Enhanced Labelling in Active Learning for Coreference Resolution

Vebjørn Espeland
School of Informatics
University of Edinburgh &
Opus 2 International
s1471720@ed.ac.uk

Benjamin Bach
School of Informatics
University of Edinburgh
bbach@inf.ed.ac.uk

Beatrice Alex
School of Literature, Languages and Cultures
Edinburgh Futures Institute
University of Edinburgh
balex@ed.ac.uk

Abstract

In this paper we describe our attempt to increase the amount of information that can be retrieved through active learning sessions compared to previous approaches. We optimise the annotator's labelling process using active learning in the context of coreference resolution. Using simulated active learning experiments, we suggest three adjustments to ensure the labelling time is spent as efficiently as possible. All three adjustments provide more information to the machine learner than the baseline, though a large impact on the F1 score over time is not observed. Compared to previous models, we report a marginal F1 improvement on the final coreference models trained using for two out of the three approaches tested when applied to the English OntoNotes 2012 Coreference Resolution data. Our best-performing model achieves 58.01 F1, an increase of 0.93 F1 over the baseline model.

1 Introduction

Coreference resolution (CR) is the task of resolving which noun phrases (NP) in a text are referring to the same entity. It is related to entity linking, but does not involve an external knowledge base. It is an important task in information extraction, as a step in structuring the unstructured information in natural language. CR has traditionally been a difficult problem, as it is hard to accurately predict coreference links without extensive real-world knowledge.

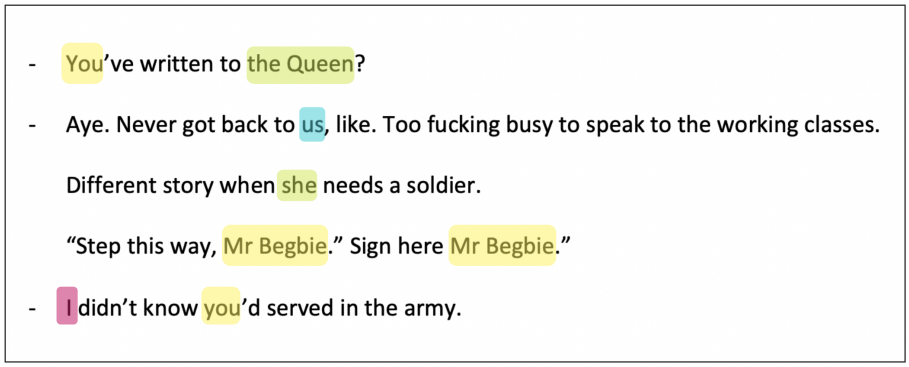

Figure 1: Different types of coreference resolution. An anaphoric pair of noun phrases is marked in green, and a cataphoric pair is marked in yellow. From "T2: Trainspotting" (Boyle, 2017)

An example of different levels of CR is shown in Figure 1. The mentions "us" and "I" are both singletons, and are not coreferring with anything in this text. The noun phrase "she" is anaphoric (where the pronoun points *backwards* to its antecedent) with "the Queen". The pronoun "You" in "You've" is coreferring with "Mr Begbie", but the pronoun is pointing *forward* to its coreferent, this type of coreference is cataphoric coreference.

Proceedings of the 3rd Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2020), pages 111–121,
Barcelona, Spain (online), December 12, 2020.

Many of the most successful coreference resolution approaches have used hand-crafted corpora, such as ACE (NIST, 2004), GAP (Webster et al., 2018) and OntoNotes (Pradhan et al., 2012). Models trained using these datasets, though comparatively successful, do not necessarily generalise to domain specific data, or noisy data. Making these big datasets is also a very expensive task, which is very difficult for low resource languages.

Active learning is a human-in-the-loop approach to machine learning, where a sample selection algorithm chooses the most informative samples for a human to annotate. This approach will reduce the total amount of samples which need to be labelled to achieve high accuracy, and in some cases it accelerates the otherwise expensive process of hand-crafting fully labelled datasets. Iteratively training and labelling this way would lead to higher accuracy models faster than training with random sampling.

