Lifelong Machine Learning for Topic Modeling and Beyond

Zhiyuan Chen
Department of Computer Science
University of Illinois at Chicago
czyuanacm@gmail.com

Abstract

Machine learning has been popularly used in numerous natural language processing tasks. However, most machine learning models are built using a single dataset. This is often referred to as one-shot learning. Although this one-shot learning paradigm is very useful, it will never make an NLP system understand the natural language because it does not accumulate knowledge learned in the past and make use of the knowledge in future learning and problem solving. In this thesis proposal, I first present a survey of *lifelong machine learning* (LML). I then narrow down to one specific NLP task, i.e., topic modeling. I propose several approaches to apply lifelong learning idea in topic modeling. Such capability is essential to make an NLP system versatile and holistic.

1 Introduction

Machine learning serves as a prevalent approach for research in many natural language processing tasks. However, most of existing machine learning approaches are built using a single dataset, which is often referred to as one-shot learning. This kind of one-shot approach is useful but it does not usually perform well to various datasets or tasks. The main shortcoming of such one-short approach is the lack of continuous learning ability, i.e., learning and accumulating knowledge from past tasks and leveraging the knowledge for future tasks and problem solving in a lifelong manner.

To overcome the above shortcoming, *lifelong machine learning* (LML) has attracted researchers' attention. The term was initially introduced in 1990s (Thrun, 1995, Caruana, 1997). LML aims to design and develop computational systems and algorithms that learn as humans do, i.e., retaining the results learned in the past, abstracting knowledge from them, and using the knowledge to help future learning. The motivation is that when faced with a new situation, we humans always use our previous experience and learned knowledge to help deal with and learn from the new situation, i.e., we learn and accumulate knowledge continuously. The same rationale can be applied to computational models. When a model is built using a single dataset for a task, its performance is limited. However, if the model sees more datasets from the same or similar tasks, it should be able to adjust its learning algorithm for better performance. There are four components in a LML framework: knowledge representation, knowledge extraction, knowledge transfer, and knowledge retention and maintenance. These components are closely connected. I will illustrate each component using examples from topic modeling in Section 3.

Compared to the significant progress of machine learning theory and algorithm, there is relatively little study on lifelong machine learning. One of the most notable works is Never-Ending Language Learner (NELL) (Carlson et al., 2010) which was proposed to extract or read information from the web to expand the knowledge base in an endless manner, aiming to achieve better performance in each day than the previous day. Recently, we proposed **lifelong Topic Modeling** (**LTM**) that extracts knowledge from topic modeling results of many domains and utilizes the knowledge to generate coherent topics in the new domains (Chen and Liu, 2014b). In (Ruvolo and Eaton, 2013), the authors proposed a method that tackles online multi-task learning in the lifelong learning setting. Some other LML related works include (Silver, 2013, Raina et al., 2007, Pentina and Lampert, 2014, Kamar et al., 2013, Kapoor and Horvitz, 2009). Note that LML is different from transfer learning which usually considers one single source domain where the knowledge is coming from and one target domain where the knowledge is applied on (Pan and Yang, 2010).

Proceedings of NAACL-HLT 2015 Student Research Workshop (SRW), pages 133–139,
Denver, Colorado, June 1, 2015.

In this thesis proposal, I narrow down the scope and focus on LML in topic modeling. Topic modeling has been successfully applied to extract semantic topics from text data. However, the majority of existing topic models (one exception is the LTM model mentioned before) belong to the one-shot approach, i.e., they are proposed to address a specific problem without any knowledge accumulation. To leverage the idea of LML, I propose several new approaches to advance topic modeling. I believe that the proposed approaches can significantly advance LML in topic modeling. More broadly, this thesis proposal aims to encourage the community to apply LML in a variety of NLP tasks.

This thesis proposal makes the following three contributions:

1. It studies and discusses lifelong machine learning (LML) in natural language processing. It identifies several important components in LML: knowledge representation, knowledge extraction, knowledge transfer, knowledge retention and maintenance. As there is relatively little study on LML compared to classic machine learning, I believe this thesis proposal will shed some light on the area and encourage the NLP community to advance the area of LML.
2. It reviews the LTM model and discusses the model in terms of LML components. In each component, the model mechanism as well as the shortcomings are discussed.
3. It proposes several new approaches to improve LML in the context of topic modeling. It proposes to enrich the knowledge representation, address knowledge conflicts, select domains and make the algorithm scalable. It further proposes new evaluation frameworks for LTM.

