Semi-Unsupervised Lifelong Learning for Sentiment Classification

Xianbin Hong
Xianbin.Hong@liverpool.ac.uk
Research Institute of Big Data Analytics, Xi'an Jiaotong-Liverpool University
Suzhou, Jiangsu, China

Gautam Pal
Gautam.Pal@xjtlu.edu.cn
Research Institute of Big Data Analytics, Xi'an Jiaotong-Liverpool University
Suzhou, Jiangsu, China

Sheng-Uei Guan
Steven.Guan@xjtlu.edu.cn
Research Institute of Big Data Analytics, Xi'an Jiaotong-Liverpool University
Suzhou, Jiangsu, China

Prudence Wong
PWong@liverpool.ac.uk
Department of Computer Science, The University of Liverpool
Liverpool, Merseyside, UK

Dawei Liu
Dawei.Liu@xjtlu.edu.cn
Research Institute of Big Data Analytics, Xi'an Jiaotong-Liverpool University
Suzhou, Jiangsu, China

Ka Lok Man
Ka.Man@xjtlu.edu.cn
Research Institute of Big Data Analytics, Xi'an Jiaotong-Liverpool University
Suzhou, Jiangsu, China

Xin Huang
Xin.Huang@xjtlu.edu.cn
Research Institute of Big Data Analytics, Xi'an Jiaotong-Liverpool University
Suzhou, Jiangsu, China

ABSTRACT

Lifelong machine learning is a novel machine learning paradigm which continually learns tasks and accumulates knowledge for reusing. The knowledge extracting and reusing abilities enable the lifelong machine learning to understand the knowledge for solving a task and obtain the ability to solve the related problems. In sentiment classification, traditional approaches like Naïve Bayes focus on the probability for each words with positive or negative sentiment. However, the lifelong machine learning in this paper will investigate this problem in a different angle and attempt to discover which words determine the sentiment of a review. We will pay all attention to obtain knowledge during learning for future learning rather than just solve a current task.

CCS CONCEPTS

• **Computing methodologies** → *Theory of mind.*

KEYWORDS

lifelong machine learning, sentiment classification

ACM Reference Format:
Xianbin Hong, Gautam Pal, Sheng-Uei Guan, Prudence Wong, Dawei Liu, Ka Lok Man, and Xin Huang. 2019. Semi-Unsupervised Lifelong Learning for Sentiment Classification. In *BDAI 2019: The 2nd International Conference on Big Data and Artificial Intelligence , June 22-24, 2019 ,Guangzhou, China.* ACM, New York, NY, USA, 6 pages.

1 INTRODUCTION

Over the past 30 years, machine learning have achieved a significant development. However, we are still in a era of "weak AI" rather than "strong AI" which due to the algorithms of AI only know how to solve a problem but have no idea why these approaches work and when can reuse them to solve other problems. Hence, the lifelong machine learning (simply said as lifelong learning or "LML" below) [7] was raised to build a new learning paradigm to learn with knowledge accumulation and reusing. With the knowledge, the AI is able to solve new problems totally unsupervised or semi-supervised. Thinking forward, the knowledge discovering and reusing becomes the core learning goal rather than to solve a specific problem under the lifelong learning setting.

For instance, in the sentiment classification we can to predict the sentiment (positive or negative) of a sentence or a document by Naïve Bayes. To solve this problem, we need to know the probability of each word that appears in positive or negative content. For different sentiment classification tasks with different domain, traditional approaches will calculate the probability of each word to be positive or negative in individual domain to achieve a good performance due to one word can be positive in a domain while be negative in another domain. Hence, for each domain we need to collect data for supervised learning. In this way, the algorithm will

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
BDAI 2019, June 22-24, 2019, Guangzhou, China
© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.

never know how to solve a problem without new labeled data and teaching. This is what a typical "weak AI".

To achieve the goal of "strong AI", we need to convert our learning goal to discover which words have sentiment orientation and check whether the orientation just valid in specific domains. If we can achieve this learning goal, the algorithms will be able to solve new tasks without teaching and explore new domain to find some new words with sentiment orientation. Zhiyuan Chen[2] ever proposed a approach to determine which domain dose a word have the sentiment orientation to achieve the goal of lifelong learning. He made a big progress but the supervised learning still is needed. Hence, we will make it forward to let the learning to star with supervised learning but continue with unsupervised learning in the future tasks.