The most expensive part of dataset creation is the labelling effort of the annotators. Therefore using the annotator's time as efficiently as possible should be a key focus in developing active learning techniques. As previous research (Section 2.2) has focused on which samples to label, this article will focus on improving the use of the annotator's time. The objective of this research is to improve the amount of information that can be retrieved through the active learning sessions.

Aiming to use the annotator's time as efficiently as possible, this article suggests three improvements to recent developments in active learning for coreference resolution. We investigate whether it is effective to label all the instances of an entity once the user has been asked to provide the first label of the entity. We also suggest an improvement based on allowing the user to edit an incorrectly identified mention and then provide coreference information, rather than disregarding that candidate coreferent pair. Finally, for mentions which are the first instances of their entity, such as the example of "Mr Begbie" above, we allow the user to provide cataphoric labels. We use the English OntoNotes 2012 Coreference Resolution dataset provided by the CoNLL 2012 shared task (Pradhan et al., 2012) to simulate dataset creation using active learning techniques.

In this paper we firstly review the related work on coreference resolution and active learning in Section 2. Then in Section 3 and 4 we explain the experimental methodology and review the results. Finally in Section 5 and 6 we analyse the results before our conclusions and directions for future work.

2 Related work

2.1 Coreference resolution

A detailed review of the early research in coreference resolution was made by Ng (2010). I will summarise this in short in this section, and move on to reviewing the later research, especially the approaches using deep learning.

Past coreference resolution research can be divided into two approaches: mention-pair and mention-ranking. The mention-pair models attempt to reduce the coreference resolution challenge to a binary problem, whether two NPs are coreferring or not. Aone and Bennett (1995) and McCarthy and Lehnert (1995) were early proponents of this method. The mention-ranking models aim to rank the candidate antecedent mentions according to likelihood of coreferring. Connolly et al. (1997) were the first to apply this approach. Other mention ranking approaches include Iida et al. (2003), Yang et al. (2003), and Yang et al. (2008).

Durrett and Klein (2013) tried to reduce the amount of expensive hand-crafted features. This idea was picked up by Wiseman et al. (2015). The benefit of using neural networks is that the fine-tuning of these features is left in the hidden layers of the network. With the arrival of word-embedding techniques after the very influential paper by Mikolov et al. (2013), much of the research in natural language processing (NLP), including coreference resolution, took a step in the direction of using neural networks.

Clark and Manning (2016a) used a deep neural network to capture a larger set of learned, continuous features indicating that more entity-level information is beneficial to the coreference task. Based on this finding, they trained a neural mention-ranking model using reinforcement learning (Clark and Manning, 2016b). They claimed that, despite being less expressive than the entity-centric models of Haghighi and Klein (2010; Clark and Manning (2015), their model is faster, more scalable and simpler to train.

Lee et al. (2017) presented a neural end-to-end coreference resolution system, without using a syntactic parser or a mention detector to extract the candidate mentions. They combined context-dependent boundary representations with an attention mechanism for NP head finding, inspired by Durrett and Klein (2013) to treat aggregated spans of words as a unit. The likelihood of two spans being coreferent is determined by merging the likelihood of either span being a mention with the likelihood of them coreferring.

Finally, with the arrival of transformers and BERT (Devlin et al., 2018), the field of NLP took another leap forward. Coreference resolution approaches using BERT include Joshi et al. (2019) and Joshi et al. (2020).

2.2 Active learning

When building a dataset for NLP tasks, a human annotator would normally have to label every single sample in the dataset which is a very expensive process. The use of active learning is an appealing solution to creating and labelling datasets, as the human annotator would only have to annotate the most informative samples. There are two main considerations in the active learning process outside of user interface design: how to choose which samples to label, and how to label them. The first consideration has been the most researched, the second is the focus of this article.