2 Background of Topic Modeling

Topic modeling, such as LDA (Blei et al., 2003) and pLSA (Hofmann, 1999), have been popularly used in many NLP tasks such as opinion mining (Chen et al., 2014), machine translation (Eidelman et al., 2012), word sense disambiguation (Boyd-Graber et al., 2007), phrase extraction (Fei et al., 2014) and information retrieval (Wei and Croft, 2006). In general, topic models assume that each document is a multinomial distribution over topics, where each topic is a multinomial distribution over words. The two types of distributions in topic modeling are document-topic distributions and topic-word distributions respectively. The intuition is that words are more or less likely to be present given the topics of a document. For example, "sport" and "player" will appear more often in documents about sports, "rain" and "cloud" will appear more frequently in documents about weather.

My work is mainly related to knowledge-based topic models (Chen and Liu, 2014a, Andrzejewski et al., 2009) which incorporate different types of prior knowledge into topic models. Supervised label information was considered in (Blei and McAuliffe, 2010, Ramage et al., 2009). Some works also enable the user to specify prior knowledge as seed words/terms for some topics (Mukherjee and Liu, 2012). Interactive topic modeling was proposed in (Hu et al., 2011) to improve topics with the interactive help from the user. However, these works require labeled data or user manual guidance while my proposed approaches do not.

3 Lifelong Topic Modeling

This section introduces the LTM model (Chen and Liu, 2014b). It first presents the overall algorithm of LTM. Then it reviews the model using the four components in the LML framework: knowledge representation, knowledge extraction, knowledge transfer, and knowledge retention and maintenance.

3.1 Overall Algorithm

The basic idea of LTM is that it extracts knowledge from the topic results obtained by topic models in the previous domains or tasks. The knowledge should reflect the correct semantic relationship by investigating different topic model results. By exploiting such knowledge, the LTM model can generate more coherent topics. It consists of 3 main steps:

1. Given a set of document corpora $D = \{D_1, \ldots, D_n\}$ from n domains, LTM runs a topic model (e.g., LDA) on each $D_i \in D$ to produce a set of topics S_i. Such topics are called the *prior topics* (or *p-topics* for short), forming the *topic base* in LTM.
2. A set of *pk-sets* (prior knowledge sets) K are mined from all the p-topics $S = \cup_i S_i$ in the topic

base. The *knowledge base* in LTM is composed of such pk-sets.

3. The knowledge, i.e., pk-sets K, is used in LTM to generate topics for a test document collection D^t (D^t may or may not be from D).

3.2 Knowledge Representation

The prior knowledge set (pk-sets) K for LTM is represented by *must-links*, i.e., if a pair of words form a must-link, they are more likely to belong to the same topic. For example, words "price" and "expensive" can form a must-link. Such knowledge representation is also used in other topic models such as (Andrzejewski et al., 2009). However, they did not model in the lifelong setting. The must-links indicate a positive semantic relationship while some other existing models (Chen and Liu, 2014a, Andrzejewski et al., 2009) also used the negative relationship called *cannot-links*. Cannot-links express that two words do not share the semantic meaning, e.g., words "price" and "beauty". Note that for topic modeling, semantics related knowledge is mostly beneficial as topic modeling tries to group words into topics with different semantics.

3.3 Knowledge Extraction

To extract pk-sets from all the prior topics (Step 2 in Section 3.1, LTM utilizes frequent itemset mining (FIM) (Agrawal and Srikant, 1994). The goal of FIM is to identify all itemsets (an itemset is a set of items) that satisfy some user-specified frequency threshold (also called minimum support) in a set of transactions. The identified itemsets are called frequent itemsets. In the context of LTM, an item is a word and an itemset is a set of words. Each transaction consists of the top words in a past topic. Note that top words ranked by the topic-word distribution from topic modeling are more likely to represent the true semantics embedded in the latent topic. The frequent itemsets of length 2 are used as pk-sets. The rationale for using frequency-based approach is that a piece of knowledge is more reliable when it appears frequent in the prior topics.

