PhD thesis proposal: Active learning for semantic annotation of large media archives, in the presence of structured background knowledge

Tim Cowlishaw

June 12, 2014

Abstract

This document proposes a novel approach to topic classification within archives of textual and audiovisual material. By applying an active learning approach to training a relational model, it leverages both the structure of the topic space and the knowledge of users of the archive in order to better model the allocation of documents to topics and the relationships between documents.

1 Introduction

The BBC's *Redux* (Butterworth 2008) and *World Service Archive* (Raimond and Ferne 2013) projects are large archives of audiovisual material (along with transcripts and descriptive text) with tremendous value to programme-makers and researchers both within the BBC and further afield. However, effective search and discovery within the archive is often hampered by the relative sparsity of semantic metadata describing the items in the archive, their topics, and relationships. Due to the size of the archives, manual annotation of each item by professional archivists or subject experts would be prohibitively time-consuming and expensive. There are many existing approaches to automatically identifying topics and relationships in large text corpora in the machine learning literature. However, traditional supervised learning approaches require an authoritative source of ground truth labels for a subset of the corpus, in the form of a training set - the manual preparation of which can also be expensive and time consuming.

We propose an *Active learning* approach (Settles 2012) to this task, which leverages the fact that the users of this system are, in general, qualified to

perform this annotation themselves (for topics within their area of interest). By interactively querying the user for ground-truth labels related to their own topic of interest, this approach would assist users with their immediate information retrieval needs by learning a relevance metric for their particular query, while also making use of the same ground-truth labels to build a general topic model of the entire archive for use in future retrieval tasks. More specifically, we will focus on learning within topic spaces which have some existing, known structure, in the form of relations or hierarchies.

This project would build upon existing work in the field of active learning, natural language processing and topic modelling (reviewed in section 4) to develop a novel active-learning approach to inference in topic models possessing some known inter-topic structure, motivated by the specific challenges encountered by users of the BBC Redux and World Service Archives.

This contribution would also have immediate practical application for users of the BBC's archive (and related projects such as the effort to semantically annotate news articles described in Shearer 2013), as well as wider applications to fields such as E-discovery for the legal profession, computer-assisted reporting, and more generally, any information retrieval task concerning a structured ontology of topics.

2 Background

The challenge of automatically annotating documents with structured semantic metadata has been addressed in previous projects by BBC Research and Development, including work on automated concept tagging and document linking. However, to date, this work has consisted of using supervised learning models which require a training set which must be compiled by hand in advance, or substituted by some heuristic or external source of information which serves as a suitable proxy for ground truth (as in Raimond and Lowis 2012).

Given that a large proportion of the users of the BBC Redux and World Service Archive systems are journalists, researchers and programme makers with a significant degree of knowledge around the topics of the programmes that they are seeking, it would be advantageous to leverage this knowledge in identifying topic structure within the corpus. Active learning provides a principled framework for this, by allowing a model to be incrementally trained by querying a human oracle for ground-truth classifications of documents from the corpus.

This approach has already been used with some success on related information retrieval problems. Lang 1995 describes a use of an actively learnt model based on the minimal description length principle to perform content-based filtering and recommendation on a corpus of Usenet posts, and reports an improvement in precision over TF-IDF (a modified version of which was used in Raimond and Lowis 2012). In addition, McCallum and Nigam 1998 describes a novel use of active learning to improve the accuracy of a naive Bayes classifier where prelabelled training data is sparse.

We propose to build upon these techniques (and further related work described in section 4), incorporating techniques from statistical relational learning (Getoor and Taskar 2007) in order to devise a novel method for active learning of hierarchical or relational structured topic models.

3 Methodology

This thesis would provide a large-scale empirical evaluation of the use of active learning for semantic annotation in the BBC's Redux and World Service archive, and its effectiveness in solving information retrieval challenges for the users of these services. This would include:

3.1 Evaluation of existing active learning approaches for information retrieval in the BBC's Archive

A thorough, empirical analysis of the effectiveness of active learning techniques already published in the literature, including a comprehensive appraisal of different query strategies and error measures, as well as a comparison to established (passive) supervised and semi-supervised learning algorithms. An overview of some of the techniques to be evaluated is given in section 4.

