Preliminary proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)
Distributed under a CC BY-NC-SA 4.0 license.

Visualising and evaluating the effects of combining active learning with word embedding features

Maria Skeppstedt[1], **Rafal Rzepka**[2,3], **Kenji Araki**[2], **Andreas Kerren**[4]
[1]The Language Council of Sweden, the Institute for Language and Folklore, Sweden
[2]Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Japan
[3]RIKEN Center for Advanced Intelligence Project (AIP), Tokyo, Japan
[4]Department of Computer Science and Media Technology, Linnaeus University, Växjö, Sweden
maria.skeppstedt@sprakochfolkminnen.se, andreas.kerren@lnu.se,
{rzepka,araki}@ist.hokudai.ac.jp

Abstract

A tool that enables the use of active learning, as well as the incorporation of word embeddings, was evaluated for its ability to decrease the training data set size required for a named entity recognition model. Uncertainty-based active learning and the use of word embeddings led to very large performance improvements on small data sets for the entity categories PERSON and LOCATION. In contrast, the embedding features used were shown to be unsuitable for detecting entities belonging to the ORGANISATION category. The tool was also extended with functionality for visualising the usefulness of the active learning process and of the word embeddings used. The visualisations provided were able to indicate the performance differences between the entities, as well as differences with regards to usefulness of the embedding features.

1 Introduction

To acquire large training data sets by the use of low-cost crowdsourcing is not a universal solution for all annotation tasks. The ethical aspect could be one concern, as the concept of low-cost crowd annotations implies low-paid annotators (Martin et al., 2017). Other obstacles might be data privacy restrictions (e.g., when annotating clinical health records), or a lack of specialised competence among crowd workers, e.g., competence in the annotation task or in a specific language. Strategies for facilitating annotation are therefore important, also in the age of crowdsourcing.

A possible strategy for facilitating annotation is to minimise the amount of manually annotated data required, e.g., data required for the task of training a machine learning model. This could be achieved by (i) using active learning to actively select training samples useful to the model and (ii) training the model on information that has been derived in an unsupervised fashion. There is a large body of research that has shown the effectiveness of using each one of these strategies individually, and there are also annotation tools/annotation tool extensions that incorporate these two strategies (Skeppstedt et al., 2016; Kucher et al., 2017). However, to the best of our knowledge, there are no studies that evaluate the effectiveness of this combined data reducing strategy provided by the tools. The first aim of this study is therefore to evaluate the effectiveness of one such tool, i.e., to evaluate whether using the tool leads to the expected decrease in data size required to train a machine learning model.

Also the annotation of a smaller data set can however be a time-consuming, and potentially boring, task. Gamification of the task is one previously explored strategy for solving this problem (Dumitrache et al., 2013; Venhuizen et al., 2013).

Another potential strategy for increasing the intrinsic motivation for the annotation task, is to make the annotator aware of the usefulness of the data that is being annotated. The second aim of this study is to take a first step towards exploring this strategy in the context of an active learning process. We aim to provide a suggestion for a visualisation of how the increasingly larger training data set, which results from the manual annotation effort, changes the model that is trained on this annotated data set. That is, a visualisation that has the potential to increase the human understanding of the active learning-based annotation process.

2 Background

The tool whose performance we have evaluated, and whose active learning process we have visualised, is the tool "PAL – a tool for Pre-annotation and Active Learning" (Skeppstedt et al., 2016). PAL is meant to be used as an extension to another annotation tool, e.g., BRAT (Stenetorp et al., 2012), for annotating data to be used for training a named entity recognition (NER) model. While high performance is often reported for the NER task, e.g., for newswire texts (Sang and Meulder, 2003), the task is more difficult for noisy texts and when small training data sets are used. For instance, the best system on the ACL 2015 Workshop on Noisy User-generated Text achieved an F-score of 0.74 for PERSON, 0.50 for COMPANY, and 0.66 for GEO-LOCATION when using a training set of 2,950 tweets (Baldwin et al., 2015; Yamada et al., 2015).

PAL provides functionality for active data selection, as well as for incorporating unsupervised data in the form of word embeddings when training the models that are used for active data selection. The tool also offers annotation support in the form of pre-annotations. This is achieved by repeatedly retraining a NER model on the data that the annotator produces in BRAT and on information incorporated from word embeddings. The trained model can then be used for two purposes: (i) to actively import new annotation data into BRAT, i.e., to actively select data useful for improving the model, and (ii) to simplify the annotation by providing the annotator with pre-annotations in BRAT format. To allow the annotator to add, delete or change the span length of pre-annotated entities — instead of annotating from scratch — has been shown to reduce annotation time (Lingren et al., 2014).

PAL could, for instance, be used according to the annotation process suggested by Olsson (2008). That is, to first annotate an actively selected subset of a corpus to achieve a model that can perform pre-annotations with acceptable accuracy, and thereafter use this model for providing the annotator with pre-annotations when a larger corpus is annotated. Such a corpus might, for instance, be used for training a model that requires a large training data set to perform well. The current study focuses on the first part of such a use case, that is on the process of actively selecting training samples to achieve a model that recognises named entities with acceptable performance.