3.4 Knowledge Transfer

For topic modeling, Gibbs sampling is a popular inference technique (Griffiths and Steyvers, 2004). The Gibbs sampler for LDA corresponds to the *simple Pólya urn* (*SPU*) model (Mimno et al., 2011). In SPU, a ball of a color (each color denotes each word) is randomly drawn from an urn (each urn corresponds to each topic) and then two balls of the same color are put back into the urn. It increases the probability of seeing a ball of the drawn color in the future, which is known as "the rich get richer".

LTM instead uses the *generalized Pólya urn* (*GPU*) model (Mahmoud, 2008). The difference is that after sampling the ball of a certain color, two balls of that color are put back along with a certain number of balls of some other colors. This flexibility is able to change the probability of multiple colors in each sampling step. Based on the GPU model, LTM increases the probabilities of both words in a pk-set when seeing either of them. For example, given the pk-set {price, expensive}, seeing word "price" under topic t will increase the probability of seeing word "expensive" under topic t; and vice versa. In other words, word "price" promotes word "expensive" under topic t. The extent of promotion of words is determined by the promotion scale parameter μ. This mechanism can transfer the information from the knowledge to the topics generated by LTM.

Since the knowledge is automatically extracted, to ensure the knowledge quality, LTM proposes two additional mechanisms. First, for each topic in the current domain, it uses KL-Divergence to find the matched topics from the topic base. Note that in topic modeling, a topic is a distribution over words. In addition, LTM proposes to use Pointwise Mutual Information (PMI) to estimate the correctness of the knowledge towards the current task/domain. The intuition is that if a piece of knowledge, i.e., must-link, is appropriate, both words in the must-link should have reasonable occurrences in the corpus of the current domain, which means the PMI value of both words is positive. On the other hand, a non-positive PMI value indicates little or no semantic correlation, and thus making the knowledge unreliable.

3.5 Knowledge Retention and Maintenance

LTM simply retains knowledge by adding the topics of a new domain into the topic base which contains all prior topics (Step 1 in Section 3.1). Then, the knowledge is extracted from the new topic base by using FIM mentioned in Section 3.3. There is no

knowledge maintenance.

4 Shortcomings of LTM

This section presents the shortcomings of LTM that corresponds to each of the four LML components.

4.1 Knowledge Representation

There are two shortcomings in terms of knowledge representations in LTM:

1. Since must-links only contain two words, the information contained is limited. The knowledge in the form of sets (containing multiple words) may be more informative.
2. The knowledge does not have a confidence value. The prior knowledge is represented and treated equally. Due to the different frequency of each piece of knowledge (i.e., each pk-set), there should be an additional value indicating confidence attached to each pk-set.

4.2 Knowledge Extraction

Knowledge extraction in LTM also has two main shortcomings:

1. The frequent itemset mining (FIM) used in LTM only extracts frequent itemsets that appear more than a uniformed support threshold. However, due to the power law distribution of natural language (Zipf, 1932), only a small portion of words in the vocabulary appears very frequently while the majority of words are relatively rare. Since the frequencies of words are different, the corresponding support threshold should also be distinct.
2. FIM cannot easily produce knowledge of richer forms. For example, as mentioned above, each piece of knowledge should contain an additional value, e.g., confidence. It is unclear how FIM generates such value, especially if the value needs to be a probability.

4.3 Knowledge Transfer

The shortcoming here is that depending on the promotion scale parameter μ set by the user (Section 3.4), the GPU model may over-promote or under-promote the words in the pk-sets. That means that if the promotion scale parameter μ is set too low, the knowledge may not influence the topics much. In contrast, if this parameter is set too high, the words in the knowledge may dominate the topics resulting inscrutable topics. So the manual setting of this parameter requires expertise from the user.

4.4 Knowledge Retention and Maintenance

Since LTM does not focus on this component, it has three main issues:

1. It is unclear how to retrieve knowledge efficiently when the number of prior topics is huge. This issue is ignored in the LTM model.
2. How a user interacts with the knowledge base (i.e., pk-sets) to improve the quality of knowledge base is also unknown. Since the knowledge is automatically extracted in LTM, the assistance from human beings should contribute to improving the quality of the knowledge base.
3. If the time factor is considered, the new added topics in the topic base may better represent emerging topics while old prior topics may not fit the new tendency anymore. In that case, the knowledge base should weight the new topics more than old topics.

5 Proposed Approaches

The previous section pointed out the shortcomings of LTM. In this section, I propose several approaches to address some of them. Additional strategies are proposed to deal with issues beyond the knowledge components.

