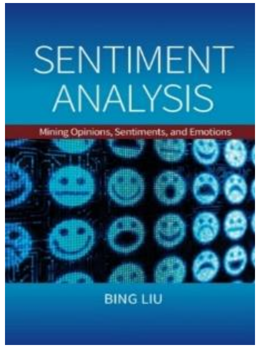
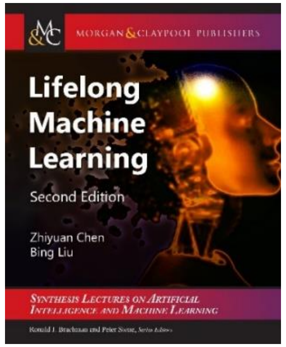
Lifelong Learning and Sentiment Analysis

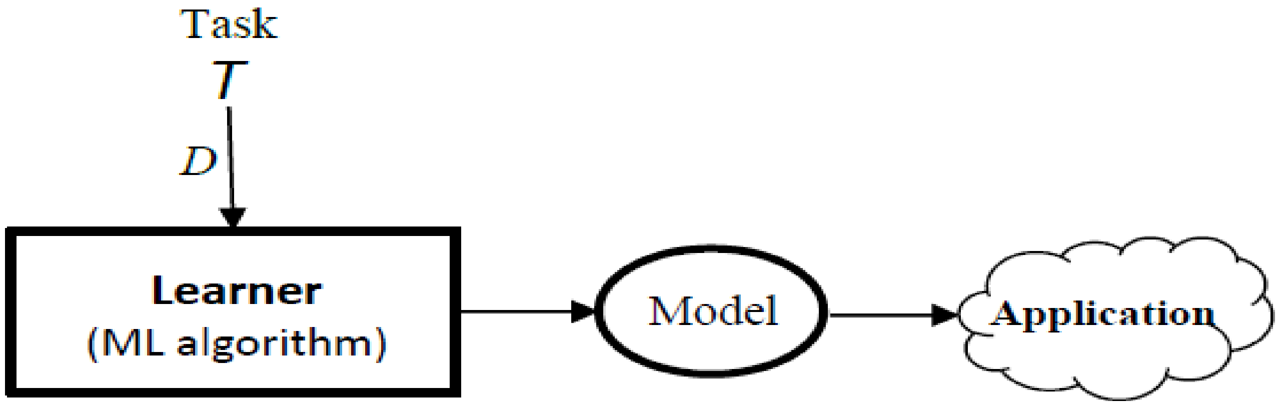




Bing Liu
Department of Computer Science
University of Illinois at Chicago

Introduction

- Learning is key to all human (and AI agent) activities
- Classic machine learning
  - Isolated single-task learning



- Knowledge learned not accumulated or used in new learning
  - Needs a large number of training examples:
    - cannot expect humans to label everything in the world.
  - Suitable for well-defined tasks in restricted and narrow environments

Lifelong learning

(Chen and Liu, 2016, 2018 book)

- Humans never learn in isolation or from scratch. We
  - learn continuously and incrementally,
  - accumulate knowledge learned in the past, and
  - use/adapt the knowledge to learn more & to learn better.

- *Lifelong Learning* (LL): imitate this human learning capability
- **Goal:** Create a machine that learns like humans
  - Without it, hard to achieve Artificial General Intelligence (AGI)
    - (what about *consciousness?*)

Practical applications **need** LL

- Chatbots, personal assistants, self-driving cars, and other physical robots, working in real-life environments need LL.
  - Chatbots will not be intelligent if they cannot learn during chatting
  - Self-driving cars aren't going to fly with only rules or off-line training
  - …
- They face the real dynamic and open world (not closed)
  - They have to continuously and incrementally learn and accumulate knowledge and adapt to new situations in a self-supervised manner.

Outline

- What is lifelong learning?
- Sentiment analysis and scale-up challenges
- Lifelong sentiment classification
- Lifelong topic models for aspect extraction and grouping
- Lifelong label propagation for identifying entities and aspects
- Learning on the job for aspect extraction
- Continual learning, meta-learning and open world learning
- Summary

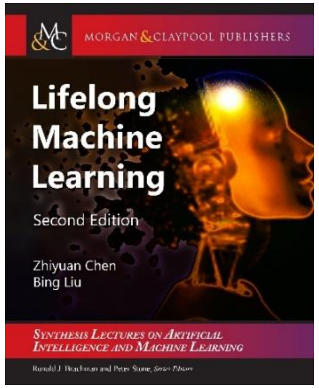
Old definition of lifelong learning (LL)

(Thrun 1995, Silver et al 2013; Chen and Liu, 2016)