There is an array of techniques to choose which samples to label next. Using an informativeness measure such as entropy enables an algorithm to choose the samples with the highest uncertainty. Lewis and Gale (1994), Gasperin (2009) and Schein and Ungar (2007) use this technique with varying degrees of success. Other methods include ensemble models like query-by-committee (QBC) and cluster-outlier methods. Sachan et al. (2015) reviewed these and found that all these methods performed better than random sampling, and that the ensemble model is the best performing one. Settles (2009) reviewed general active learning literature, and Olsson (2009) reviewed the AL literature within the scope of NLP. Recently, Shen et al. (2017) used active learning for named entity recognition, achieving close to state-of-the-art results with only 25% of the training data.

For deciding what to do with the selected samples, the dominant approach has been binary pairwise selection for potential manual coreference annotation (Gasperin, 2009; Laws et al., 2012; Zhao and Ng, 2014; Sachan et al., 2015). This approach pairs up candidate mentions with candidate antecedents, and the annotator can discard or accept a mention-pair dependent on whether they are coreferring or not. Sachan et al. (2015) introduced *must-link* (ML) and *cannot-link* (CL) constraints as a method of storing user annotations. The mention-pairs which where deemed coreferent received the ML constraint, and the ones deemed not coreferent received the CL constraint, where the coreference likelihood of those pairs was set to 1 and 0 respectively. Applying transitivity (if A is coreferent with B, and B with C, then A and C must also be coreferent) to these constraints means more labels can be distributed without extra labelling.

Li et al. (2020) improved on the mention-pair constraints by using span embeddings instead of mentions, as successfully applied to coreference resolution in Lee et al. (2017). They also augmented the pair-wise annotation with a second step of marking the first occurrence of the entity if the span pair is not coreferent, introducing the notion of discrete annotations.

The marking of the first occurrence of the entity allows the annotator to cluster the entities. Together with the notion of transitivity, this makes annotation more efficient, as it makes use of some false negatives. However, this approach, though better than pairwise decision, still does not make use of the false positives. It also ignores readily available information about other occurrences of the entity in question.

It takes time for an annotator to find the *first* sample of the highlighted entity, particularly if the document they are labelling is more than a few sentences. When the annotator has spent the time finding the first occurrence of the entity, they will have identified many, if not all, of the other occurrences of that entity, and it will be relatively cheap to annotate all the occurrences in the document. A good interface will have predicted and highlighted these occurrences.

If the sample turns out to be negative, e.g. by the proform span (the span in question, as opposed to the antecedent span) being the first span in the document, then allowing the annotator to label cataphoric

spans would also contribute towards the goal of increasing annotator efficiency.

The setup in Li et al. (2020) allows a candidate coreferent pair to be disregarded in three ways, where only the third way should be a valid reason for disregarding:

1. The span is incorrectly identified, and is not a valid noun phrase.
2. The span is the first mention of that entity (and thus has no antecedent).
3. The span is the only mention of that entity in the document.

The following section will elaborate on the experiments to improve upon these shortcomings.

3 Methodology

The experiments reported in this paper investigate a set of different methods for conducting manual annotation during an active learning scenario.

3.1 Discrete annotation with cataphoric links

Previous approaches to active learning for coreference resolution have focused primarily on antecedent labelling, ignoring potential occurrences following an entity. The OntoNotes dataset is not made with specific cataphoric linkings. This makes it more difficult to test how well the system performs when adding cataphoric data. It is still however possible to retrieve cataphoric mentions of an entity from the dataset.

Even though the sample selection algorithm will only select entities with a candidate *antecedent*, it should be possible for the annotator to choose cataphoric occurrences. Our simulated experiment will test whether allowing the annotator to select cataphoric mentions will have an impact on how many label queries are disregarded.

3.2 Annotating all spans for the queried entity in the document

This is motivated by the experience that it is easier to label multiple spans of the same entity in the same document than it is to annotate just one instance, even if the document contains several occurrences of that entity. Even though more samples are being labelled, and those samples are not necessarily the most informative ones, they will still provide more information per query and per clock-time than strictly pair-wise or discrete annotation.

The improvement would be made by adding multiple ML and CL constraints for each query. Every time a suggested pair is not the final pair of that query a CL constraint is applied, and every label the annotator selects receives a ML constraint. This, combined with transitivity constraints (elaborated in Li et al. (2020)), is hypothesised to increase the amount of information available to the learner.

