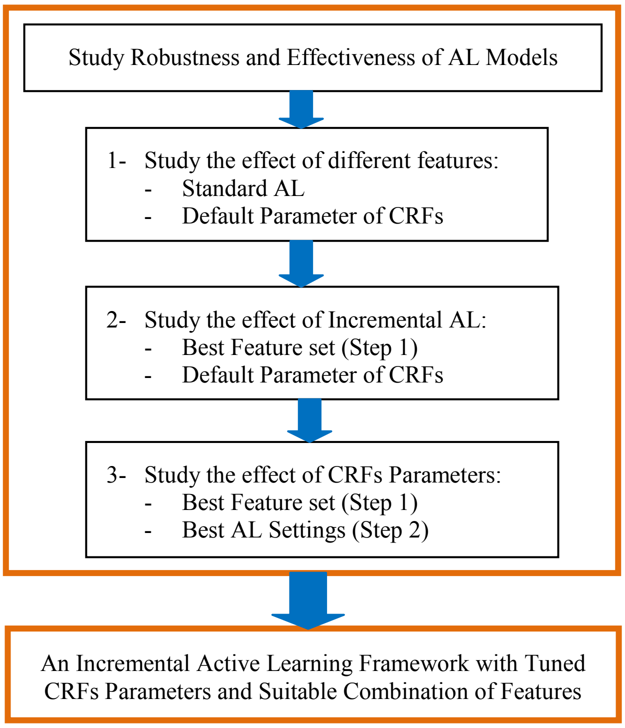

Study Robustness and Effectiveness of AL Models
1- Study the effect of different features:
- Standard AL
- Default Parameter of CRFs
2- Study the effect of Incremental AL:
- Best Feature set (Step 1)
- Default Parameter of CRFs
3- Study the effect of CRFs Parameters:
- Best Feature set (Step 1)
- Best AL Settings (Step 2)
An Incremental Active Learning Framework with Tuned CRFs Parameters and Suitable Combination of Features

Figure 4: Experimental framework.

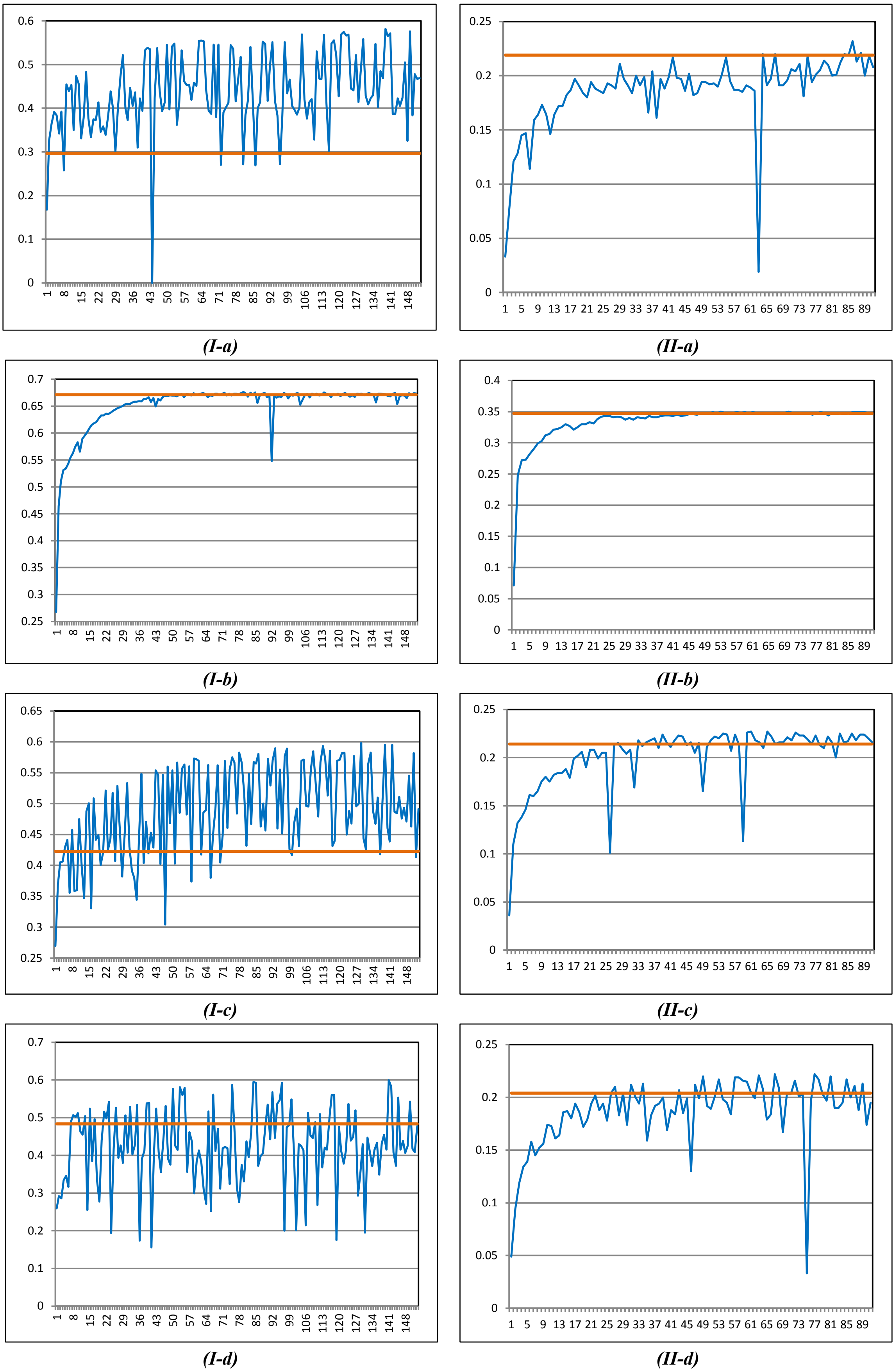
	Supervised Learning			Active Learning		
	P	**R**	**F1**	**P**	**R**	**F1**
O	0.252	0.36	0.296	0.625	0.544	0.581
A	0.725	0.624	**0.671**	0.728	0.63	**0.676**
B	0.416	0.429	0.428	0.64	0.561	0.598
C	0.501	0.466	0.483	0. 654	0.552	0.599