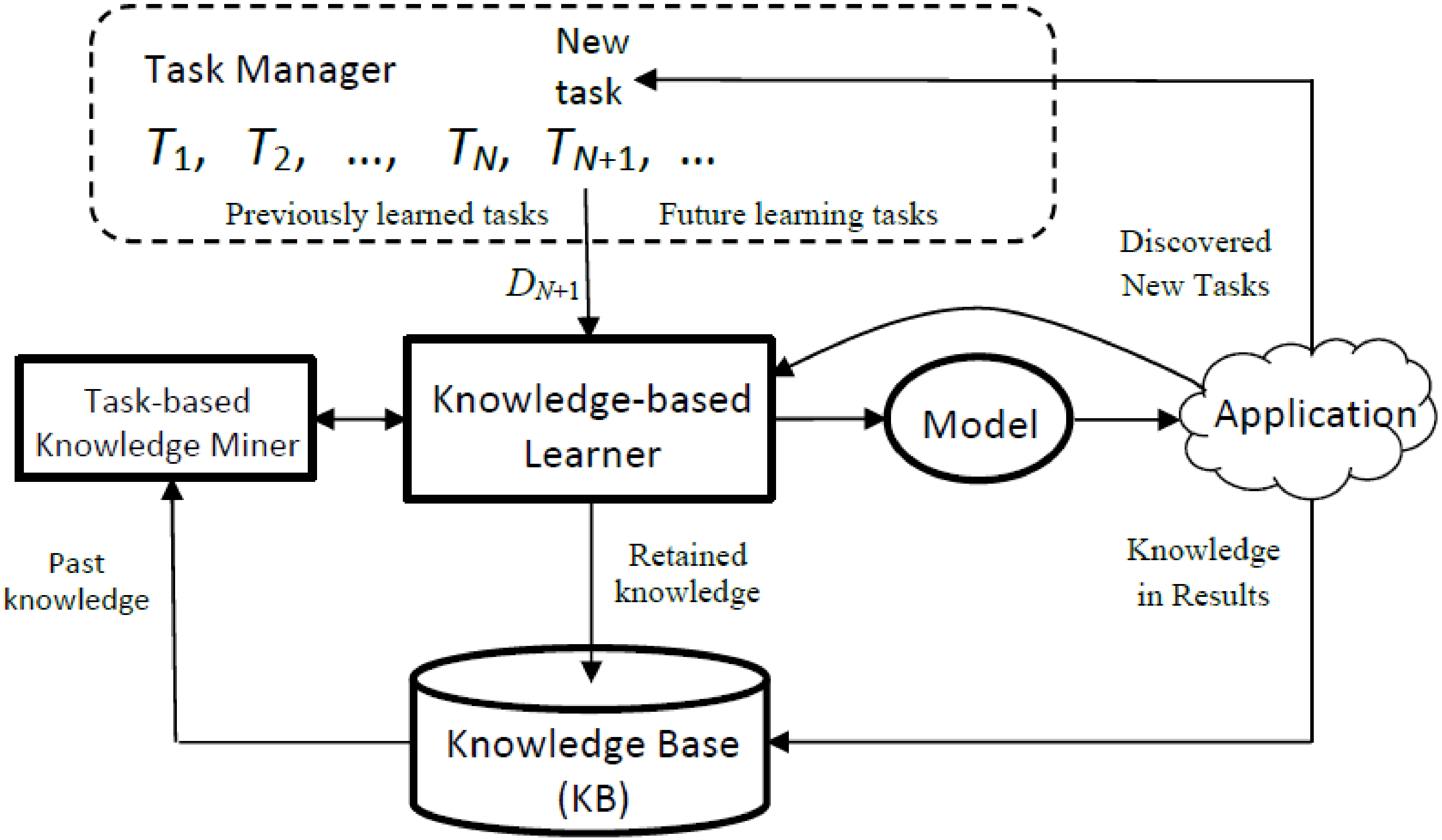
- The learner has performed learning on a sequence of tasks, from 1 to N.
- When faced with the $(N+1)$th task, it uses the relevant knowledge in its *knowledge base* (KB) to help learn the $(N+1)$th task.
- After learning $(N+1)$th task, KB is updated with the learned results from $(N+1)$th task.

New definition of lifelong learning (LL)

(Fei et al 2016, Shu et al 2017a, 2017, **Chen & Liu, 2018)**



2nd edition



Key characteristics of LL

(Chen and Liu, 2018-book)

1. **Continuous learning process** (w/o supervision)
2. **Knowledge accumulation in KB** (long-term memory)
3. **Using/adapting the past knowledge** to help future learning
4. **Learning in the open world**, discovering new or unseen tasks & learning them incrementally
5. **Learning on the job** or learning while working, during model application or testing

- Both (4) and (5) need self-supervision using the agent's own knowledge and/or environmental feedback.

Two types of shared knowledge

- *Global knowledge*: These methods assume a *global latent structure* among tasks that are shared (tasks are almost the same) (Thrun, 1996, Ruvolo and Eaton, 2013, Bou Ammar et al., 2014, …)
  - Global structure **L**: learned and leveraged in the new task learning.

$$\boldsymbol{\theta}^t = \mathbf{L}\mathbf{s}^t$$

- *Local knowledge:* do not assume a *global latent structure* among tasks (Chen and Liu, 2014, Fei et al., 2016, Liu et al., 2016, Shu et al., 2016, 2017,…).
  - During the new task learning, they meta-mine and selectively use those pieces of prior knowledge that are applicable.
    - Tasks can be fairly different.

Open problems/challenges

- Is the past knowledge actually correct?
- Is the past knowledge actually applicable?
- What is past knowledge and how to represent it?
- …

- LL forces us to think about the issue of knowledge and the role it plays in learning.
  - knowledge representation, acquisition, reasoning, maintenance, etc.

LL research is ramping up

- Many related topics and names
  - Lifelong learning
  - Continual learning (continuous learning)
  - Never-ending learning
  - Open-world learning
  - Meta-learning
  - Developmental learning (in robotics)
- **DARPA program** (2018): Lifelong learning machines (L2M)
- **DARPA program** (2019): Open-world learning/AI (SAIL-ON)

Transfer, Multitask → Lifelong

- Transfer learning: using source domain to help target domain,
  - Learning is not continuous
  - No accumulation of knowledge except data
  - Only one directional: help target domain
- Multitask learning: Jointly optimize multi. tasks
  - No accumulation of knowledge except data
  - Hard to re-learn all when tasks are numerous

- Both no discovery of new problems or learning in testing

Outline

- What is lifelong learning?
- Sentiment analysis and scale-up challenges
- Lifelong sentiment classification
- Lifelong topic models for aspect extraction and grouping
- Lifelong label propagation for identifying entities and aspects
- Learning on the job for aspect extraction
- Continual learning, meta-learning and open world learning
- Summary

Sentiment Analysis

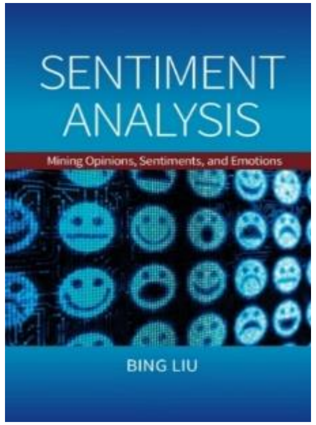
- Sentiment analysis (SA) (or opinion mining)
  - computational study of opinion, sentiment, emotion, appraisal, and evaluation and their applications.
- Why is it important? Hundreds of companies have worked on it
  - Opinions are key influencers of our behaviors.
    - Our beliefs/perceptions of reality are conditioned on how others see the world.
      - Whenever we make a decision we often seek out others' opinions.
- Started from CS: still a very challenging problem.
  - Spread to management science, finance, economics, health, political science, history, and many other fields.