3.3 Annotation error

Whether the annotator is helped by interface highlighting of predictions or not, a potential challenge with asking an annotator to label all occurrences of an entity in the document is that they are susceptible to losing focus due to boredom or time pressure. In these situations it is plausible that there will be a certain amount of error. Taking inspiration from Sachan et al. (2015), which included user labelling error as a hyperparameter, we include labelling error in our experiments.

3.4 Enabling span editing and annotating all spans

In previous approaches to active learning for coreference resolution, when an annotator is queried with a span which is incorrectly identified as a span, that query is disregarded. There is no difference between a CL constraint because of correctly identified spans not linking, and a CL constraint caused by correctly linked but incorrectly identified spans. These kinds of boundary errors are common in entity recognition, and these frequent errors can have a big impact on downstream performance. In the discrete annotation, Li et al. (2020) improved this problem by making the user click all the words in the antecedent span,

building the span word by word. However, they did not allow the user to correct the proform span. This limitation also applies to their simulated experiments.

We therefore allow the user to correct the proform span. The method for manually correcting the proform span is letting the annotator choose which words belong to the span. In the simulated experiment we scan the indeces of all spans in that document for the closest span that belongs to a coreference cluster in the dataset. We then find an antecedent to the new proform, and make a new ML constraint, leaving a CL constraint to the initial candidate pair. If the nearest span is not coreferent with any other span in the document, the incorrectly identified span is unlikely to be a boundary error, and the query is therefore disregarded as not coreferring.

4 Evaluation

We compare the baseline discrete labelling system versus enhanced labelling using the standard English CoNLL-2012 coreference resolution dataset (Pradhan et al., 2012). Following both Li et al. (2020) and Sachan et al. (2015), user labelling is simulated from the gold standard labels in the CoNLL dataset.

4.1 Evaluation metric

In the field of coreference resolution there are multiple ways of scoring a system, each with their own benefits and drawbacks. A somewhat standardised option, and the one chosen to evaluate the experiments reported in this paper, is to combine the recall and precision from MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998) and CEAFe (Luo, 2005) as an average F1 score. We compute this score with the official CONLL-2012 evaluation scripts.

We also compare the amount of successful queries in each AL session as a metric of how successful the annotation approach is at providing positive training examples. A successful query is a query which returns a coreferent pair, regardless of whether the original proform or antecedent candidate were coreferent or not. This way, there will be at least one ML constraint from that query. An unsuccessful query does not return a coreferent pair, and the only thing that can be learnt from that query is that the original proform and antecedent candidates are not coreferent, resulting in only one CL constraint.

4.2 Neural network architecture

For the sake of comparison we use the same coreference model as in (Li et al., 2020). They use the AllenNLP implementation of Lee et al. (2017), which keeps all the hyperparameters, except that it excludes speaker features, variational dropout and limits the maximum number of considered antecedents to 100. In Lee et al. (2017), they use GloVe embeddings (Pennington et al., 2014) as word embeddings. They use a bidirectional LSTM (Hochreiter and Schmidhuber, 1997), where the hidden states have 200 dimensions, to represent the aggregated word spans. The model internal scoring for determining whether a span is a mention, and whether two mentions are coreferring, is using feed-forward neural networks consisting of two hidden layers with 150 dimensions and rectified linear units (Nair and Hinton, 2010). The optimiser used is ADAM (Kingma and Ba, 2014).

4.3 Experiments

We ran simulated AL experiments with the OntoNotes 2012 Coreference Resolution dataset using the following setup. Each experiment is based on Li et al. (2020), using their entropy selector as sample selection algorithm, selecting 20 queries from each document. The OntoNotes is split into 2802 training documents, 343 validation documents and 348 testing documents. The validation set is used to compute F1 score while training, whereas the test set is used only for final F1 score computation after training has finished.

A 700-document subset of the training data is set aside, and the initial model is trained on this subset. The model trains until convergence with a patience of 2 epochs, up to 20 epochs, before adding more data. Then 280 documents are labelled in an AL session. After these 280 documents are labelled, they are added to the 700 documents, and training continues on the now 980 documents in the set aside training subset.