2 LIFELONG MACHINE LEARNING

It was firstly called as lifelong machine learning since 1995 by Thrun [6, 8]. Efficient Lifelong Machine Learning (ELLA) [5] raised by Ruvolo and Eaton. Comparing with the multi-task learning [1], ELLA is much more efficient. Zhiyuan and Bing [2] improved the sentiment classification by involving knowledge. The object function was modified with two penalty terms which corresponding with previous tasks.

2.1 Components of LML

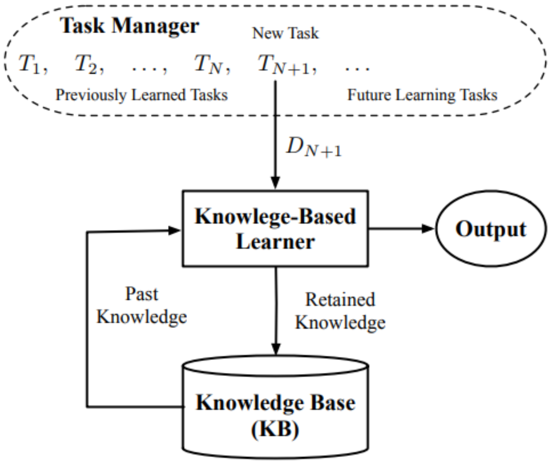

Figure 1: Knowledge System in the Lifelong Machine Learning [2]

The knowledge system contains the following components:

- Knowledge Base (KB): The knowledge Base[2] mainly used to maintain the previous knowledge. Based on the type of knowledge, it could be divided as Past Information Store (PIS), Meta-Knowledge Miner (MKM) and Meta-Knowledge Store (MKS).
- Knowledge Reasoner (KR): The knowledge reasoner is designed to generate new knowledge upon the achieve knowledge by logic inference. A strict logic design is required so the most of the LML algorithms lack of the component.
- Knowledge-Base Learner (KBL): The Knowledge-Based Learner[2] aims to retrieve and transfer previous knowledge to the current task. Hence, it contains two parts: task knowledge miner and leaner. The miner seeks and determines which knowledge could be reused, and the learner transfer such knowledge to the current task.

2.2 Sentiment Classification

Hong and etc.[3] ever discussed that the NLP field is most suitable for the researches of the lifelong learning due to it is easier to extract knowledge and be understood by human. Previous classical paper[2] chose the sentiment classification as the learning target because it is could be regarded as a task as well as a group of subtasks in different domain. These sub-tasks related to each other but a model trained on a domain is unable to perform well in rest domains. The sub-tasked is related means that the knowledge transform among tasks is possible to improve performance. And the distribution of distribution is different requires that our algorithms could know when the knowledge can be used and when can not. Known these, an algorithm can be called as "lifelong" due to it is able to transfer previous knowledge to new tasks to improve performance.

By using Naive Bayes to solve sentiment classification, we need to know the probability of each words that shows in positive or negative content. We also need to know well that some words may only have sentiment orientation in some specific domains. "Lifelong Sentiment Classification" ("LSC" for simple below) [2] records that which domain does a word have the sentiment orientation. If a word always has sentiment orientation or has significant orientation in current domain, a high weight will sign to it more than other words. This approach contains a knowledge transfer operation and a knowledge validation operation.

3 CONTRIBUTION OF THIS PAPER

Although LSC[2] already raised a lifelong approach, it only aims to improve the classification accuracy. It will not deliver a summary that which words are influence sentiment which is most important to us and can be used for future learning. In addition, it still limits in supervised learning and unable to handle the tasks without labeled data.

Our paper advances the lifelong learning in sentiment classification and have two main contributions:

- **We introduce an novel approach to discover and store the words with sentiment orientation for reuse**
- **A improved lifelong learning paradigm is proposed to solve the sentiment classification problem under unsupervised learning setting with previous knowledge**

4 SENTIMENT ORIENTATION WORDS

4.1 Naïve Bayesian Text Classification

In this paper, we define a word has sentiment orientation by calculating the probability that it will appears in a positive or negative content (sentence or document). If a word has high probability with sentiment orientation, it also will leads to the document have higher probability of sentiment orientation based on the Naïve Bayesian (NB) formula.