Core Sentiment Analysis (SA) problem

(Hu and Liu 2004; Liu 2006, 2012)

- Id: John on 5-1-2008 -- "*I bought an iPhone yesterday. It is such a nice phone. **The touch screen is really cool**. The voice quality is great too. It is much **better** than my old Blackberry. …*"
- **Definition**: An *opinion* is a quadruple,
  (***target***, *sentiment*, *holder*, *time*)
- A more practical definition:
  (*entity*, *aspect*, *sentiment*, *holder*, *time*)
  - E.g., (iPhone, touch_screen, +, John, 5-1-2008)
- SA goal: Given an opinion doc, mine all quintuples

SA is a rich problem



- (***entity***, ***aspect***, ***sentiment***, *holder*, *time*)
  - target entity: Entity extraction & resolution
  - aspect of *entity*: Aspect extraction & resolution
  - sentiment: Aspect sentiment classification
  - *opinion holder:* *Information/data extraction*
  - *time:* *Information/data extraction*
- About all NLP problems
  - Aspect grouping (price = cost = expensive)
  - Lexical semantics
  - Coreference resolution
  - …..

Opinion summary

(Hu and Liu, 2004)

Aspect/feature based summary of opinions about a phone:

Aspect: **Touch screen**

Positive: 212

- *The touch screen was really cool.*
- *The touch screen was so easy to use and can do amazing things.*

…

Negative: 6

- The screen is easily scratched.
- I have a lot of difficulty in removing finger marks from the touch screen.

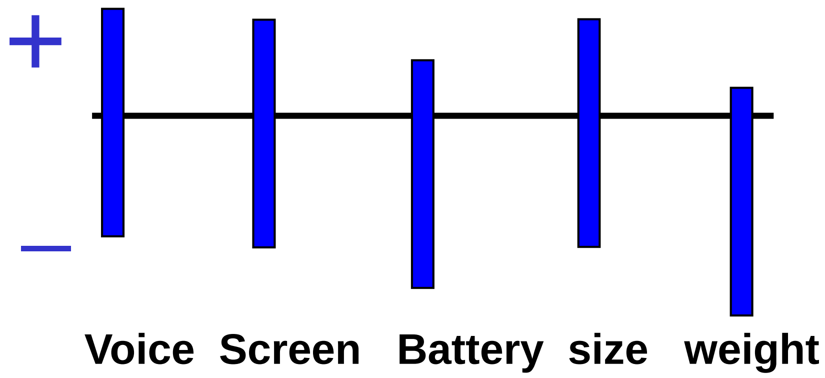
…

Aspect: **voice quality**

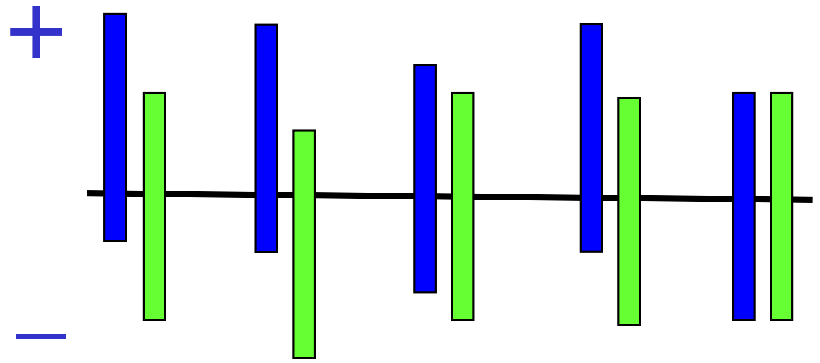
…

(Liu et al. 2005)