This continues until all the 2802 documents in the training set have been labelled. Finally, a new model trained on all the 2802 training documents with all the model and training parameters reset. This last step is to make the final model comparable to other models trained without AL, and use the same hyperparameter as Lee et al. (2017). There are 20 span-pair queries per document in the AL session, meaning 5600 queries per AL session, and a total of 39200 queries over the 8 AL sessions.

For labelling with error, 10% of the labels retrieved in the annotation session are set to a random span in the document. We implement this by introducing a 10% chance of having a random span chosen instead of a coreferring span. This is to prevent the erroneous labels systematically having the same index each AL session.

We include one baseline experiment from Li et al. (2020). The experiment is using discrete annotation with the same parameters as our experiments, but we report the F1 score for the baseline with the best performing experiment from Li et al. (2020), which uses a query-by-committee system with three models. This is done to compare the results of our experiments to the currently best performing coreference resolution system using AL.

In the baseline experiment and Experiments 1 and 3, the annotator is only allowed to select one occurrence of the proform entity. In Experiment 2 the annotator labels all the anteceding occurrences of the proform, whereas in 4 and 5 the annotator labels all the occurrences of that entity.

We also perform a timed annotation exercise with the same setup as in Li et al. (2020). We recruited 10 annotators with experience in text processing, who annotated for 30 minutes each. Li et al. (2020) used annotators with NLP experience, whereas our annotators did not that but are skilled in working with speech transcripts. This might impact the absolute annotation time, but the relative annotation time within our group of annotators should still be informative. The annotators in Li et al. (2020) were asked a pair-wise question first, and in the case of non-coreference they were asked to annotate the *first* instance of the entity. In contrast, we asked our annotators to label *all* instances of the entity in the case. When an annotator provided only one extra instance of the entity, that was noted as a "follow-up question", whereas when they labelled more than one extra instance of the entity it was noted as a "multi-response". We used the same annotation interface as in Li et al. (2020), but altered it to allow cataphoric labelling as well as multiple labels per query.

4.4 Results

Table 1 shows the results from our timed annotation exercise. In our experiment the annotators spent longer on the initial question (20.66 s), but were faster on supplying answers for the follow-up question (12.61 s). When annotating more than one extra occurrence, the time taken for each of those occurrences was lower than answering the initial question.

The average normalised annotation time per occurrence was 16.57 seconds. In contrast, the annotators' median normalised annotation time was only 10.26 seconds per occurrence. This indicates that the distribution of annotation times is higher at the lower end, and that there were a few queries with very

		Avg. Time per query
Li et al. (2020)	Initial question	15.96s
	Follow-up question	15.57s
	ONLY Follow-up question	28.01s
Our experiments	Initial question	20.66s
	Follow-up question	12.61s
	Normalised multi-response	16.57s

Table 1: Results for the timed annotation exercise. We first list the results from the corresponding timed exercise reported in Li et al. (2020). The fourth and fifth results for our equivalent experiments, with the exception that the annotators were allowed to select any instance of the entity in the follow-up, not just the first. The final time in the table is the average time taken for the annotators to label every instance of the entity, normalised by the number of labels in each query.