NB text classification [4] will calculate the probability of each word w given a sentiment orientation (positive or negative). Use use the same formula as LSC[2] used below. $P(w|c_j)$ is the probability of a word appears in a class:

$$P(w|c_j) = \frac{\lambda + N_{c_j,w}}{\lambda\,|V| + \sum_{v=1}^{V} N_{c_j,v}} \tag{1}$$

Where c_j is either positive (+) or negative (-) sentiment orientation. $N_{c_j,w}$ is the frequency of word w in documents of class c_j . |V| is the size of vocabulary V and $\lambda(0 \leqslant \lambda \leqslant 1)$ is used for smoothing (set as 1 for Laplace smoothing in this paper).

Given a document, we can calculate the probability of it for different classes by:

$$P(c_j|d_i) = \frac{P(c_j)\prod_{w\in d_i} P(w|c_j)^{n_w,d_i}}{\sum_{r=1}^{C} P(c_r)\prod_{w\in d_i} P(w|c_r)^{n_w,d_i}} \tag{2}$$

Where d_i is the given document, n_w, d_i is the frequence of a word appears in this document.

To predict the class of a document, we only need to calculate $P(c_+|d_i) - P(c_-|d_i)$. If the difference is lager than 0, the document should be predict as positive orientation:

$$\begin{aligned} P(c_+|d_i) - P(c_+|d_i) = {} & \frac{P(c_+)\prod_{w\in d_i} P(w|c_+)^{n_w,d_i}}{\sum_{r=1}^{C} P(c_r)\prod_{w\in d_i} P(w|c_r)^{n_w,d_i}} - \\ & \frac{P(c_-)\prod_{w\in d_i} P(w|c_-)^{n_w,d_i}}{\sum_{r=1}^{C} P(c_r)\prod_{w\in d_i} P(w|c_r)^{n_w,d_i}} \end{aligned} \tag{3}$$

As we only need to know whether $P(c_+|d_i) - P(c_-|d_i)$ is lager that 0, so the formula could be simplify to:

$$\begin{aligned} P(c_+|d_i) - P(c_+|d_i) = P(c_+) \prod_{w\in d_i} P(w|c_+)^{n_w,d_i} - \\ P(c_-) \prod_{w\in d_i} P(w|c_-)^{n_w,d_i} \end{aligned} \tag{4}$$

4.2 Discover Sentiment Orientation Words

Ideally, if we know the $P(c_+)$, $P(c_-)$ and $P(w|c_j)$ for all words, we can predict the sentiment orientation for all documents. However, above three key components are different in different domains. LSC [2] discussed a possible solution of $P(w|c_j)$. As we known, not all words have sentimental orientation like "a", "one" and etc. while some words always have like "good", "hate", "excellent" and so on. In addition, some words only have sentiment orientation in specific domains. For example, "tough" in reviews of the diamond may mean that the diamond have a good quality while "tough" in the comments for food normally shows that it is hard to chew. Hence, in order to achieve the goal of the lifelong learning. We need to find the words always have sentiment orientation and be careful for those words only shows orientation in specific domains.

5 LIFELONG SEMI-SUPERVISED LEARNING FOR SENTIMENT CLASSIFICATION

Although LSC [2] considered the difference among domains, it still is a typical supervised learning approach.In this paper, we proposed to learn as two stages:

(1) Initial Learning Stage: to explore a basic set of sentiment orientation words. After that, the model should be able to basically classify a new domain with a good performance.
(2) Self-study Stage: Use the knowledge accumulated from the initial stage to handle new domains, also fine-tune and consolidate the knowledge generated from initial learning stage.

5.1 Initial Learning Stage

In this stage, we need to train the model to remember some sentiment orientation words. This requires us to find the words with sentiment orientation in each domain. We need to answer two question here:

(1) How to determine whether a word has orientation?
(2) How much domain do we need for the initial learning stage?