3.2 Advancing the state-of-the-art in active learning for text classification, topic modelling and related natural language processing challenges

We will use the use-case of the BBC archives to motivate further advances in active learning for topic modelling, focusing on probabilistic models which incorporate some known structure over the topic space. For example, we will investigate active methods of performing inference for the parameters of topic models such as Latent Dirichlet Allocation (Blei, Ng, and Jordan

2003) and related extensions which model topic hierarchies (Blei et al. 2003), Gaussian Process Topic Models (Agovic and Banerjee 2012), and statistical relational models such as Markov Logic Networks (Domingos and Richardson 2007).

By combining active learning approaches to inference with models that reflect the structure of the topic space in question, we hope to make a novel contribution to the field of automated topic modelling and classification. Specific approaches could include the use of query-by-committee (Argamon-Engelson and Dagan 2011) to estimate classification variance in a model where computing such statistics directly would be intractable (as described for naive-Bayes classifiers in McCallum and Nigam 1998), combining active learning with complementary semi-supervised and unsupervised learning techniques, and the use of active learning to infer a topic distribution over documents, by defining a topic model over users of the system and propagating this to documents via their training decisions.

3.3 Investigation of human-computer interaction challenges related to active learning

The use of machine learning techniques for information retrieval tasks poses specific challenges in terms of user interface design and human computer interaction (Brajnik, Guida, and Tasso 1990). Active learning, in particular, leads to a specific set of interaction and interface design trade-offs (Rubens, Kaplan, and Sugiyama 2011). An active learning system, used for real-world information retrieval tasks (as we are proposing) would provide an excellent test subject for further research into user interface issues related to active learning, and particularly active learning approaches to text classification, search and topic modelling; an area which has not yet been examined in detail in the HCI literature. In particular, we propose examining approaches to the new user problem (Rashid, Albert, and Cosley 2002) in relation to active learning, as well as methods of modelling the effort needed to label a given document, in order to optimise an active learning system to jointly minimise classification error and user effort expended on labelling tasks. (Settles, Craven, and Friedland 2008).

4 Related work

Active learning (and specifically active learning for natural-language processing) is a well-studied and active area of research. In particular, advances have been made in the use of active learning for classification tasks with

a variety of model classes, including naive Bayes classifiers (McCallum and Nigam 1998) and SVMs (Tong and Koller 2002), as well as named entity recognition (Olsson 2008) and word-sense disambiguation (Fujii et al. 1998). A broad overview of the active learning literature is given by (Settles 2010), and a more specific review of literature related to active learning for NLP is given by (Olsson 2009). More generally, active learning has been used successfully to estimate the parameters of families of probabilistic model of use in a natural language processing context, including HMMs (Argamon-Engelson and Dagan 2011) and conditional random fields (Settles 2008 and Olsson 2008).

Of particular relevance is existing work which leverages existing domain knowledge in the task of learning a topic model. Andrzejewski 2010 describes several extensions to the Latent Dirichlet Allocation topic model which do exactly this, and which could be built upon to derive analogous active learning approaches. In addition, existing approaches which treat individual users of an active learning system as draws from a probability distribution, most notably Sheng, Provost, and Ipeirotis 2008's work on learning with multiple noisy oracles offer a useful stating point for research on modelling the interests users of an active learning system as a probabilistic topic model.

In addition, there is a large body of existing work on statistical topic models which incorporate some sort of topic structure. Blei et al. 2003 extends the LDA model to incorporate hierarchies of topics, while Agovic and Banerjee 2012 models covariances between individual topics through the use of Gaussian processes. The field of statistical relational learning (Getoor and Taskar 2007) also potentially offers a large amount of useful background on modelling richly structured topic ontologies around documents, including Bayesian logic programs (Kersting and De Raedt 2001), Relational dependency networks (Neville and Jensen 2007) and Markov logic networks (Domingos and Richardson 2007). In particular, Fischer 2011 describes how relational learning can be used for incorporating background knowledge into semantic web data mining tasks. Thompson, Califf, and Mooney 1999 describes an active approach to information extraction based on inductive logic, which would provide a useful starting point for further research into active learning for relational models.