2.1 Approaches for minimising training data

To use active learning, instead of a random sampling of training data, has led to a reduction of the number of samples needed to train classifiers to recognise different entity types (Shen et al., 2004; Tomanek et al., 2007). The technique builds on the following idea: Data samples estimated to be useful to a machine learning model are actively selected from a pool of unlabelled data. The selected samples are presented to an annotator for manual annotation, and the newly annotated data is then added to the set of labelled data that is available for training the model. This expanded training data set is then used to retrain the model, which in turn is applied in the next iteration in the process of actively selecting data. The estimate of a sample's usefulness can, for instance, be based on the level of disagreement among different classifiers (Olsson, 2008, pp. 25–29), or on properties specific to the type of model used, e.g., a confidence measure provided by the model (Settles, 2009).

The other technique included in PAL for reducing the training data size is to incorporate features gathered in an unsupervised fashion, through the use of text distributional properties of word types. There is a large body of research that shows this technique to be effective for named entity recognition, e.g., the use of features in the form of Brown clusters (Miller et al., 2004) and more recently in the form of different types of word vectors automatically derived from large corpora (Sahlgren, 2006; Mikolov et al., 2013). Word vectors have for instance been incorporated in the feature set when using conditional random fields classifiers (Turian et al., 2010; Guo et al., 2014; Henriksson, 2015; Copara et al., 2016), or used as input to different types of neural network-based classifiers (Godin et al., 2015; dos Santos and Guimarães, 2015; Yang et al., 2016; Lample et al., 2016; Reimers and Gurevych, 2017). There is, however, less research that investigates the effects of using the two strategies of unsupervised features and active learning in tandem; in particular their effects on small data sets, i.e., the use case that we explore here.

2.2 Functionality of PAL

Each iteration in PAL is run in two steps. First, data positioned in PAL's "folder for labelled data" is used for training a machine learning model; a model which is then used for selecting new data samples from PAL's "folder for unlabelled data." The model also provides BRAT-format pre-annotations for the selected data, enabling it to be directly imported into BRAT (Figure 3b). In the second step, which takes place after the data has been manually annotated, the data annotated in BRAT is moved into PAL's "folder for labelled data", to enable the next active learning iteration.

A basic feature vector for training the model, x_n, is constructed through representing each token by a concatenation of (i) the one-hot encoding for the token with (ii) the one-hot encoding for a configurable number of neighbours to the token.

The functionality of incorporating features derived in an unsupervised fashion is provided in PAL through an extension of the basic vector by a vector derived from pre-trained word embeddings. This is achieved by concatenating the basic feature vector with the word embedding vector that represents the token, as well as with the word embedding vectors that represent the neighbours of the token.

Information from gazetteers or information on which words were capitalised were not included in the feature set, to focus the experiment on the effects of the different strategies compared. This also makes the results somewhat more generalisable, e.g., to entity types that are not typically capitalised or for which gazetteers do not typically exist, or to languages that do not use an initial capital letter as a signal for names.

With the focus on making the data selection and model training process as comprehensible as possible for a human, we used the main classification method included in PAL, which is a token-level logistic regression classifier. That is, a classifier for which a human-interpretable confidence measure can be returned for each token in the pool of unlabelled data. The output of this unstructured predictor, is then post-processed into B/I-labels for tokens classified as an entity.

The confidence is then used for carrying out *uncertainty sampling* from the pool of unlabelled data (Settles, 2009). More specifically, the measure used is the difference in certainty level between the two most probable classifications for each of the tokens in the data pool. Given c_{p1} as the most probable classification and c_{p2} as the second most probable classification for the observation x_n, the uncertainty measure would be:

$$M_n = P(c_{p1}|x_n) - P(c_{p2}|x_n) \tag{1}$$

The smaller M_n, the higher is the uncertainty of the classifier and the higher is the sample ranked in the active selection process (Schein and Ungar, 2007).

PAL represents each training sample by the lowest M among the tokens it includes. For each iteration in the active selection process, samples that contain tokens with the lowest M-values are thereby selected. To achieve a variation among the samples selected, PAL also imposes the constraint of not allowing the selected texts to include the same word twice, if this word is predicted by the model to be included in a named entity.

PAL accesses embeddings through Gensim (Řehůřek and Sojka, 2010) and uses Scikit-learn's (Pedregosa et al., 2011) logistic regression classifier with a regularisation strength determined through cross-fold validation.

3 Method

The evaluation of PAL was carried out using the Broad Twitter Corpus (Derczynski et al., 2016), which consists of English tweets annotated for the three entities PERSON, LOCATION, and ORGANISATION. The corpus is sampled across different regions, temporal periods, and from different types of Twitter users, to ensure a large diversity of the entities included. Each of the three entity types was annotated separately.