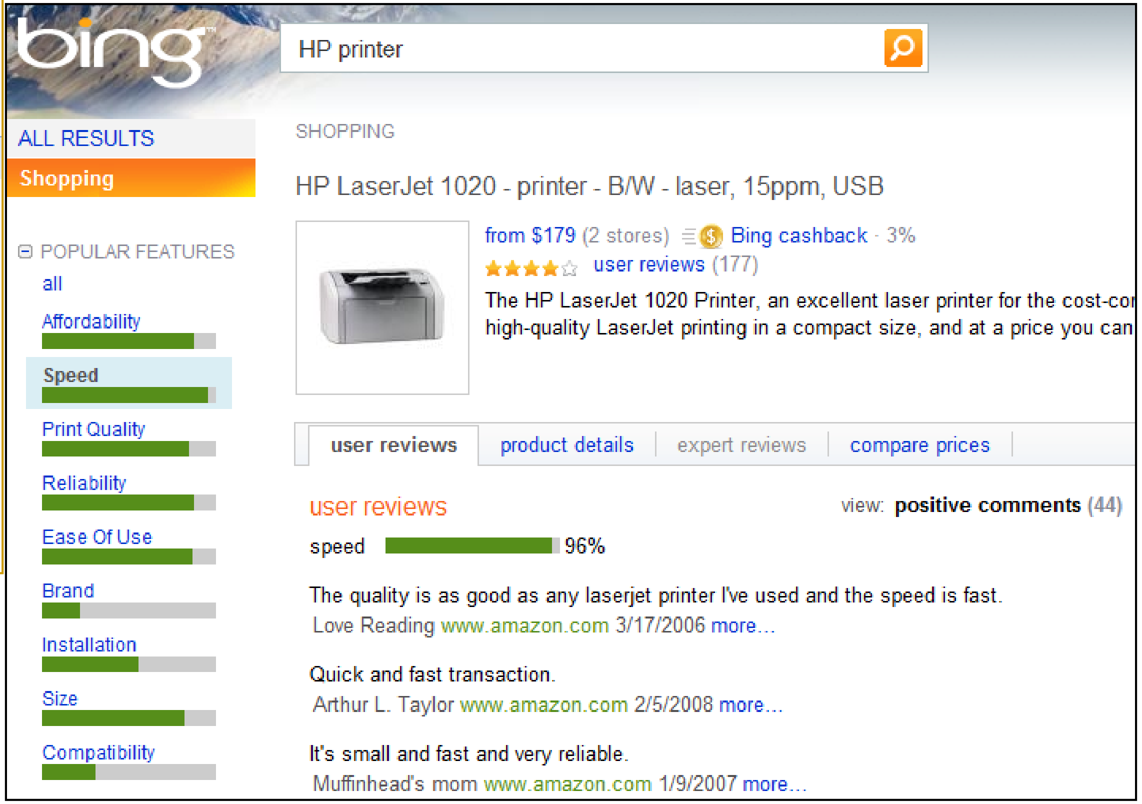
- Opinion Summary of 1 phone



- Opinion comparison of 2 phones



Aspect-based opinion summary



Scale-up challenges

- Not so hard to build a SA system for applications involving one or two domains (e.g., products categories).
- Major challenge in the real world: scale up to
  - thousands of categories (e.g., BestBuy), and
  - tens of thousands of categories (e.g., Amazon.com).
- Problems:
  - Low accuracy 60-70% for sentence-level sentiment classification
  - Aspect-level sentiment classification is worse.
    - Aspect extraction is a headache

Scale-up problem (cont'd)

- Supervised learning: simply labeling a huge amount of data does not work well.
  - Labeling for aspect sentiment analysis in large scale is hard.
  - There are existing systems, but not accurate.
- Unsupervised learning: every domain is different.
  - Words indicating positive sentiment in one domain may indicate negative (or no) sentiment in another domain
    - "This car is quiet" v.s. "This earphone is quiet"

Knowledge is needed

- A lot of prior knowledge is needed.
  - Depending on expert input will not scale.
- Can the system learn the prior knowledge by itself?
  - It is possible!
- We want to scale up the following subtasks of SA
  - Aspect sentiment classification
  - Aspect extraction and aspect grouping
  - Entity and aspect identification

Observation: knowledge shared across domains

- Experience in 2 (3) startups: After working on numerous SA projects for clients, I realized
  - a lot of sharing of concepts across domains (product categories)
- Easy to see this about sentiment words,
  - e.g., *good, bad, poor, terrible*, etc.
- Aspect grouping is shared too
  - picture = photo and expensive = price
- A great deal of product features/aspect sharing too …

Scale-up sentiment analysis (SA)!

- A great deal of product features (or aspects) sharing across domains, e.g.,
  - Every product review domain has the aspect *price*
  - Most electronic products share the aspect *battery*
  - Many also share the aspect of *screen*.
  - ….
- As we see more and more, fewer and fewer things are new.
- How to scale up SA to 1400 (BestBuy) or more product categories
  - Leveraging the sharing

My motivation to work on LL

- How to systematically leverage such sharing?
  - Retain/accumulate knowledge learned in the past.
  - use the knowledge for new task learning
- I.e.: ***lifelong learning***
  - Do better and better with more and more domains!
- This leads to our work (not just in SA)
  - Lifelong topic modeling and aspect extraction (Chen and Liu 2014a, b)
  - Lifelong sentiment classification (Chen et al. 2015)
  - others in 2016, 2017, 2018, 2019

Outline

- What is lifelong learning?
- Sentiment analysis and scale-up challenges
- Lifelong sentiment classification
- Lifelong topic models for aspect extraction and grouping
- Lifelong label propagation for identifying entities and aspects
- Learning on the job for aspect extraction
- Continual learning, meta-learning and open world learning
- Summary

Lifelong Sentiment Classification (LSC)
(Chen, Ma, and Liu 2015)

- "*I bought a cellphone a few days ago. It is such a nice phone. The touch screen is really cool. The voice quality is great too. ....*"
- Goal: classify docs or sentences as + or -.
  - Need to manually label a lot of training data for each domain, which is highly labor-intensive
- Can we not label for every domain or at least not label so many docs/sentences?

A simple LL approach

Assuming we have worked on a *large number of past domains* with all their training data *D*

- Build a classifier using *D*, test on new domain
- In many cases – improve accuracy by as much as 19% (= 80%-61%). Why?
- In some others cases – not so good, e.g., it works poorly for toy reviews. Why? “toy”

Lifelong learning: objective function

- Key idea: using past knowledge to revise probabilities (parameters of the generative model) of current domain
- Maximize the probability difference

$$\sum_{i=1}^{|D^t|} \left(P\left(c_j|d_i\right) - P\left(c_f|d_i\right)\right)$$

- c_j: labeled class in ground truth
- c_f: all classes other than c_j

Exploiting knowledge via penalties

- Penalty terms for two types of knowledge
  - Document-level knowledge
  - Domain-level knowledge

$$\frac{1}{2}\alpha \sum_{w \in V_S} \left(X_{+,w} - R_w \times X^0_{+,w}\right)^2 + \frac{1}{2}\alpha \sum_{w \in V_S} \left(X_{-,w} - (1 - R_w) \times X^0_{-,w}\right)^2$$

  - R_W : ratio of #tasks where *w* is positive / #all tasks
  - $X^0_{+,w} = N^t_{+,w} + N^{KB}_{+,w}$ and $X^0_{-,w} = N^t_{-,w} + N^{KB}_{-,w}$

Some results

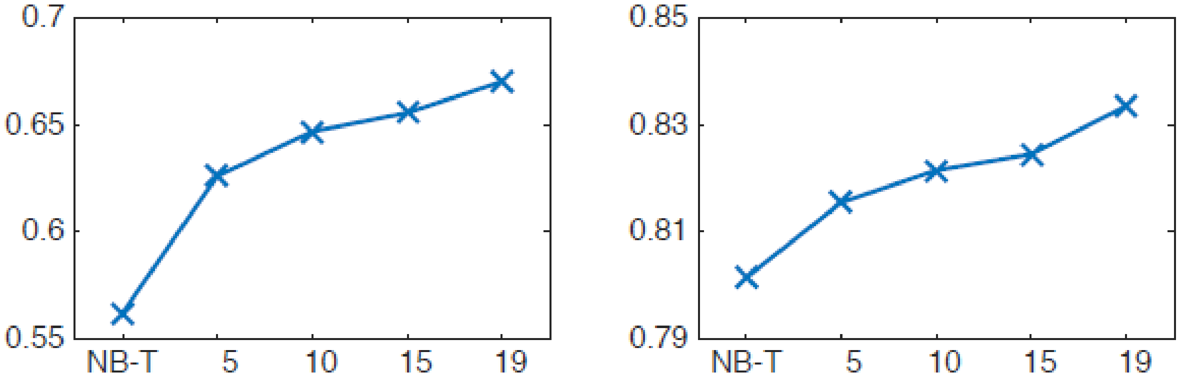


Figure 1: (Left): Negative class F1-score of LSC with #past domains in natural class distribution. (Right): Accuracy of LSC with #past domains in balanced class distribution.