For the first question, we need to find which words mainly show at the positive or negative documents. This means for a word w with positive orientation, $P(+|w) >> P(-|w)$ or $P(+|w) >> P(+)$. In this paper, we will use $O(w) = P(+|w)/P(+)$ to represent the orientation. This because that the $P(c_j|w)/P(w)$ is easy to extend to multi-classes classification problem. According to the Bayesian formula, $P(+|w)/P(+) = P(w|+)/P(w)$.

5.2 Self-study Stage

In this stage, our main task is to explore which words have orientation. We will use the these words to predict the new domains and assign the pseudo-labels to them. With the pseudo labels, we are able to discover the new words with orientation. Following the the procedure for self-study:

(1) Using the orientation words accumulate from previous tasks to predict a new domain, and assign the prediction result as pseudo labels.
(2) Using the reviews and pseudo labels the new domain as new training data to run Naïve model.
(3) Update the orientation words knowledge base.

6 EXPERIMENT

6.1 Datasets

In the experiment, we use the same datasets as LSC [2] used. It contains the reviews from 20 domains crawled from Amazon.com and each domain has 1,000 reviews (the distribution of positive and negative reviews is imbalanced).

6.2 Word Orientation Analysis

To answer the first question for the initial learning stage, we need to know which words exactly influence the sentiment classification.

F1 Score Percentage Datasets	100%	90%	80%	70%	60%	50%	40%	30%	20%	10%
AlarmClock	0.8342	0.8342	0.8342	0.8342	0.8342	0.8342	0.8342	0.8342	**0.8342**	0.4425
Baby	0.6814	0.6814	0.6814	0.6814	0.6814	0.6814	0.6814	0.6814	**0.6814**	0.2411
Bag	0.7236	0.7236	0.7236	0.7236	0.7236	0.7236	0.7236	0.7236	**0.7236**	0.3706
CableModem	0.6462	0.6462	0.6462	0.6462	0.6462	0.6462	0.6462	0.6462	**0.6462**	0.2997
Dumbbell	0.7395	0.7395	0.7395	0.7395	0.7395	0.7395	0.7395	0.7395	**0.7395**	0.198
Flashlight	0.6398	0.6398	0.6398	0.6398	0.6398	0.6398	0.6398	0.6398	**0.6398**	0.3113
Gloves	0.6837	0.6837	0.6837	0.6837	0.6837	0.6837	0.6837	0.6837	**0.6837**	0.322
GPS	0.7503	0.7503	0.7503	0.7503	0.7503	0.7503	0.7503	0.7503	**0.7503**	0.3308
GraphicsCard	0.7287	0.7287	0.7287	0.7287	0.7287	0.7287	0.7287	0.7287	**0.7287**	0.4118
Headphone	0.7289	0.7289	0.7289	0.7289	0.7289	0.7289	0.7289	0.7289	**0.7289**	0.4137
HomeTheaterSystem	0.8631	0.8631	0.8631	0.8631	0.8631	0.8631	0.8631	0.8631	**0.8631**	0.2912
Jewelry	0.7065	0.7065	0.7065	0.7065	0.7065	0.7065	0.7065	0.7065	**0.7065**	0.5016
Keyboard	0.7161	0.7161	0.7161	0.7161	0.7161	0.7161	0.7161	0.7161	**0.7161**	0.2432
MagazineSubscriptions	0.8082	0.8082	0.8082	0.8082	0.8082	0.8082	0.8082	0.8082	0.8082	**0.8082**
MoviesTV	0.6916	0.6916	0.6916	0.6916	0.6916	0.6916	0.6916	0.6916	**0.6916**	0.6576
Projector	0.7604	0.7604	0.7604	0.7604	0.7604	0.7604	0.7604	0.7604	**0.7604**	0.3183
RiceCooker	0.7932	0.7932	0.7932	0.7932	0.7932	0.7932	0.7932	0.7932	**0.7932**	0.1841
Sandal	0.6714	0.6714	0.6714	0.6714	0.6714	0.6714	0.6714	0.6714	**0.6714**	0.3715
Vacuum	0.7842	0.7842	0.7842	0.7842	0.7842	0.7842	0.7842	0.7842	**0.7842**	0.2095
VideoGames	0.7643	0.7643	0.7643	0.7643	0.7643	0.7643	0.7643	0.7643	**0.7643**	0.4916
Average	0.7341	0.7341	0.7341	**0.7341**	0.7328	0.7192	0.7125	0.7105	0.7084	0.3601

Table 1: F1 Score of Naïve Bayesian Classifiers under Decreasing Word Usage Percentage

Firstly, we calculate $P(w|+)$ and $P(w|-)$ for each words. Then, we define the orientation by $O(w) = P(w|+)/P(w)$. Finally, we only choose a specific percentage words to predict and see whether the performance decreases. In addition, we also only consider the words that at least show over average 5 times in per domain. This because that we did not delete the symbols and numbers in the data, and these characters may be noise in the training data.