We removed metadata in the form of hashtags and usernames starting with @, to make the task more similar to most previous NER tasks, where entities are mentioned in a textual context. The corpus is divided into six segments, each of them with a different signifying property, e.g., tweets from popular individuals, tweets from mainstream news, or tweets focused on one specific event. For performing the experiments we, however, sampled randomly from the corpus (as described below), without taking this structure into account.

3.1 Simulation of active learning

The active learning process in PAL was used in simulated mode as follows: the machine learning model was first trained on a small labelled data set consisting of 200 randomly selected tweets, i.e., a set representing an initial seed set. The task of the active learning algorithm would then be to select the most informative data points from the pool of unlabelled data. In the experiment, the "pool of unlabelled data" was simulated by the texts from the pre-labelled tweets in the Broad Twitter Corpus, and the corpus labels were used to simulate input in the form of manual annotations performed by the annotator.

For the experiment performed, we selected 20 tweets in each iteration. These 20 tweets and their corresponding labels were thus added to the set of labelled data, to simulate the process of them being manually annotated. The model was, thereafter, retrained, and a new iteration in the process of actively selecting tweets was then carried out, until the set of labelled data contained 1,000 tweets.

A context window of the two most immediate neighbours was used, with a frequency cut-off of three occurrences for a neighbour to be included. Word embeddings from a word2vec skip-gram model, which had been pre-trained by Godin et al. (2015) on 400 million tweets, were used as unsupervised features.

3.2 Evaluating the active learning simulation

The strategies used in PAL for decreasing the training data size required were compared to a baseline strategy. A total of four different strategies were thus evaluated for their performance on a small training data set: (i) the baseline, with *random* data selection and a *basic feature* vector, (ii) data selection through *active learning* and the *basic feature* vector, (iii) *random* data selection and the feature vector extended with *word2vec features*, and finally (iv) data selection through *active learning* and the feature vector extended with *word2vec features*.

4,000 tweets were randomly selected from the Broad Twitter Corpus to simulate the pool of unlabelled data, and 2,000 other tweets were randomly selected to be used as evaluation data. From the simulated pool of data were then 200 tweets randomly selected to form the seed set.

Starting with this seed set, an evaluation was carried out of the four different strategies investigated. For one of the active learning strategies, the basic feature vector was used, and for the other, the word2vec extension. For every step in the iteration, the performance of the model was evaluated against the 2,000 tweets that formed the evaluation data, i.e., after 20 new training data samples had been actively added to the training data set.

For the two strategies that did not include active learning, each iteration instead consisted of a random selection of 20 new tweets from the simulated data pool. A new model was trained on data including these newly selected tweets, and then evaluated against the 2,000 tweets in the evaluation set. The same randomly selected data sets were used both for the setting with word2vec features and the setting without these features.

As results of the study were heavily dependent on the random selection of a number of small data sets, it was particularly important to make sure that results achieved were not due to chance. The entire experiment was therefore repeated 10 times, each time with a new random selection of data pool, evaluation and seed set, as well as training data for the strategies not using active data selection. A separate experiment was carried out for each one of the three entity types LOCATION, ORGANISATION and PERSON, i.e., matching the manner in which the evaluation corpus had been annotated. Entities were represented by the BIO-encoding, and the classifications were evaluated using the CoNLL 2000 NER script (Tjong Kim Sang and Buchholz, 2000).

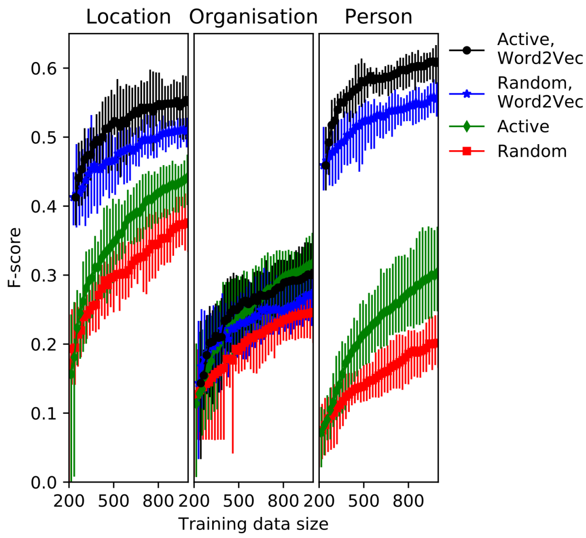

Figure 1: Average **F-score** for the ten experiment reruns. The error bars show the interval between the minimum and maximum of the F-scores measured, and the x-axes show the number of training samples.