(a)

	Supervised Learning			Active Learning		
	P	**R**	**F1**	**P**	**R**	**F1**
O	0.266	0.186	0.219	0.334	0.178	0.232
A	0.419	0.296	**0.347**	0.424	0.298	**0.35**
B	0.257	0.184	0.214	0.324	0.174	0.227
C	0.233	0.182	0.204	0.304	0.174	0.222

(b)

Table 1: The effectiveness of the supervised and the active learning approach (the one exhibiting the highest effectiveness) with respect to O, A, B, and C feature sets (P = Precision, R = Recall, and F1 = F1-measure) (a) i2b2/VA 2010 (b) ShARe/CLEF 2013.



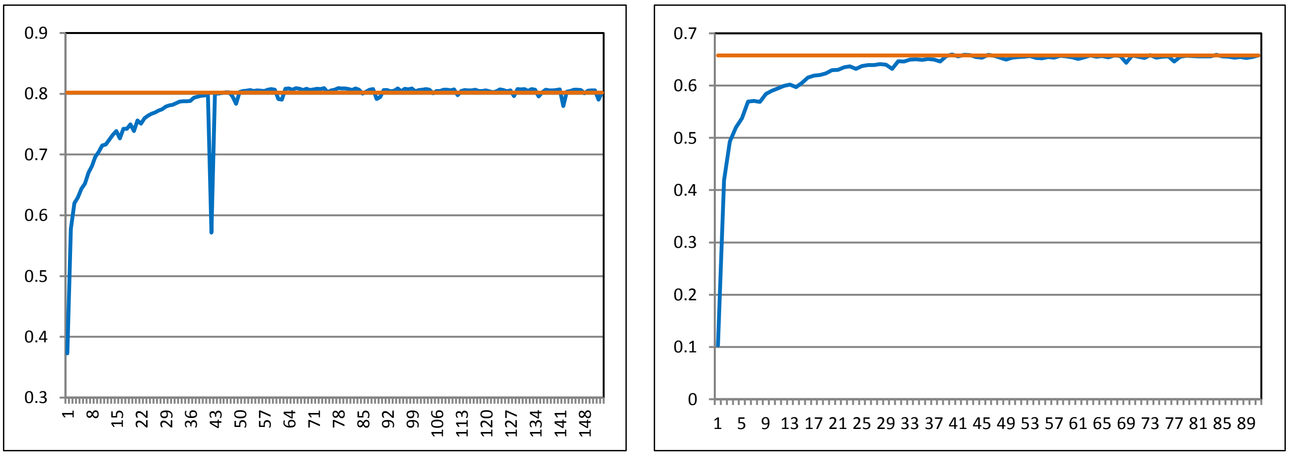
(I-a) *(II-a)*

(I-b) *(II-b)*

(I-c) *(II-c)*

(I-d) *(II-d)*

Figure 5: Active learning curves across batches (blue line) and supervised effectiveness (orange straight line). The horizontal axis reports the number of batches used to train the classifier in the AL setting, while the vertical axis reports the value of F1-measure obtained by applying the classifier from the corresponding batch (or the whole training data in the case of the supervised classifier) on the test data (I: i2b2/VA 2010 dataset, II: ShARe/CLEF 2013 dataset, a: O, b: A, c: B, d: C).



(a) **(b)**

Figure 6: Active learning curves across the batches (blue line) and supervised effectiveness (orange straight line), using the combination of whole features. The horizontal axis reports the number of batches used to train the classifier in the AL setting, while the vertical axis reports the value of F1-measure obtained by applying the classifier from the corresponding batch (or the whole training data in the case of the supervised classifier) on the test data (a) i2b2/VA 2010 (b) ShARe/CLEF 2013.

	Supervised Learning			Active Learning		
	P	**R**	**F1**	**P**	**R**	**F1**
All	0.816	0.788	0.802	0.824	0.795	0.809

(a)

	Supervised Learning			Active Learning		
	P	**R**	**F1**	**P**	**R**	**F1**
All	0.759	0.581	0.658	0.763	0.58	0.659

(b)

Table 2: The effectiveness of the supervised and the best active learning approach (the one exhibiting the highest effectiveness) using all features (A, B, and C) (P = Precision, R = Recall, and F1 = F1-measure) (a) i2b2/VA 2010 (b) ShARe/CLEF 2013.