The HCI and Information Retrieval literature also includes a number of works which are of interest, most notably Chang et al. 2009's research into human interpretations of probabilistic topic models (the effects of which are particularly pertinent in an active learning context), as well as Rashid, Albert, and Cosley 2002's work on the new user problem in recommender systems. In addition, there is a large amount of research on *relevance feed-*

back (Koenemann and Belkin 1996) in the information retrieval literature which is relevant to the challenge of actively training a topic model. In particular one potentially fruitful avenue of research could involve investigating the use of implicit feedback from a user's behaviour after performing a query, analogous to that described in Kelly and Belkin 2001.

References

[1] Amrudin Agovic and A Banerjee. "Gaussian process topic models". In: *CoRR* (2012). URL: http://arxiv.org/abs/1203.3462.

[2] DM Andrzejewski. "Incorporating domain knowledge in latent topic models". PhD thesis. 2010. URL: http://pages.cs.wisc.edu/~andrzeje/publications/dave-thesis.pdf.

[3] S Argamon-Engelson and I Dagan. "Committee-Based Sample Selection For Probabilistic Classifiers". In: *Journal of Artificial Intelligence Research* 11 (2011), pp. 335–360. URL: http://arxiv.org/abs/1106.0220.

[4] DM Blei, AY Ng, and MI Jordan. "Latent dirichlet allocation". In: *the Journal of machine Learning research* 3 (2003), pp. 993–1022. URL: http://dl.acm.org/citation.cfm?id=944937.

[5] DM Blei et al. "Hierarchical Topic Models and the Nested Chinese Restaurant Process." In: *NIPS* (2003). URL: https://papers.nips.cc/paper/2466-hierarchical-topic-models-and-the-nested-chinese-restaurant-process.pdf.

[6] G. Brajnik, G. Guida, and C. Tasso. "User modeling in expert man-machine interfaces: a case study in intelligent information retrieval". In: *IEEE Transactions on Systems, Man, and Cybernetics* 20.1 (1990), pp. 166–185. ISSN: 00189472. DOI: 10.1109/21.47819. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=47819.

[7] Brandon Butterworth. *History of the 'BBC Redux' project*. 2008. URL: http://www.bbc.co.uk/blogs/legacy/bbcinternet/2008/10/history_of_the_bbc_redux_proje.html.

[8] Jonathan Chang et al. "Reading Tea Leaves: How Humans Interpret Topic Models." In: *NIPS* (2009). URL: https://papers.nips.cc/paper/3700-reading-tea-leaves-how-humans-interpret-topic-models.pdf.

[9] Pedro Domingos and Matthew Richardson. "Markov Logic: A Unifying Framework for Statistical Relational Learning". In: *Statistical Relational Learning*. MIT Press, 2007. URL: http://books.google.com/books?hl=en\&lr=\&id=lSkIewOw2WoC\&oi=fnd\&pg=PA339\&dq=Markov+Logic+:+A+Unifying+Framework+for+Statistical+Relational+Learning\&ots=T0xES3dmt_\&sig=DZFlAuhpdfyyY-Z46U6Wq1mktjc.

[10] Thomas Fischer. "Relational Learning and Optimization in the Semantic Web". In: *Tagungsband zum 14. Interuniversitären Doktorandenseminar Wirtschaftsinformatik* (2011), pp. 1–10. URL: http://www.qucosa.de/fileadmin/data/qucosa/documents/7073/Tagungsband_gesamt.pdf?origin=publication_detail\#page=16.

[11] A Fujii et al. "Selective sampling for example-based word sense disambiguation". In: *Computational Linguistics* (1998), pp. 2–12. URL: http://dl.acm.org/citation.cfm?id=972766.

[12] Lisa Getoor and Ben Taskar. *Introduction to Statistical Relational Learning*. The MIT Press, 2007, p. 608.

[13] Diane Kelly and NJ Belkin. "Reading Time, Scrolling and Interaction: Exploring Implicit Sources of User Preferences for Relevance Feedback". In: *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. 2001, pp. 408–409. URL: http://dl.acm.org/citation.cfm?id=384045.

[14] Kristian Kersting and Luc De Raedt. "Bayesian Logic Programs". In: *CoRR* (2001). arXiv: 0111058 [arXiv:cs]. URL: http://www.ke.tu-darmstadt.de/m/lehre/archiv/ws0910/ml-sem/Brech-Mark-BayesianLP.pdf.