3.3 Visualising the active learning process

We extended PAL by enabling it to record statistics for the pool of unlabelled data for each iteration of active data selection. We also extended the tool by adding a command which allows the user to generate a visualisation of this recorded data. The visualisation aimed to increase the human understanding of the active learning and classifier training by (i) showing why a particular set of samples are chosen for manual annotation in each iteration, (ii) showing an indication of the usefulness of the embedding features used, through visualising how clusters formed by the embeddings correspond to the entity categories investigated, and (iii) showing how the classification uncertainty for the pool of unlabelled data changes when more data is annotated and used for training the model.

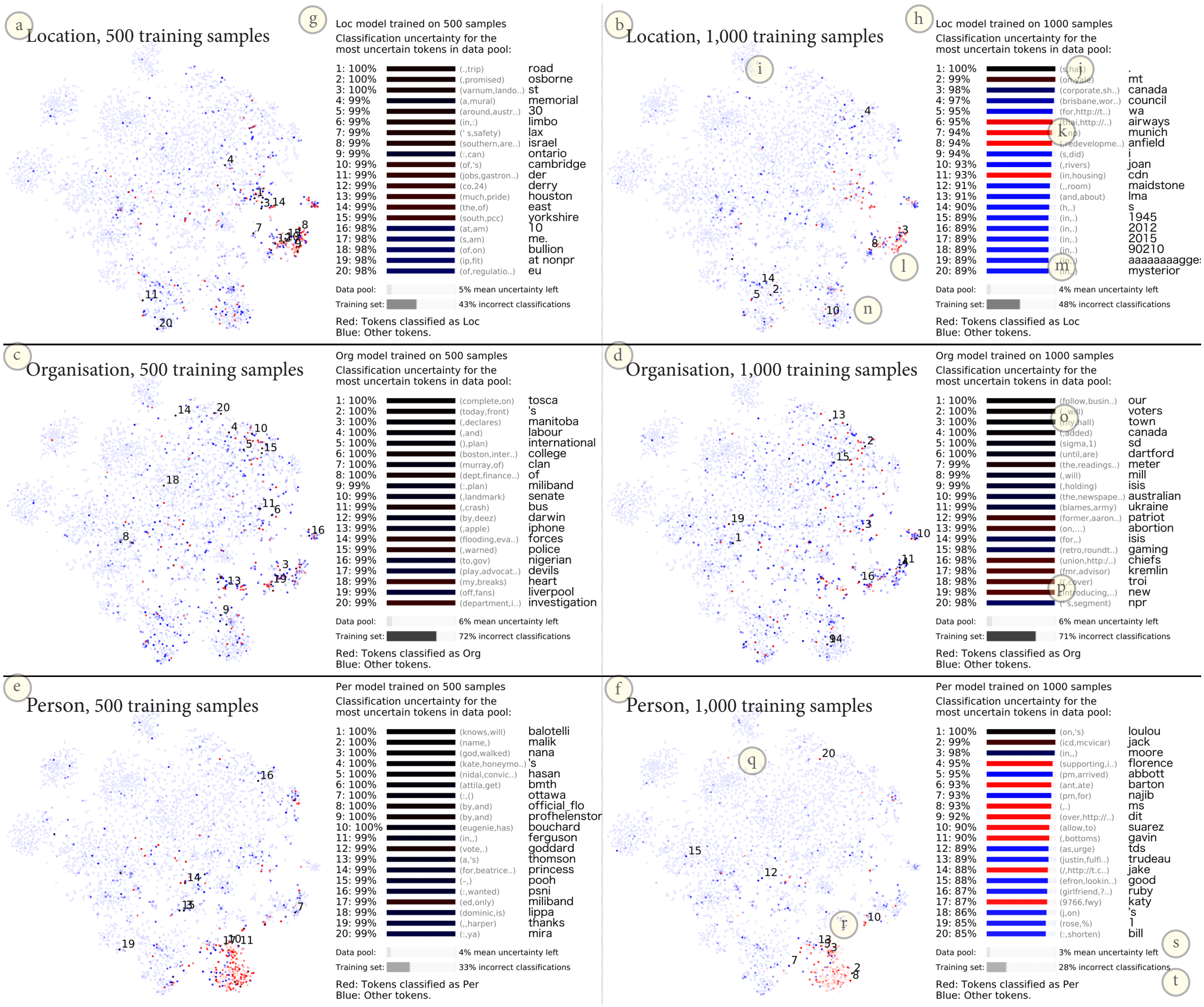

Figure 2: **(a-f)** Six subplots, two for each of the three entity categories. **(g)** The left-hand column: The model's uncertainty for classifying tokens in the pool of unlabelled data when *500* samples have been removed from the pool, labelled, and then used as training data for the model. **(h)** The right-hand column: Same as *g*, but with a training data size of *1,000* samples. **(i)** A t-SNE plot is displayed to the left in each of the six subplots, showing word embeddings that correspond to words included in the pool of unlabelled data. Words that occur in similar contexts are positioned close to each other in the plot. **(j)** The 20 most uncertain tokens in the pool of unlabelled data, together with a bar chart showing their level of uncertainty, is displayed to the right in each subplot. That is, the 20 tokens for which the machine learning model, trained on the set of labelled data available, is most uncertain. (The two closest neighbouring tokens are shown in parenthesis.) **(k-l)** The colour *red* is used for signifying that a token has been classified by the model as belonging to the entity category in question (i.e., classified as a LOCATION, ORGANISATION or PERSON entity). **(m-n)** The colour *blue* is used for signifying that a token is *not* classified as belonging to the entity category in question. **(o-p)** The t-SNE plot and the bar chart use the same colour-coding for signifying the output of the machine learning model. The larger the uncertainty with which a token is classified by the model, the darker (i.e., closer to black) is the red or blue in which it is displayed. **(q)** In contrast, tokens that the model classifies with a low uncertainty are displayed in a bright colour with low saturation. **(r)** The numbers can be used for locating the position in the t-SNE plot for those among the most uncertain tokens that occurred at least twice in the pool of unlabelled data. **(s)** Bar chart indicating mean model uncertainty for all words left in the pool of unlabelled data. **(t)** Bar chart indicating the proportion of incorrectly classified tokens when conducting cross-fold validation on the training set.

The advantage of applying the functionality in PAL that uses a token-level, logistic regression classifier for the data selection, and that selects samples based on their most uncertain token, is that the selection process is easily explainable. That is, the first of the visualisation goals can be met by conveying a list of these tokens, for which the model was most uncertain, together with the model's classification uncertainty for these tokens.

The second visualisation goal can be met by plotting a t-distributed stochastic neighbour embedding plot, t-SNE (van der Maaten and Hinton, 2008), of the word embeddings that were used as features. Plotted word embeddings can then be colour-coded according to how the word which they represent most often is classified. Thereby, a comparison between classifications by the trained model and clusters of word embeddings, as shown by the t-SNE plot, can be carried out.