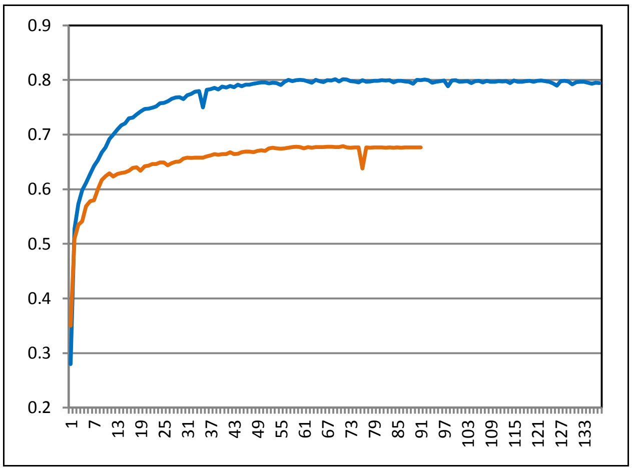
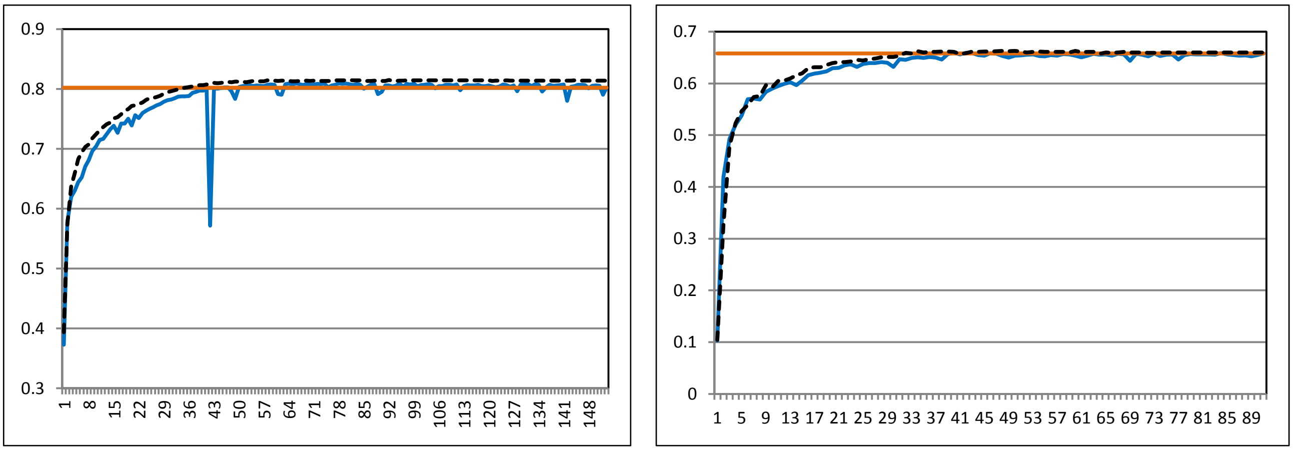
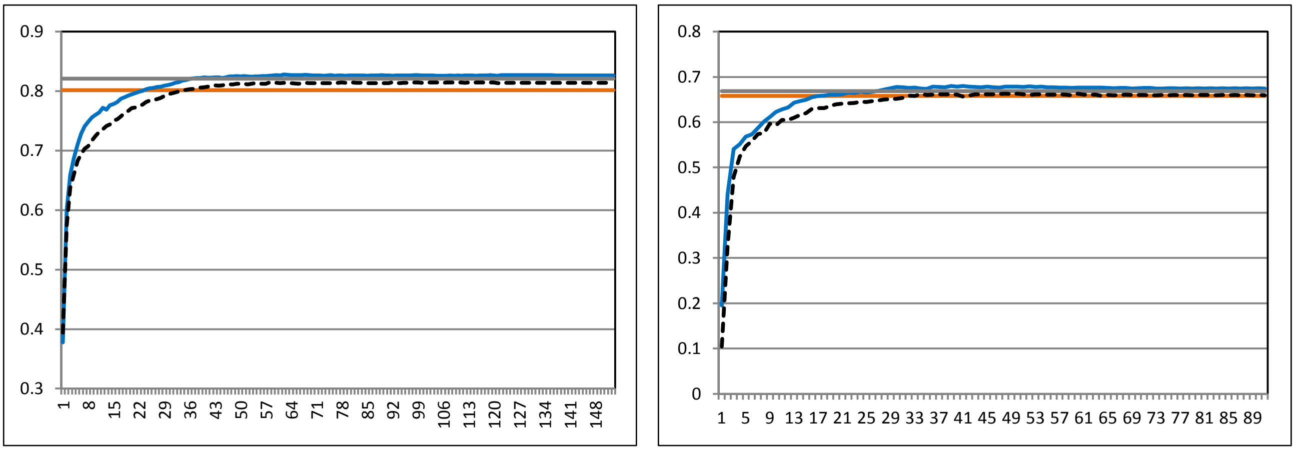


Figure 7: 10-fold cross validation results across the active learning batches on i2b2/VA 2010 (blue curve) and ShARe/CLEF 2013 (orange curve) datasets. The horizontal axis corresponds to the number of batches used for training and the vertical axis reports F1-measure values.



(a) **(b)**

Figure 8: Standard vs. Incremental active learning (ALCE vs. InALCE). The dashed black curve and blue curve represent the effectiveness of InALCE and ALCE, respectively, and the orange line represents the effectiveness of the supervised classifier. The horizontal axis reports the number of batches used to train the classifier in the AL setting, while the vertical axis reports the value of F1-measure obtained by applying the classifier from the corresponding batch (or the whole training data in the case of the supervised classifier) on the test data (a) i2b2/VA 2010 (b) ShARe/CLEF 2013.



(a) **(b)**

Figure 9: Incremental non-tuned (dash black curve) vs. Incremental (blue curve) active learning with tuned CRFs parameters (InALCE vs. InALCE-Tun) against the un-tuned (orange line) and tuned (grey line) supervised effectiveness. The horizontal axis reports the number of batches used to train the classifier in the AL setting, while the vertical axis reports the value of F1-measure obtained by applying the classifier from the corresponding batch (or the whole training data in the case of the supervised classifier) on the test data (a) i2b2/VA 2010 (b) ShARe/CLEF 2013.