Not in this talk: *Lifelong aspect sentiment classification (Wang et al., 2018)*

Outline

- What is lifelong learning?
- Sentiment analysis and scale-up challenges
- Lifelong sentiment classification
- Lifelong topic models for aspect extraction and grouping
- Lifelong label propagation for identifying entities and aspects
- Learning on the job for aspect extraction
- Continual learning, meta-learning and open world learning
- Summary

Lifelong Topic Modeling (LTM) (Chen and Liu, 2014)

Aspect extraction: *"the battery is great, but pictures are poor."*

- Aspect terms: battery, picture

- Aspect extraction actually has two tasks:

  (1) extracting aspect terms

  - "picture," "photo," "battery," "power"

  (2) grouping synonymous aspect terms.

  - {"picture," "photo"}, {"battery," "power"}

- Top modeling performs both tasks at the same time.
  - E.g., {*price*, *cost*, *cheap*, *expensive*, …}

What knowledge?

(Chen and Liu, 2014, 2014)

- Should be in the same topic

  => Must-Links

  e.g., {picture, photo}

- Should not be in the topic

  => Cannot-Links

  e.g., {battery, picture}

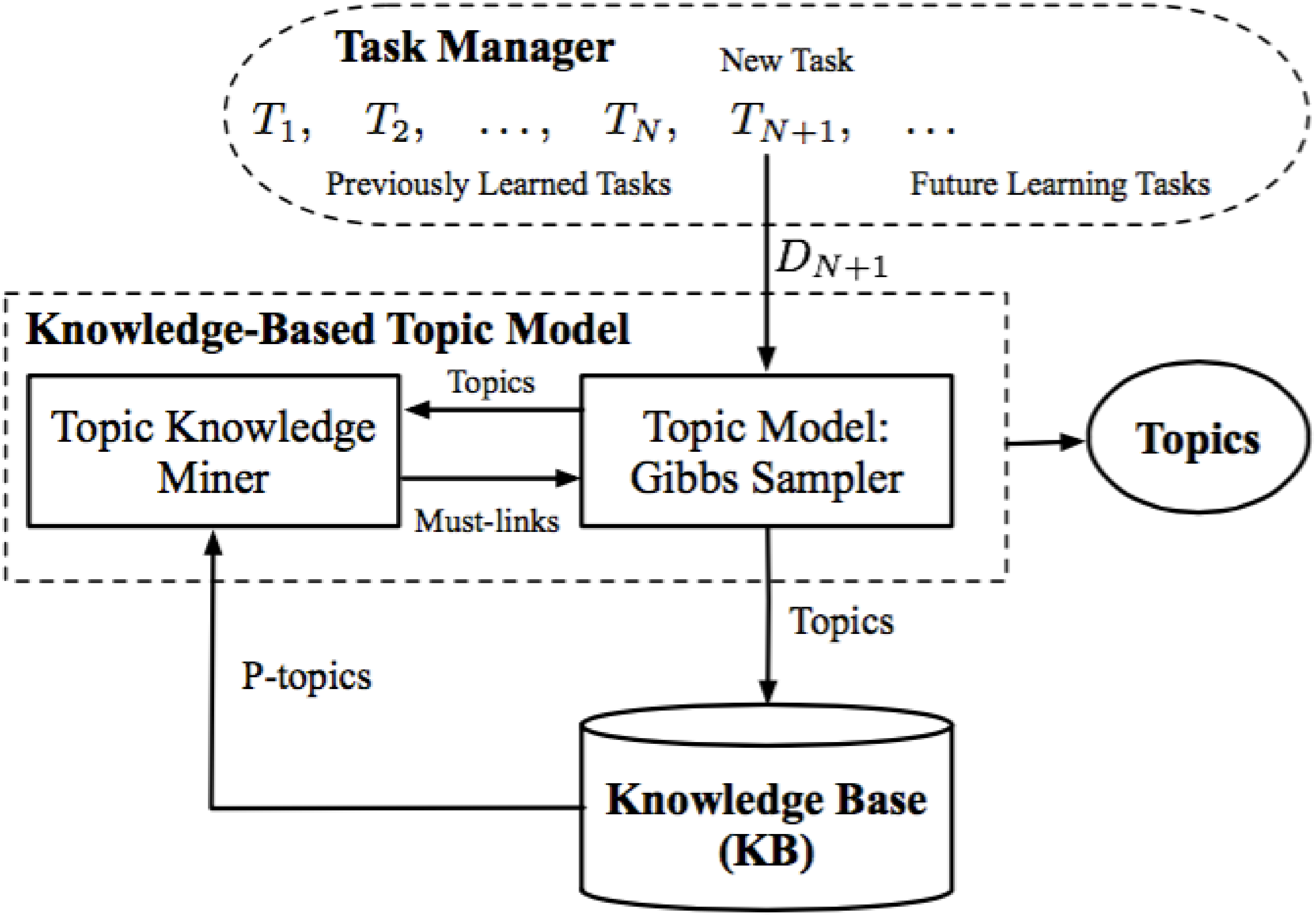
LTM: Lifelong Topic Modeling

(Chen and Liu, 2014)

Algorithm 2 LTM(D^t, S)