To show the classification uncertainty of the most uncertain tokens also helps meeting the third visualisation goal. That is, changes in uncertainty for these most uncertain tokens indicate changes in model performance when the training data size increases. In addition, the colour-coding of the t-SNE plot can also be used for indicating whether the classification uncertainty for the tokens in the pool of unlabelled data changes when more data is labelled and used as training data.

3.4 Visualisations for another corpus

To verify that the visualisation also functions on another corpus than the English corpus that we used during development and for simulation of the process, we performed a small annotation experiment on a corpus of Japanese microblogs.[1]

As white space is not normally used in Japanese text, we first performed a pre-processing using the MeCab tool (Kudo, 2006). That is, the text segments generated by MeCab was used, and white space was inserted between these segments. Thereby, the white space-based tokenisation included in Scikit-learn could be used as-is. As unsupervised features, we used word embedding vectors from a word2vec model that had been trained on Japanese texts, which had been segmented by MeCab and merged with the help of a dictionary[2].

For this corpus, we did not perform a simulation, but instead applied PAL for the authentic use case of annotating raw text data. That is, we used the facilities of active learning and pre-annotation that are available in PAL for annotating text, and generated a visualisation after each iteration. We imported the pre-annotations generated by PAL into the BRAT annotation tool, as shown in Figure 3, to modify or delete incorrect annotations and to add omitted ones. We used annotation guidelines for entity detection and tracking (EDT)[3].

4 Results

Evaluation results in the form of an F-score measurement when evaluating against an external evaluation set are shown in Figure 1, while Figures 2 and 3 show the output of the proposed visualisations for the active learning process.

4.1 Evaluation results

The main lines in Figure 1 show the average F-scores for the ten re-runs for each training data size included in the experiment. The error bars show the minimum and maximum F-scores for the ten re-runs, i.e., giving an indication of the variation in the results achieved.

For the entity categories LOCATION and PERSON, average F-scores for the four different strategies produce four well-separated lines. Results are often separated, or close to separated, also when taking the lowest/highest value measured for the ten folds into account. Active data selection gives better results than random selection, and incorporating unsupervised features gives better results than not using them. The incorporation of unsupervised features is a more useful strategy than active data selection, and, more importantly, combining the two strategies is the overall most useful method.

Figure 1 further shows that while active learning was useful also for the category ORGANISATION, the use of word embeddings instead had a small negative impact on this category for a data set containing more than 600 samples.

4.2 Visualisation output

The visualisation functionality, with which we extended the PAL tool in this study, provides one visualisation of the unlabelled data pool for each iteration in the active learning process. The left-hand column in Figure 2 shows three visualisations, one for each of the three entity categories investigated. Each of them was generated in an active learning iteration when the training data set contained 500 samples. The right-hand column in the

[1] http://www.cs.cmu.edu/~lingwang/microtopia/#twittergold Microblogs collected with the criterium that they should contain the same content written in Japanese and in English (Ling et al., 2014), from which we used the Japanese parts.

[2] https://github.com/shiroyagicorp/japanese-word2vec-model-builder

[3] www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-edt-v4.2.6.pdf

figure shows visualisations for the three categories, that instead were generated when the training data set contained 1,000 data samples. All six subplots visualise the state of the pool when using active learning and the word2vec features.