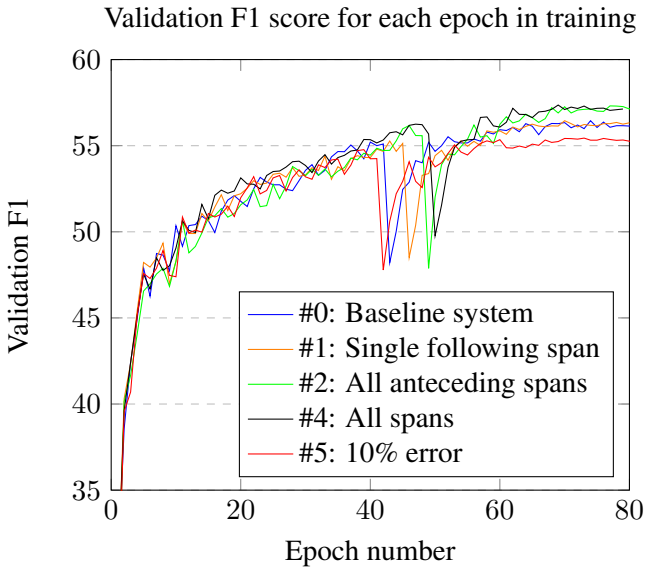

Figure 2: The F1 score while training for each experiment. This score is computed using the validation dataset. As expected, the scores are similar at the earlier stages, when the model is trained on the same number of labels. For the later epochs the models trained on more labels, Experiment 2 and 4, perform marginally better than the other models. The dip in F1 score around epoch 50 represent the retraining of the model from scratch after all the documents have been labelled.

long times which might have skewed the average. The fastest annotations for the multi-response queries were made in 2.07 when normalised for the number of labels annotated in that query. The slowest annotations took 124.95 seconds.

Figure 2 plots the F1 score over the training epochs, using the validation data. The improvements in F1 over the epochs are very similar for each of the training methods in the early stages, but in the later stages the active learning approaches which allow multiple labelling come out on top.

In the baseline experiment 49% of the queries return a coreferent label pair, which means over half of the queries did not result in a ML constraint. In Experiment 1 that number is increased to 54%, as can be seen in Table 2. This is a reduction of disregarded queries by 11%. In Experiment 2 and 4 the simulated annotator is instructed to label all the occurrences of the entity in the given document, which results in several label pairs per query. For Experiment 2 there are 0.93 label pairs per query, whereas for Experiment 4 there are 1.41 label pairs per query.

There was no difference between the labels retrieved for Experiment 3, where the annotator was allowed to edit proform spans and the results for the baseline experiment. A total of 6 spans were edited

#	**Experiment**	Successful labels per query	CONLL F1 score
0	Discrete annotation (Li et al., 2020)	0.51	57.08
1	Allowing following spans	0.54	**58.01**
2	Annotating all anteceding spans	0.93	57.18
3	Allowing proform edit	0.51	56.09
4	Combining 1 and 2	**1.41**	57.37
5	Combining 1 and 2 with 10% error	0.52	55.48

Table 2: Experiments for the AL models, with the F1 score representing the performance on the final models on the test set. The "Successful label per query" column explains how many queries returned with positive coreferent pairs. The F1 score for the baseline (Experiment 0) is achieved using a sample selector with the query-by-committee approach. When Experiment 2 and 4 are close to and exceeding 1 that is because they are returning more than one label pair per query.

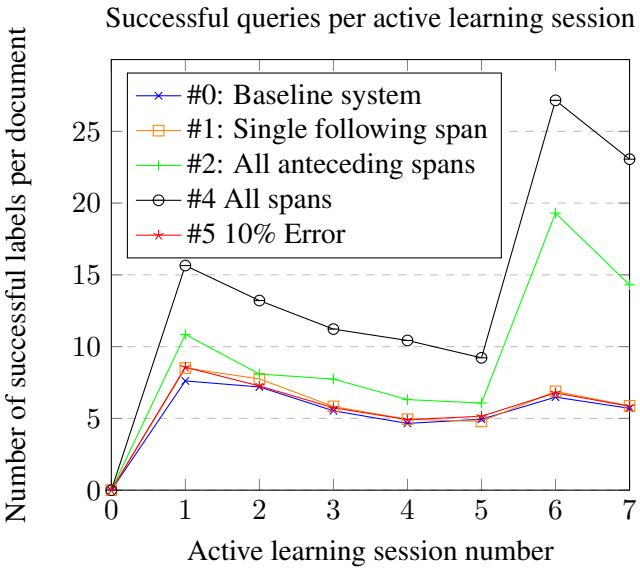

Figure 3: The number of successful queries for each AL session. The sessions have been normalised for document length, as some of the sessions have significant longer document lengths. Experiment 3 is not included, as it was overlapping with the baseline system. The approaches in Experiment 2 and 4 are more effective at providing successful label pairs than the other experiments, particularly with longer documents.

under the simulated experiment.