5.1 Expanding Knowledge Base

As mentioned above, each piece of knowledge in the knowledge base (i.e., pk-set) is stored and treated equally. However, a piece of knowledge may be more reliable if it gets supports from a large number of domains or it is extracted from the domains or data of higher quality with less noise. In such case, it is more informative to assign a value to the knowledge to indicate its confidence. I propose to add this additional value to each piece of knowledge in the knowledge base. The value is obtained from the normalized support of the knowledge, i.e., the normalized frequency of the knowledge in multiple domains. This expansion can also benefit the knowledge estimation part because the confidence field can provide the prior information to the model for knowledge filtering and estimation.

Another useful expansion is to consider cannot-links with confidence value (Chen and Liu, 2014a, Andrzejewski et al., 2009). Cannot-links express the negative semantic relationship between words, which can lead the model to separate them in different topics. Same as for must-links, cannot-links can also be attached with a confidence value, indicating its prior reliability.

5.2 Knowledge Conflict

After expanding the knowledge base, knowledge retention and maintenance needs additional attention. As we know, must-links express positive semantic correlations while cannot-links express the negative correlations, which means must-links and cannot-links are completely exclusive. Apparently, two words can form a must-link or cannot-link, but not both. The extracted knowledge can contain noise due to 3 reasons below:

1. The corpora which topic models are built on contain noise. This becomes a more serious problem if the corpora are coming from social media with informal languages.
2. Topic modeling is an unsupervised learning method and thus it can generate illogical topics containing words without any semantic correlation. Such topics will then produce incorrect knowledge.
3. The knowledge extraction step is not perfect either. The knowledge extracted using frequency-based FIM approach may include noisy must-links as some words are very frequent that their pairs can also pass the support threshold and form must-links.

The noise in knowledge base means that the newly extracted knowledge may have conflict with the ones in knowledge base. For example, the knowledge base contains the must-link $\{A,\ B\}$. However, the new knowledge contains cannot-link $\{A, B\}$. In such a case, we should not simply merge such knowledge into the knowledge base as it will make the knowledge base nonsensical. It requires us to propose a new strategy when such conflict happens. I propose two approaches to deal with the above situations:

1. Leverage the confidence assigned to each piece of knowledge. Intuitively, when a must-link and a cannot-link forms a conflict, the knowledge base should remain the type of knowledge (must-link or cannot-link) if its confidence is significantly higher than the conflicted one. By doing so, I make sure that the knowledge base does not contain conflicted knowledge and the knowledge piece in the knowledge base has the highest confidence among its conflicted ones.
2. If the confidence is same or similar between two types of knowledge having conflicts, I use the words that share must-links to make the decision. Let us say the must-link is $\{A,\ B\}$, I denote the set of words in which each word shares a must-link with A (or B) as S_A (or S_B). Then I use the overlapping percentage of S_A and S_B as estimation that how likely words A and B share the positive semantic correlation. This is intuitive since if words A and B are truly semantically correlated, they should share a lot of words in their must-links. For instance, words "price" and "expensive" can form must-links with words such as "cheap", "cost", "pricy", etc.

5.3 Domain Selection

I also notice an important issue that LTM struggles with, i.e., LTM uses all the domains as the source from which the knowledge is extracted. In other words, LTM assumes all the domains are relevant and helpful to the current domain. However, this assumption may not always hold. For example, the topics from the domain "Politics" may not contribute much to the domain "Laundry" as they are very different in terms of both word usage and word semantics. Simply using all the domains as LTM has two major drawbacks:

1. The knowledge extracted from all the domains may contain some inappropriate knowledge towards a particular domain. Although LTM has a mechanism to estimate and filter knowledge, it is still not perfect. For a more effective knowledge transfer, a domain selection step is indispensable to make sure the knowledge is more relevant and beneficial.
2. Extracting knowledge from all the domains can be time-consuming given a huge number of domains. Many of the extracted knowledge is useless as a particular domain only contains a limited set of words. So domain selection can also improve the knowledge extraction efficiency.