Each subplot shows the state of the pool of unlabelled data. That is, each subplot contains an uncertainty colour-coded t-SNE visualisation of word embeddings that correspond to tokens present in the data pool, as well as a bar chart displaying the classification uncertainty for the 20 most uncertain tokens in the pool. Red colours in the t-SNE plot and the bar chart signify tokens that the model, trained on the currently available labelled data, classifies as belonging to the entity category in question, whereas blue colours indicate that this model classifies the token as outside of an entity. Darker colours in the t-SNE plot and the bar chart signify higher uncertainty for the classification.

In particular, the colours and lengths of the bars for PERSON and LOCATION show that there is a higher uncertainty for a model trained on 500 data samples than for a model trained on 1,000 samples. Also the colour coding of the t-SNE plot gives a slight indication of this difference in uncertainty. In contrast, for the ORGANISATION entity, there is a large uncertainty also for a training data set containing 1,000 samples. The bars that indicate mean uncertainty left in the data pool corroborate this difference.

The visualised differences in model uncertainty for different entities correspond to differences found in the evaluations against the gold standard, as shown in Figure 1. That is, the model trained to recognise ORGANISATION, which is visualised as uncertain, still yields a very low F-score when trained on 1,000 training samples. Similarly, that better results were achieved for PERSON and LOCATION when evaluating against the gold standard, is reflected by a visualisation that indicates a lower uncertainty for models trained on 1,000 training samples to detect these entity categories.

Conversely, the percentage of incorrect classifications increases when the training data set for the entity LOCATION increases. Thereby, the standard measurement, in the form of incorrect classifications when performing a cross-validation on the labelled data, fails to indicate changes in model performance.[4]

[4]This measure is equivalent to inverse accuracy. *Inverse* accuracy is used to match the uncertainty measure used, i.e., the aim for both should be to reach 0%.

The spatial information in the t-SNE plot of word embeddings correspond well to differences with regards to the usefulness of embedding features between the three entity categories evaluated. That is, tokens classified as belonging to the categories PERSON and LOCATION, for which word embeddings were useful, are shown as clusters of red dots in the t-SNE plots. In contrast, tokens classified as belonging to ORGANISATION, for which word embeddings were shown not to be useful, mainly occur as scattered dots in the plot.

The output of experiments on the Japanese data, for a model trained on 138 manually labelled microblogs, is shown in Figure 3. Figure 3a visualises the state of the pool with regards to the LOCATION category, and Figure 3b shows pre-annotations resulting from this model.[5]

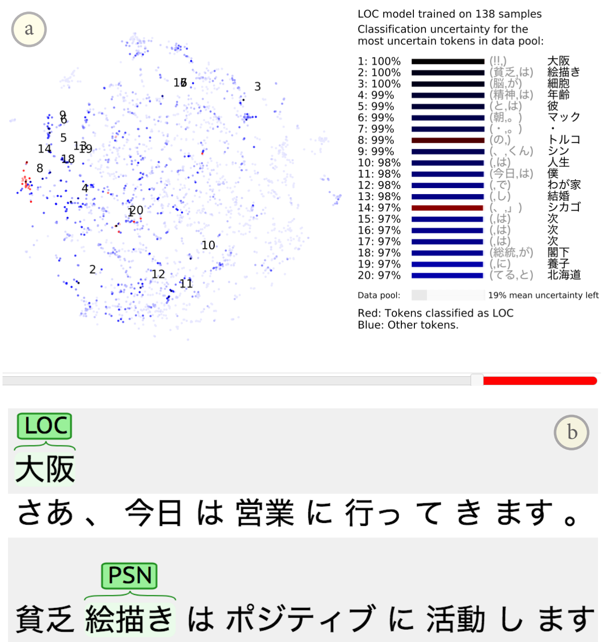

Figure 3: **(a)** The state for the LOCATION entity in the pool of unlabelled data, when the NER model has been trained on 138 manually labelled Japanese microblogs. Two potential entity clusters are shown in the t-SNE plot (close to *8*, Turkey, and *20*, Hokkaido). Which iteration is shown can be changed through the slider provided. **(b)** Pre-annotations for two samples selected for manual annotation, as they contain the two most uncertain tokens in the data pool, i.e., the tokens shown as the first two elements in the list of uncertain tokens.

[5]The code for PAL, as well as for the experiments reported here, can be found at: https://github.com/mariask2/PAL-A-tool-for-Pre-annotation-and-Active-Learning. There, a link can also be found to a video showing how the state of the pool changes with an increasing training data size.

5 Discussion

Results for the LOCATION and PERSON entities yield that the combined functionality of active learning and incorporation of unsupervised features has the potential to lead to large increases in results on small data sets. This, in turn, shows that these techniques form useful components for the use case on which we focused here, i.e., to achieve models that can give acceptable performance on small data sets and that can be applied for providing pre-annotations when annotating larger data sets.