In Figure 3 the label pair counts are separated into the active learning sessions, and normalised by average document length for that session. This measure can be seen as an average number of successful label-pairs per document. In Experiment 1 there are marginally more labels successfully identified than in the baseline system. For both Experiment 2 and 4 the AL sessions provide many more label pairs per document, up to an average 27.16 label pairs for Experiment 4 in AL session 6. The efficacy of the combined model is reduced when 10% labelling error is added in each AL session, but Experiment 5 still provided more labels than the baseline system.

5 Analysis and Discussion

The timed annotation exercise show that the cost of annotating all the labels of an entity in a text is low when the annotator has already read the text to make a judgement on the initial coreference pair. The results also show that there might be a cut tail distribution of annotation times. The majority of the multi-response annotations were faster than the initial and the single-response follow-up question responses.

On average it took our annotators longer time for the initial question in our implementation of the same timed annotation exercise as in Li et al. (2020), but shorter time for the follow-up question. People working in NLP are likely to be more experienced with seeing text containing bracketed annotation. It is possible that our set of annotators were slower at responding for the initial question because of the lack of experience in NLP.

One reason the average time for answering the follow-up question was lower in our setup might be that the annotators were allowed to label any instance of the occurrence, not just the first. Particularly for longer texts it might be faster to label an occurrence closer to the proform entity than the first occurrence.

From Figure 3 we can see that the labelling approach in Experiment 1 returns more labels per query than the baseline approach, through the AL sessions. The same is true for Experiment 4 and 2 respectively. This indicates that cataphoric occurrences contain unused information, which should be used for training. The sudden jump in successful queries in AL sessions 6 and 7 for Experiment 2 and 4 can partly be ascribed to an increase in document length in those sessions, even though the graph is normalised to

document length. This might mean that models trained on datasets with longer documents are able to benefit more from the improved label retrieval rate.

Even with 10% of the labels chosen at random the combined approach retrieved more successful label pairs than the baseline system, but the final F1 score was somewhat lower. This lower score F1 was expected, as the erroneous labelling would add confusion to the model. Care should therefore be taken when designing a labelling system to ensure that errors are minimised.

The small improvement in the validation F1 score shown in Figure 2 indicates that the added labels under the current system do not translate into having an impact on how fast high accuracy is achieved. Despite this, the final F1 score on the separate test data is marginally higher for Experiment 4 than the baseline experiment.

This lack of impact could have several causes. As the machine learning algorithm is the same as in the baseline system, it might not be best suited to make use of the extra available information. In addition, the OntoNotes dataset does not inherently support cataphoric linking of entities, so a dataset which does contain inherent cataphoric links might also contribute towards making use of the extracted data more efficiently.

The negative results for Experiment 3 can have multiple causes. One of these is that the algorithm for selecting replacement proform spans was purposefully conservative in choosing the closest span. This was to retain ecological validity in the annotation simulation, as an annotator would look close to the span to determine whether the error was a boundary error.

6 Conclusion and Future Research

The contribution of the research in this article is the improved techniques for extracting more information from user labelling. We have seen that allowing annotators to leverage cataphoric information, especially in combination with annotating several occurrences per query, can contribute to optimising the time spent by annotators hand labelling a dataset. Even though the machine learning models did not perform markedly better earlier in the training process, the amount of disregarded queries dropped by a noticeable amount just by adding cataphoric labels.

We have also seen that the amount of successful label pairs per query is over 1 for the approaches allowing multiple responses. This means that it is possible to extract much more information than with previous approaches. Our timed annotation exercise indicate that labelling several occurrences of an entity in the same query is faster than answering multiple queries with only one set of labels. It would be interesting to investigate whether choosing labels closer or further from the proform label would have an impact on the learning.

These findings are interesting for the real world application of coreference resolution systems, particularly for long form documents, such as in the legal sector, where there is a lot more information to leverage than in short form documents. A future project would look into making changes to the machine learning model for more effective use of the new data.

Future research would also look into testing which interface design would best aid the human annotator in the labelling process, especially for long form documents.