1: $A^t \leftarrow$ GibbsSampling(D^t, $\emptyset$, N); // Run N Gibbs iterations with no knowledge (equivalent to LDA).
2: **for** $i = 1$ **to** N **do**
3: $\quad K^t \leftarrow$ KnowledgeMining(A^t, S);
4: $\quad A^t \leftarrow$ GibbsSampling(D^t, K^t, 1); // Run with knowledge K^t.
5: **end for**

LTM architecture



An example of knowledge mining

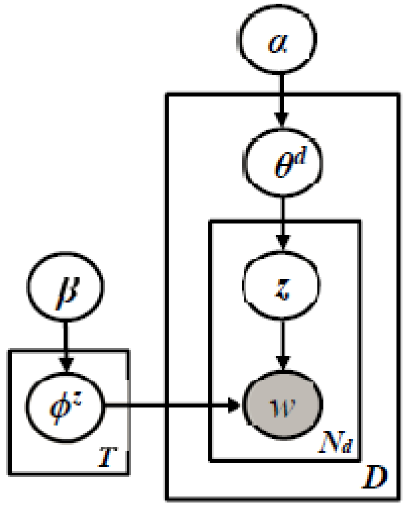
- Given a newly discovered topic:

  {*price*, *book*, *cost*, *seller*, *money*}

  - We find 3 matching topics from topic base *S*
    - Domain 1: {*price*, *color*, *cost*, *life*, *picture*}
    - Domain 2: {*cost*, *screen*, *price*, *expensive*, *voice*}
    - Domain 3: {*price*, *money*, *customer*, *expensive*}
- If we require words to appear in at least two domains, we get two must-links (knowledge):
  - {*price*, *cost*} and {*price*, *expensive*}.
  - Each set is likely to belong to the same aspect/topic.

Model inference: Gibbs Sampling

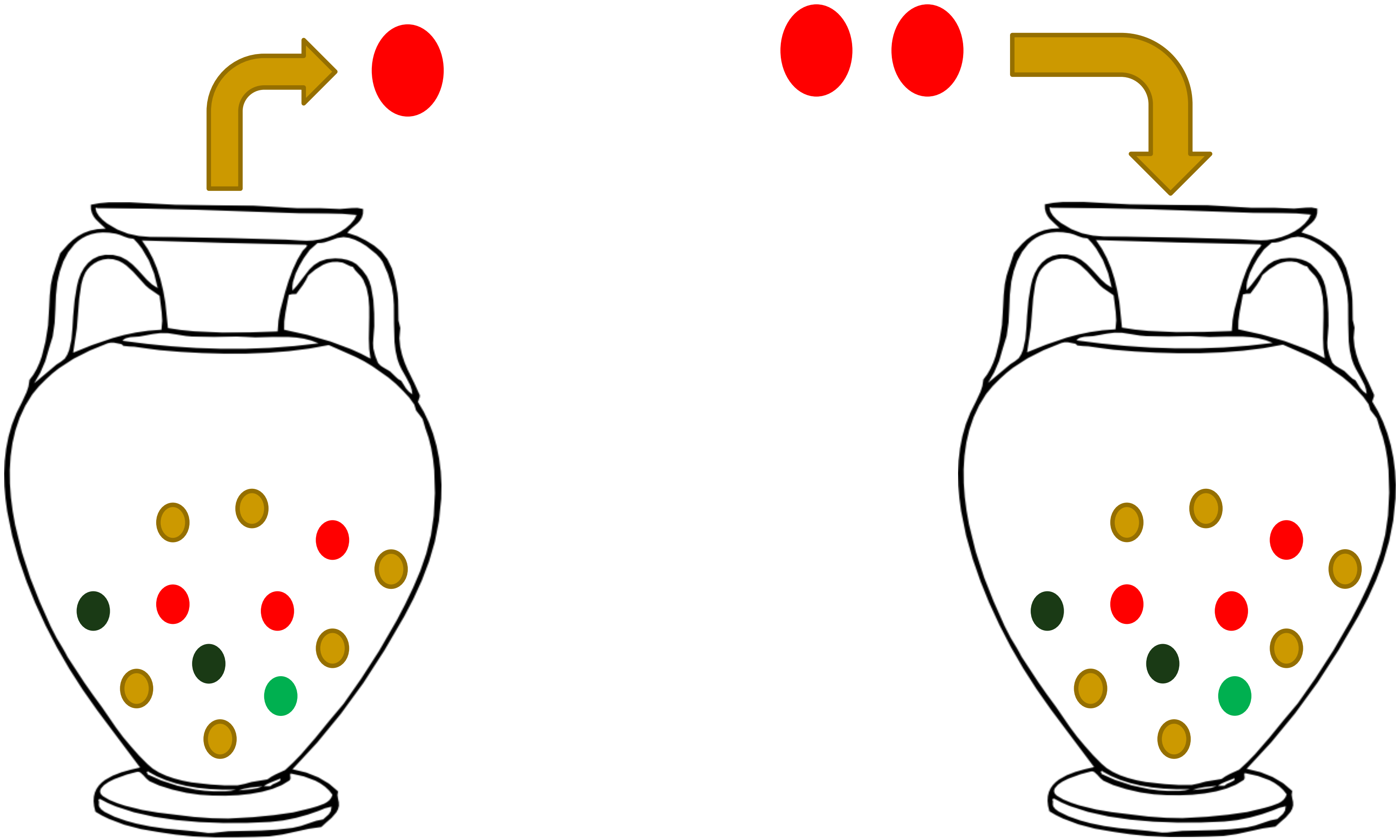
- How to use the *must-links* knowledge?
  - e.g., {*price*, *cost*} & {*price*, *expensive*}
  - How to know they are applicable?
- Graphical model: same as LDA, but different inference
  - Generalized Pólya Urn Model (GPU)