We firstly sorted the words or symbols (no data pre-processing to the corpus in this paper) by the orientation $O(w)$ and then choose a specific percentage words or symbols from the whole words to only 10%. From Table 1 we can see that using more than 70% obtains the best average result. We also noticed that the performance will not significantly decrease until we removed more than 80% of words and symbols. This means that the most of words and symbols do not have obvious sentiment orientation.

Hence, we will only keep 20% of words for Naïve Bayes model and it still keeps around 96% f1 score. Although the performance decrease on a single domain, the better global performance will achieve with the orientation words. In addition

6.3 Requirement for the Initial Learning

For the second question of the initial learning stage, the answer depends on the tasks. In the practice, all of the labeled data definitely need to be used for training. The only question should be conceded is that how much domains is insufficient if there only are a few labeled data. For this sentiment classification task, one domain is absolutely insufficient. Based on the experiment result, the initial learning stage at least needs two domains, and can achieve much better performance with three domains. Increase more domains will not significant influence performance. Hence, three domains is enough for this task. For different tasks, two labeled domains are necessary. More labeled domains are suggested to continue collect until the performance on the new domains tends to steady.

6.4 Self-study Learning

In the self-study learning stage, our assume that we are under unsupervised learning setting. In this stage, there is not any labeled data. Instead of that, we will use the model generate from the initial learning stage to predict each use domain and assign pseudo labels. After that, the model will regards the pseudo labels as real label and continue training on the new domain. With this method, self-study learning stage can learn new domains well without any labeled data.

Table 2 is the F1 score of three models on 17 domains. The first three domains was used for the initial learning stage. And we use the Macro-F1 score because the datasets are imbalanced and it can show the performance on the minor classes. We compared our model (Semi-Unsupervised Learning, SU-LML for short) with Naïve Bayes model which only trained on the first three (source) domains (NB-S) and Naïve Bayes model trained on each domain with labels by 5-fold cross validation (NB-T). We can see that our approach is significantly better than other two approaches. It even perform better than the NB-T, a typically supervised learning. The figure 2 shows the result more clearly.

F1 Score Model Datasets	NB-S	NB-T	SU-LML
CableModem	0.4774	0.6633	**0.8694**
Dumbbell	0.6539	0.764	**0.8748**
Flashlight	0.6536	0.6251	**0.8259**
Gloves	0.5973	0.6943	**0.785**
GPS	0.6447	0.7465	**0.9121**
GraphicsCard	0.4797	0.7346	**0.8768**
Headphone	0.5938	0.7356	**0.8858**
HomeTheaterSystem	0.6242	0.8611	**0.9236**
Jewelry	0.6927	0.7088	**0.7599**
Keyboard	0.6905	0.7289	**0.8707**
MagazineSubscriptions	0.6284	0.8056	**0.8932**
MoviesTV	0.4991	0.6785	**0.8381**
Projector	0.6565	0.7525	**0.8575**
RiceCooker	0.6833	0.8027	**0.8475**
Sandal	0.6972	0.6904	**0.8059**
Vacuum	0.7728	0.8	**0.8992**
VideoGames	0.5665	0.7564	**0.9068**
Average	0.6242	0.7381	**0.8607**