References

Chinatsu Aone and Scott William Bennett. 1995. Evaluating automated and manual acquisition of anaphora resolution strategies. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 122–129. Association for Computational Linguistics.

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference*, volume 1, pages 563–566. Granada.

D. Boyle. 2017. T2: Trainspotting. United Kingdom: TriStar Pictures, Inc., Sony Pictures Releasing.

Kevin Clark and Christopher D Manning. 2015. Entity-centric coreference resolution with model stacking. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1405–1415.

Kevin Clark and Christopher D Manning. 2016a. Deep reinforcement learning for mention-ranking coreference models. *arXiv preprint arXiv:1609.08667*.

Kevin Clark and Christopher D Manning. 2016b. Improving coreference resolution by learning entity-level distributed representations. *arXiv preprint arXiv:1606.01323*.

Dennis Connolly, John D Burger, and David S Day. 1997. A machine learning approach to anaphoric reference. In *New methods in language processing*, pages 133–144.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Greg Durrett and Dan Klein. 2013. Easy victories and uphill battles in coreference resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1971–1982.

Caroline Gasperin. 2009. Active learning for anaphora resolution. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 1–8.

Aria Haghighi and Dan Klein. 2010. Coreference resolution in a modular, entity-centered model. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 385–393. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Ryu Iida, Kentaro Inui, Hiroya Takamura, and Yuji Matsumoto. 2003. Incorporating contextual cues in trainable models for coreference resolution. In *Proceedings of the 2003 EACL Workshop on The Computational Treatment of Anaphora*.

Mandar Joshi, Omer Levy, Daniel S Weld, and Luke Zettlemoyer. 2019. Bert for coreference resolution: Baselines and analysis. *arXiv preprint arXiv:1908.09091*.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Florian Laws, Florian Heimerl, and Hinrich Schütze. 2012. Active learning for coreference resolution. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 508–512, Montréal, Canada, June. Association for Computational Linguistics.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. *arXiv preprint arXiv:1707.07045*.

David D Lewis and William A Gale. 1994. A sequential algorithm for training text classifiers. In *SIGIR'94*, pages 3–12. Springer.

Belinda Li, Gabriel Stanovsky, and Luke Zettlemoyer. 2020. Active learning for coreference resolution using discrete annotation. *arXiv preprint arXiv:2004.13671*.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 25–32. Association for Computational Linguistics.

Joseph F McCarthy and Wendy G Lehnert. 1995. Using decision trees for coreference resolution. *arXiv preprint cmp-lg/9505043*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *ICML*.

Vincent Ng. 2010. Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 1396–1411. Association for Computational Linguistics.

NIST. 2004. Automatic content extraction (ace).

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

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 1–40. Association for Computational Linguistics.

Mrinmaya Sachan, Eduard Hovy, and Eric P Xing. 2015. An active learning approach to coreference resolution. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Andrew I Schein and Lyle H Ungar. 2007. Active learning for logistic regression: an evaluation. *Machine Learning*, 68(3):235–265.

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

Yanyao Shen, Hyokun Yun, Zachary C Lipton, Yakov Kronrod, and Animashree Anandkumar. 2017. Deep active learning for named entity recognition. *arXiv preprint arXiv:1707.05928*.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th conference on Message understanding*, pages 45–52. Association for Computational Linguistics.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the gap: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Sam Wiseman, Alexander Matthew Rush, Stuart Merrill Shieber, and Jason Weston. 2015. Learning anaphoricity and antecedent ranking features for coreference resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics.

Xiaofeng Yang, Guodong Zhou, Jian Su, and Chew Lim Tan. 2003. Coreference resolution using competition learning approach. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 176–183. Association for Computational Linguistics.

Xiaofeng Yang, Jian Su, and Chew Lim Tan. 2008. A twin-candidate model for learning-based anaphora resolution. *Computational Linguistics*, 34(3):327–356.

Shanheng Zhao and Hwee Tou Ng. 2014. Domain adaptation with active learning for coreference resolution. In *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)*, pages 21–29, Gothenburg, Sweden, April. Association for Computational Linguistics.