To select domains, I propose to measure the domain distance by utilizing JS-Divergence. Given two distributions P and Q, JS-Divergence between them is defined as below:

$$JS(P,Q) = \frac{1}{2}KL(P,M) + \frac{1}{2}KL(Q,M) \quad (1)$$

$$M = \frac{1}{2}(P+Q) \quad (2)$$

$$KL(P,Q) = \sum_i ln\left(\frac{P(i)}{Q(i)}\right)P(i) \quad (3)$$

Since each topic produced by topic models is a distribution over words, I can use JS-Divergence to measure the distance between topics. The problem is defined as given two domains D_1 and D_2, the goal is to estimate the domain distance by estimating their corresponding topic distance. I propose the following algorithm: for each topic t in domain D_1, I find the most similar topic (say t') in domain D_2 that has the smallest JS-Divergence with t. I denote this smallest JS-Divergence by $e(t)$. Then, the distance between domain D_1 and domain D_2 is defined as below:

$$DIST(D_1,D_2) = \sum_{t\in D_1} e(t) + \sum_{t'\in D_2} e(t') \quad (4)$$

Note that to make the distance symmetric, I calculate function $e()$ for each topic in domain D_1 as well as domain D_2. After the domain distance is calculated, given a new domain D', I can rank all existing domains by Equation 4 and pick up top K most relevant domains.

5.4 Scalability

In this sub-section, I also consider the scalability issue. There are generally 2 bottlenecks in LTM.

The first one is frequent itemset mining (FIM). There are some proposed scalable versions of FIM such as (Chester et al., 2009, Moens et al., 2013).

The second one is Gibbs sampling in topic models. Gibbs sampling (Griffiths and Steyvers, 2004) is a popular inference technique for topic modeling. However, it is not scalable to large datasets as it needs to make pass over the corpus many times. Some promising frameworks have been proposed (Yao et al., 2009, Zhai et al., 2012, Hu et al., 2014) to solve this issue. Since the GPU model used in LTM is a natural extension to that in LDA, these proposed methods are also applicable to LTM.

6 Evaluation

This section proposes a new evaluation framework that suits our proposed approaches. In (Chen and Liu, 2014b), the evaluation measurements are Topic Coherence (Mimno et al., 2011) and Precision@n which asks annotators to label both topics and words. A more comprehensive evaluation framework can contain the following two measurements:

1. Knowledge Evaluation. In order to evaluate each piece of knowledge (must-link or cannot-link) in the knowledge base, PMI score of both words using a large standard text corpus (Newman et al., 2010) can be applied. Human annotation can also be used to label the correctness of each piece of knowledge. This is to evaluate the effectiveness of knowledge handling in the model.
2. Domain Evaluation. As mentioned in 5.3, not all the prior domains are suitable to a new domain. It is important to evaluate the model performance by providing different sets of prior domains. There could be three main sets of prior domains for an extensive evaluation: 1) all relevant; 2) all irrelevant; 3) a combination of both. The relevance of domains should be defined by experts that are familiar with these domains.

7 Conclusions

This thesis proposal studied lifelong machine learning in topic modeling. It first introduced lifelong machine learning and its important components. Then, it reviewed the LTM model and pointed out its drawbacks. The corresponding approaches were proposed to address the issues and further advance the problem. For future direction, I would like to further integrate lifelong machine learning in the context of other NLP tasks, such as word sense disambiguation. I believe that the lifelong machine learning capacity is essential to a robust NLP system to overcome the dynamics and complexity of natural language, and for the purpose of a deeper understanding of natural language.

Acknowledgments

This work was supported in part by grants from National Science Foundation (NSF) under grant no. IIS-1111092 and IIS-1407927.

References

Rakesh Agrawal and Ramakrishnan Srikant. 1994. Fast Algorithms for Mining Association Rules. In *VLDB*, pages 487–499.

David Andrzejewski, Xiaojin Zhu, and Mark Craven. 2009. Incorporating Domain Knowledge into Topic Modeling via Dirichlet Forest Priors. In *ICML*, pages 25–32.

David M. Blei and Jon D. McAuliffe. 2010. Supervised Topic Models. In *NIPS*, pages 121–128.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Jordan L Boyd-Graber, David M. Blei, and Xiaojin Zhu. 2007. A Topic Model for Word Sense Disambiguation. In *EMNLP-CoNLL*, pages 1024–1033.

Andrew Carlson, Justin Betteridge, and Bryan Kisiel. 2010. Toward an Architecture for Never-Ending Language Learning. In *AAAI*, pages 1306–1313.