The categories LOCATION and PERSON seem to be relatively coherent in terms of the contexts in which they occur, as shown by the large model performance increases achieved when word embedding features were incorporated. In contrast, that slightly better results were achieved for ORGANISATION without word embedding features, indicates that entities belonging to this category occur in semantically diverse contexts.

These differences in context coherence between different entity categories were also shown by the t-SNE plot functionality, which we provided to meet one of the visualisation goals of the PAL tool extension of this study, i.e., the goal of showing whether the word embeddings used as features formed clusters corresponding to manually annotated entity categories. Thereby, the annotator is provided with a possibility to estimate the effect of these word embedding features in the active learning process.

The t-SNE plot and the bar charts of the extended version of the PAL tool also meet the visualisation goals of showing why a particular set of samples were chosen for annotation, and of showing how the increased size of the training data set affects the performance of the trained model. An increased training data size led to that two of the classifiers achieved an F-score that might be high enough to be acceptable for pre-annotation, while the F-score remained low for the ORGANISATION category, also when the data size was increased. These differences were reflected in the visualisations of the effects of the increased training data size.

We believe that visualisations that aim to increase the human understanding of the active learning process and of the features used, and that show how the state of the data pool changes as more data is manually annotated, have the potential to increase the intrinsic motivation for the annotation task. Future work will therefore include user studies to determine how annotators perceive these visualisations that were added to the PAL tool, and how the visualisations affect the motivation for the annotation work. Such user studies should also include investigations of how the performance level of the machine learning model correlates with the perceived usefulness of the pre-annotations provided by the model.

6 Conclusion

We evaluated the ability of the PAL tool to reduce the training data size required through the use of active selection of data and through the incorporation of unsupervised features in the form of word embeddings. Results achieved for the categories LOCATION and PERSON showed that the combined functionality of active learning and incorporation of word embeddings has the potential to lead to large increases in results on small data sets. In contrast, word embeddings did not lead to any improvements in the performance for detecting the ORGANISATION entity, and low F-scores were achieved for this entity category, also when 1,000 samples were used for training the model.

The PAL tool was also extended with visualisation functionality, with the aim of increasing the human understanding of the active learning process and of the features used. The visualisations provided were able to indicate performance differences between the entities, as well as differences with regards to the usefulness of the embedding features. That is, the same differences that were shown in the formal evaluations against the gold standard annotations.

We hope that this study will inspire annotation projects to facilitate the annotation process by practically applying the methods that we have evaluated here. In particular, we hope that the application of PAL, and other tools that provide annotation support, will lead to that more annotation projects are being conducted on corpora for which crowdsourcing is not appropriate. For instance, corpora for specialised domains or smaller languages.

Acknowledgements

We would like to thank the reviewers for their valuable input. We would also like to thank the Swedish Research Council that funded this study (project numbers 2016-06681 and 2017-00626).

References

[Baldwin et al.2015] Timothy Baldwin, Marie-Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. 2015. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 126–135, Beijing, China, July. Association for Computational Linguistics.

[Copara et al.2016] Jenny Copara, Jose Eduardo Ochoa Luna, Camilo Thorne, and Goran Glava. 2016. Spanish NER with word representations and conditional random fields. In *Proceedings of the 6th Named Entities Workshop*, pages 34–40, Berlin, Germany. Association for Computational Linguistics.

[Derczynski et al.2016] Leon Derczynski, Kalina Bontcheva, and Ian Roberts. 2016. Broad twitter corpus: A diverse named entity recognition resource. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 1169–1179, December.

[dos Santos and Guimarães2015] Cicero dos Santos and Victor Guimarães. 2015. Boosting named entity recognition with neural character embeddings. In *Proceedings of the Fifth Named Entity Workshop*, pages 25–33, Beijing, China, July. Association for Computational Linguistics.

[Dumitrache et al.2013] Anca Dumitrache, Lora Aroyo, Chris Welty, Robert-Jan Sips, and Anthony Levas. 2013. "Dr. Detective": combining gamification techniques and crowdsourcing to create a gold standard in medical text. In *Proceedings of the 1st International Workshop on Crowdsourcing the Semantic Web, Sydney, Australia, October 19, 2013*, pages 16–31, Aachen, Germany. CEUR-WS.org.

[Godin et al.2015] Fréderic Godin, Baptist Vandersmissen, Wesley De Neve, and Rik Van de Walle. 2015. Multimedia lab @ acl w-nut ner sharedtask: named entity recognition for twitter microposts using distributed word representations. In *ACL 2015 Workshop on Noisy User-generated Text, Proceedings*, pages 146–153. Association for Computational Linguistics.

[Guo et al.2014] Jiang Guo, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Revisiting embedding features for simple semi-supervised learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 110–120.

[Henriksson2015] Aron Henriksson. 2015. Learning multiple distributed prototypes of semantic categories for named entity recognition. *Int. J. Data Min. Bioinformatics*, 13(4):395–411, October.