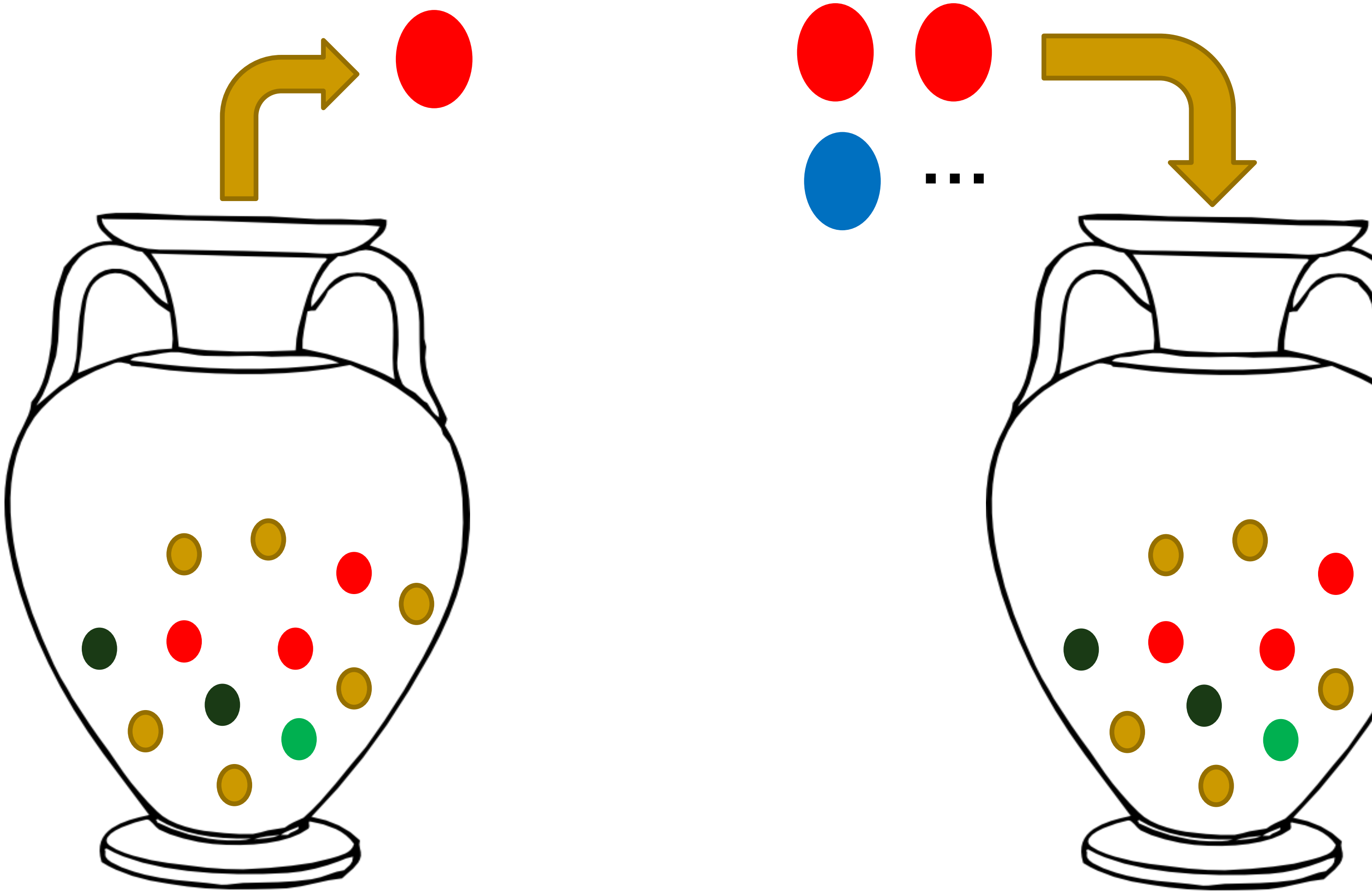
- Idea: When assigning a topic *t* to a word *w*, also assign *a fraction of t* to words in must-links sharing with *w*.

$$P(z_i = t | \boldsymbol{z}^{-i}, \boldsymbol{w}, \alpha, \beta, \mathbf{A}') \propto$$
$$\frac{n_{d,t}^{-i} + \alpha}{\sum_{t'=1}^{T}(n_{d,t'}^{-i} + \alpha)} \times \frac{\sum_{w'=1}^{V} \mathbf{A}'_{t,w',w_i} \times n_{t,w'}^{-i} + \beta}{\sum_{v=1}^{V}(\sum_{w'=1}^{V} \mathbf{A}'_{t,w',v} \times n_{t,w'}^{-i} + \beta)}$$

Simple Pólya Urn model (SPU)



Generalized Pólya Urn model (GPU)



Outline

- What is lifelong learning?
- Sentiment analysis and scale-up challenges
- Lifelong sentiment classification
- Lifelong topic models for aspect extraction and grouping
- Lifelong label propagation for identifying entities and aspects
- Learning on the job for aspect extraction
- Continual learning, meta-learning and open world learning
- Summary

Lifelong learning in graph label propagation

(Shu et al 2016)

- **Problem:** Identify entity and aspect in sentiment analysis.
  - This car has a great engine.
- Relaxation Labeling (RL): an unsupervised graph-based label propagation algorithm for classification
  - Each node n_i has a multinomial distribution
    - $P(L(n_i))$ ($L(n_i)$ is the label of n_i on a label set Y).
  - Each edge has two conditional distributions:
    - $P(L(n_i) \mid L(n_j))$ and $P(L(n_j) \mid L(n_i))$
- Augmented it with lifelong learning (*Lifelong-RL*).

Relaxation Labeling (contd)

- Start from initial $\boldsymbol{P^0(L(n_i))}$**, the algorithm iteratively** updates label distribution: Iteration r + 1 is computed as follows:

$$P^{r+1}(L(n_i)) = \frac{P^r(L(n_i)) \times (1 + \Delta P^{r+1}(L(n_i)))}{\sum_{y \in Y} P^r(L(n_i) = y) \times (1 + \Delta P^{r+1}(L(n_i) = y))}$$

- The final label of node n_i is its highest probable label.

$$L(n_i) = \underset{y \in Y}{\operatorname{argmax}}(P(L(n_i) = y))$$

Lifelong relaxation labeling

Lifelong-RL uses two forms of knowledge

- Prior edges: graphs are usually not given or fixed but are built based on text data.
  - If the data is small, many edges may be missing
  - But such edges may existing in the graph of some previous tasks
- Prior labels: initial $P^0(L(n_i))$ is quite hard to set, but results from previous tasks can be used to set it more accurately.