Table 2: F1 Score for NB-S, NB-T, SU-LML

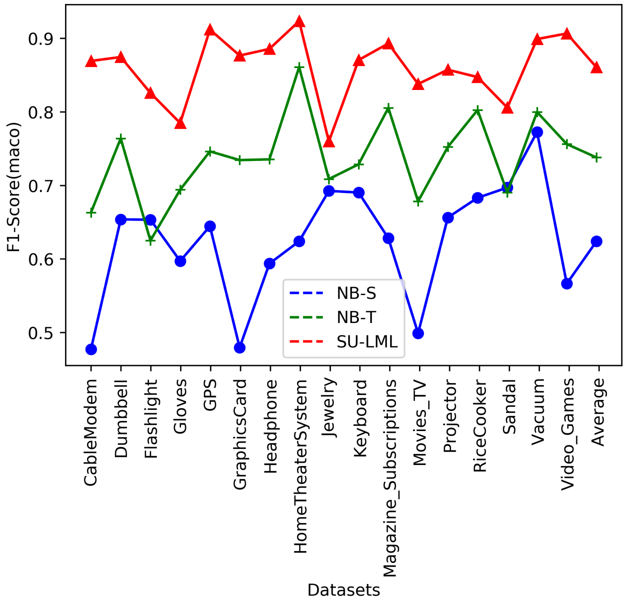

Figure 2: F1 Score in Self-Study Stage

6.5 Knowledge Generated during Learning

In this paper, we do one more important things is that we are learning which words have sentiment orientation. If a word was regarded with sentiment orientation, we increase the orientation score of it plus one. In addition, we will plus an additional score from 0 to 1 to 1 based on the $O(w)$ rank. From table 3, we can see that most top words exactly show negative emotion.

Word	Degree for Negative Sentiment
refund	32.99921813917123
garbage	32.994266353922335
junk	32.985405264529575
waste	32.984102163148286
worst	32.97185301016418
rma	32.96846494657285
poorly	32.96194943966641
terrible	32.95569455303623
disappointed	32.949960906958566
trash	32.948918425853535
useless	32.94683346364347
worthless	32.94057857701329
awful	32.92520198071411
defective	32.917904612978894
return	32.913734688558776
exchange	32.908001042481104
respond	32.90487359916601
poor	32.90409173833724
disappointment	32.90278863695596
crap	32.89653375032577

Table 3: Top 20 Words with Negative Sentiment

7 CONCLUSION AND OUTLOOK

We proposed a semi-unsupervised lifelong sentiment classification approach in this paper. It can accumulate knowledge from previous learning and turn to self-study. Very few labeled data required in our approach so it is very suitable for the industry scenario. The performance of it even exceeds the supervised learning, which shows that the knowledge reusing of the lifelong learning is useful.

Although we only show two classes classification here, but the ideal is also suitable for the multi-classes classification. All text classification can use approach, not only sentiment classification. Our model already know which words have sentiment orientation and can use them to classify, which uses the same approach of our human being. We shows that to focus the goal behind the learning tasks is more meaningful than just to find a solution. We should learn the knowledge and skills for all tasks rather than a solution for one task.

8 ACKNOWLEDGMENTS

This research is supported by the Research Institute of Big Data Analytics, Xi'an Jiaotong – Liverpool University and the CERNET Innovation Project under Grant NGII20161010.

REFERENCES

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

[2] Zhiyuan Chen, Nianzu Ma, and Bing Liu. 2015. Lifelong learning for sentiment classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Vol. 2. 750–756.

[3] Xianbin Hong, Prudence Wong, Dawei Liu, Sheng-Uei Guan, Ka Lok Man, and Xin Huang. 2018. Lifelong Machine Learning: Outlook and Direction. In *Proceedings of the 2nd International Conference on Big Data Research*. ACM, 76–79.

[4] Andrew McCallum and Kamal Nigam. 1999. Text classification by bootstrapping with keywords, EM and shrinkage. *Unsupervised Learning in Natural Language Processing* (1999).

[5] Paul Ruvolo and Eric Eaton. 2013. ELLA: An efficient lifelong learning algorithm. In *International Conference on Machine Learning*. 507–515.

[6] Sebastian Thrun. 1996. Is learning the n-th thing any easier than learning the first?. In *Advances in neural information processing systems*. 640–646.

[7] Sebastian Thrun. 1998. Lifelong learning algorithms. In *Learning to learn*. Springer, 181–209.

[8] Sebastian Thrun and Tom M Mitchell. 1995. Lifelong robot learning. *Robotics and autonomous systems* 15, 1-2 (1995), 25–46.