[Kucher et al.2017] Kostiantyn Kucher, Carita Paradis, Magnus Sahlgren, and Andreas Kerren. 2017. Active learning and visual analytics for stance classification with alva. *ACM Trans. Interact. Intell. Syst.*, 7(3):14:1–14:31, October.

[Kudo2006] Taku Kudo. 2006. Mecab : Yet another part-of-speech and morphological analyzer. http://mecab.sourceforge.jp.

[Lample et al.2016] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 260–270.

[Ling et al.2014] Wang Ling, Luis Marujo, Chris Dyer, Alan Black, and Isabel Trancoso. 2014. Crowdsourcing high-quality parallel data extraction from twitter. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, WMT '14, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Lingren et al.2014] Todd Lingren, Louise Deleger, Katalin Molnar, Haijun Zhai, Jareen Meinzen-Derr, Megan Kaiser, Laura Stoutenborough, Qi Li, and Imre Solti. 2014. Evaluating the impact of pre-annotation on annotation speed and potential bias: natural language processing gold standard development for clinical named entity recognition in clinical trial announcements. *J Am Med Inform Assoc*, 21(3):406–13.

[Martin et al.2017] David Martin, Sheelagh Carpendale, Neha Gupta, Tobias Hoßfeld, Babak Naderi, Judith Redi, Ernestasia Siahaan, and Ina Wechsung. 2017. Understanding the crowd: Ethical and practical matters in the academic use of crowdsourcing. In *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments*, pages 27–69, Cham. Springer International Publishing.

[Mikolov et al.2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

[Miller et al.2004] Scott Miller, Jethran Guinness, and Alex Zamanian. 2004. Name tagging with word clusters and discriminative training. In *Proceedings of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL HLT)*, pages 337–342, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Olsson2008] Fredrik Olsson. 2008. *Bootstrapping Named Entity Annotation by Means of Active Machine Learning*. Ph.D. thesis, University of Gothenburg. Faculty of Arts.

[Pedregosa et al.2011] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[Řehůřek and Sojka2010] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Paris, France, May. European Language Resources Association (ELRA).

[Reimers and Gurevych2017] Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348, Copenhagen, Denmark, September. Association for Computational Linguistics.

[Sahlgren2006] Magnus Sahlgren. 2006. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. thesis, Stockholm University.

[Sang and Meulder2003] Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL-2003*, pages 142–147.

[Schein and Ungar2007] Andrew I. Schein and Lyle H. Ungar. 2007. Active learning for logistic regression: an evaluation. *Mach. Learn.*, 68(3):235–265, October.

[Settles2009] Burr Settles. 2009. Active learning literature survey. Computer Sciences Technical Report #1648, University of Wisconsin–Madison, http://research.cs.wisc.edu/techreports/2009/TR1648.pdf.

[Shen et al.2004] Dan Shen, Jie Zhang, Jian Su, Guodong Zhou, and Chew-Lim Tan. 2004. Multi-criteria-based active learning for named entity recognition. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Skeppstedt et al.2016] Maria Skeppstedt, Carita Paradis, and Andreas Kerren. 2016. PAL, a tool for Pre-annotation and Active Learning. *JLCL*, 31(1):91–110.

[Stenetorp et al.2012] Pontus Stenetorp, Sampo Pyysalo, Goran Topic, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: a web-based tool for NLP-assisted text annotation. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 102–107, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Tjong Kim Sang and Buchholz2000] Erik F. Tjong Kim Sang and Sabine Buchholz. 2000. Introduction to the conll-2000 shared task: Chunking. In *Proceedings of CoNLL-2000 and LLL-2000*, pages 127–132. Lisbon, Portugal.

[Tomanek et al.2007] Katrin Tomanek, Joachim Wermter, and Udo Hahn. 2007. Efficient annotation with the Jena ANnotation Environment (JANE). In *Proceedings of the Linguistic Annotation Workshop*, pages 9–16, Stroudsburg, PA, USA, June. Association for Computational Linguistics.

[Turian et al.2010] Joseph P. Turian, Lev-Arie Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394, Uppsala, Sweden, July. Association for Computational Linguistics.

[van der Maaten and Hinton2008] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data Using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

[Venhuizen et al.2013] Noortje Venhuizen, Valerio Basile, Kilian Evang, and Johan Bos. 2013. Gamification for word sense labeling. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)*, pages 397–403, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Yamada et al.2015] Ikuya Yamada, Hideaki Takeda, and Yoshiyasu Takefuji. 2015. Enhancing named entity recognition in twitter messages using entity linking. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 136–140, Beijing, China, July. Association for Computational Linguistics.

[Yang et al.2016] Zhilin Yang, Ruslan Salakhutdinov, and William W. Cohen. 2016. Multi-task cross-lingual sequence tagging from scratch. *CoRR*, abs/1603.06270.