[15] Jurgen Koenemann and NJ Belkin. "A case for interaction: A study of interactive information retrieval behaviour and effectiveness". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (1996), pp. 205–212. URL: http://dl.acm.org/citation.cfm?id=238487.

[16] Ken Lang. "NewsWeeder : Learning to Filter Netnews". In: *Proceedings of the 12th International Machine Learning Conference (ML95)*. 1995. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.85.7363.

[17] A McCallum and K Nigam. "Employing EM and Pool-Based Active Learning for Text Classification." In: *ICML* (1998). URL: http://www-connex.lip6.fr/~amini/RelatedWorks/McC98.pdf.

[18] Jennifer Neville and David Jensen. "Relational dependency networks". In: *The Journal of Machine Learning Research* 8 (2007), pp. 653–692. URL: http://dl.acm.org/citation.cfm?id=1248683.

[19] Fredrik Olsson. *A literature survey of active machine learning in the context of natural language processing*. 2009. URL: http://eprints.sics.se/3600/.

[20] Fredrik Olsson. *Bootstrapping Named Entity Annotation by Means of Active Machine Learning*. 2008. ISBN: 9789187850370.

[21] Yves Raimond and Tristan Ferne. *The BBC World Service Archive Prototype*. 2013. URL: http://challenge.semanticweb.org/2013/submissions/swc2013_submission_5.pdf.

[22] Yves Raimond and Chris Lowis. "Automated interlinking of speech radio archives." In: *LDOW* (2012). URL: http://events.linkeddata.org/ldow2012/papers/ldow2012-paper-11.pdf.

[23] AM Rashid, Istvan Albert, and Dan Cosley. "Getting to know you: learning new user preferences in recommender systems". In: *IUI '02 Proceedings of the 7th international conference on Intelligent user interfaces*. 2002, pp. 127–134. ISBN: 1581134592. URL: http://dl.acm.org/citation.cfm?id=502737.

[24] Neil Rubens, Dain Kaplan, and Masashi Sugiyama. "Active learning in recommender systems". In: *Recommender Systems Handbook* (2011), pp. 1–31. URL: http://link.springer.com/chapter/10.1007/978-0-387-85820-3_23.

[25] Burr Settles. *Active Learning*. Vol. 6. 1. June 2012, pp. 1–114. ISBN: 9781608457250. DOI: 10.2200/S00429ED1V01Y201207AIM018. URL: http://www.morganclaypool.com/doi/abs/10.2200/S00429ED1V01Y201207AIM018.

[26] Burr Settles. *Active learning literature survey*. 2010. URL: http://csis.bits-pilani.ac.in/faculty/goel/course_material/MachineLearning/2013/ReadingMaterial/settles.activelearning.pdf.

[27] Burr Settles. *Curious Machines: Active Learning with Structured Instances*. 2008. URL: http://books.google.com/books?hl=en\&lr=\&id=dZ-yjXxoZLsC\&oi=fnd\&pg=PR2\&dq=Curious+Machines:+Active+Learning+with+Structured+Instances\&ots=AvZS9tWKb9\&sig=BEg6iedg7cTsfFpbM3GG0htfa28.

[28] Burr Settles, Mark Craven, and Lewis Friedland. "Active learning with real annotation costs". In: *Proceedings of the NIPS Workshop on Cost-Sensitive Learning*. 2008, pp. 1–10. URL: http://www.cs.cmu.edu/~bsettles/pub/settles.nips08ws.pdf.

[29] Matt Shearer. *BBC News Lab: Linked data*. 2013. URL: http://www.bbc.co.uk/blogs/internet/posts/BBC-News-Lab.

[30] VS Sheng, Foster Provost, and PG Ipeirotis. "Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers". In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (2008), pp. 614–622. URL: http://dl.acm.org/citation.cfm?id=1401965.

[31] CA Thompson, ME Califf, and RJ Mooney. "Active Learning for Natural Language Parsing and Information Extraction". In: *ICML* (1999), pp. 1–16. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.52.597\&rep=rep1\&type=pdf.

[32] Simon Tong and Daphne Koller. "Support Vector Machine Active Learning with Applications to Text Classification". In: *The Journal of Machine Learning Research* (2002), pp. 45–66. URL: http://dl.acm.org/citation.cfm?id=944793.