Rich Caruana. 1997. Multitask Learning. *Machine learning*, 28(1):41–75.

Zhiyuan Chen and Bing Liu. 2014a. Mining Topics in Documents : Standing on the Shoulders of Big Data. In *KDD*, pages 1116–1125.

Zhiyuan Chen and Bing Liu. 2014b. Topic Modeling using Topics from Many Domains, Lifelong Learning and Big Data. In *ICML*, pages 703–711.

Zhiyuan Chen, Arjun Mukherjee, and Bing Liu. 2014. Aspect Extraction with Automated Prior Knowledge Learning. In *ACL*, pages 347–358.

Sean Chester, Ian Sandler, and Alex Thomo. 2009. Scalable apriori-based frequent pattern discovery. In *CSE*, pages 48–55.

Vladimir Eidelman, Jordan Boyd-Graber, and Philip Resnik. 2012. Topic Models for Dynamic Translation Model Adaptation. In *ACL*, pages 115–119.

Geli Fei, Zhiyuan Chen, and Bing Liu. 2014. Review Topic Discovery with Phrases using the Pólya Urn Model. In *COLING*, pages 667–676.

Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *PNAS*, 101 Suppl:5228–5235.

Thomas Hofmann. 1999. Probabilistic Latent Semantic Analysis. In *UAI*, pages 289–296.

Yuening Hu, Jordan Boyd-Graber, and Brianna Satinoff. 2011. Interactive Topic Modeling. In *ACL*, pages 248–257.

Yuening Hu, Ke Zhai, Vladimir Edelman, and Jordan Boyd-Graber. 2014. Polylingual Tree-Based Topic Models for Translation Domain Adaptation. In *ACL*, pages 1166–1176.

Ece Kamar, Ashish Kapoor, and Eric Horvitz. 2013. Lifelong Learning for Acquiring the Wisdom of the Crowd. In *IJCAI*, pages 2313–2320.

Ashish Kapoor and Eric Horvitz. 2009. Principles of lifelong learning for predictive user modeling. In *User Modeling*, pages 37–46.

Hosam Mahmoud. 2008. *Polya Urn Models*. Chapman & Hall/CRC Texts in Statistical Science.

David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *EMNLP*, pages 262–272.

Sandy Moens, Emin Aksehirli, and Bart Goethals. 2013. Frequent Itemset Mining for Big Data. In *IEEE International Conference on Big Data*, pages 111–118.

Arjun Mukherjee and Bing Liu. 2012. Aspect Extraction through Semi-Supervised Modeling. In *ACL*, pages 339–348.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *HLT-NAACL*, pages 100–108.

Sinno Jialin Pan and Qiang Yang. 2010. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.*, 22(10):1345–1359.

Anastasia Pentina and Christoph H Lampert. 2014. A PAC-Bayesian Bound for Lifelong Learning. In *ICML*, pages 991–999.

Rajat Raina, Alexis Battle, Honglak Lee, Benjamin Packer, and Andrew Y Ng. 2007. Self-taught Learning : Transfer Learning from Unlabeled Data. In *ICML*, pages 759–766.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. 2009. Labeled LDA : A supervised topic model for credit attribution in multi-labeled corpora. In *EMNLP*, pages 248–256.

Paul Ruvolo and Eric Eaton. 2013. ELLA: An efficient lifelong learning algorithm. In *ICML*, pages 507–515.

Daniel L Silver. 2013. On Common Ground: Neural-Symbolic Integration and Lifelong Machine Learning. In *9th International Workshop on Neural-Symbolic Learning and Reasoning NeSy13*, pages 41–46.

Sebastian Thrun. 1995. Lifelong Learning: A Case Study. Technical report.

Xing Wei and W Bruce Croft. 2006. LDA-based document models for ad-hoc retrieval. In *SIGIR*, pages 178–185.

Limin Yao, David Mimno, and Andrew McCallum. 2009. Efficient methods for topic model inference on streaming document collections. In *KDD*, pages 937–946.

Ke Zhai, Jordan Boyd-Graber, Nima Asadi, and Mohamad L. Alkhouja. 2012. Mr. LDA: a Flexible Large Scale Topic Modeling Package using Variational Inference in MapReduce. In *WWW*, pages 879–888.

George Kingsley Zipf. 1932. *Selected Papers of the Principle of Relative Frequency in Language*. Harvard University Press.