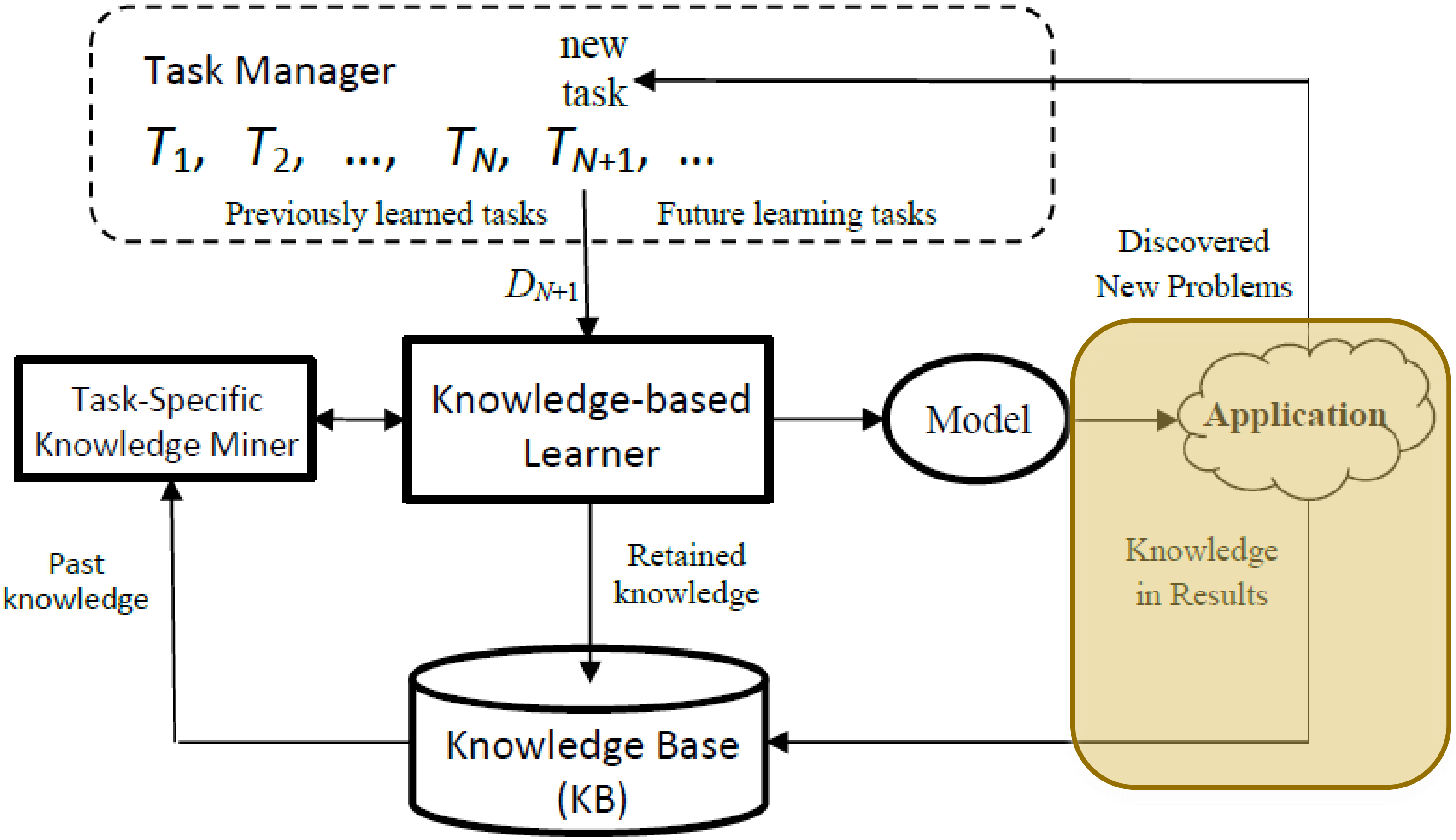
Outline

- What is lifelong learning?
- Sentiment analysis and scale-up challenges
- Lifelong sentiment classification
- Lifelong aspect extraction and grouping
- Lifelong label propagation for identifying entities and aspects
- Learning on the job for aspect extraction
- Continual learning, meta-learning and open world learning
- Summary

Learning on the job

- It is known that about 70% of our human knowledge comes from ‘on-the-job’ learning
  - Only about 10% through formal training
  - The rest 20% through observation of others
    - (These DO Not include our commonsense knowledge)
- Goal: improve model performance after it has been built
  - Using the knowledge gained in the application of the model (Shu, Xu, and Liu, 2017) and environment feedback

Lifelong learning extended



Improving model in testing or execution

(Shu et al 2017)

- Can a model's performance be improved after training?
- This paper proposes a technique to do this in the context of CRF for information extraction.
- Idea: connect features with extraction results
  - More results $\rightarrow$ better features
  - It exploits dependency features
    - As the model sees more data, more features are identified
    - These features help produce better results in the new domain using the same model.

Summary of LL and sentiment analysis and …

- There is clearly information sharing that can be leveraged.
  - Knowledge can clearly be accumulated
  - What we have learned earlier is useful later
  - As we see more and more, fewer and fewer things are new.
- Isn't this true for natural language processing (NLP) as a whole?
  - Words/phrases almost have the same meaning in different domains/fields.
  - Sentences in all domains follow the same syntax
  - NLP problems are closely related to each other
    - POS tagging, coreference resolution, entity recognition, linking, dialogue, …
- Isn't this true for any field of study beyond NLP?

Outline

- What is lifelong learning?
- Sentiment analysis and scale-up challenges
- Lifelong sentiment classification
- Lifelong topic models for aspect extraction and grouping
- Lifelong label propagation for identifying entities and aspects
- Learning on the job for aspect extraction
- Continual learning, meta-learning and open world learning
- Summary

Continual learning and catastrophic forgetting

- Continual learning (CL) is mainly for solving the *catastrophic forgetting* problem in neural networks.
- Catastrophic forgetting (CF): learning a new task will change the weights that have been learned for past tasks,
  - degrading the models for previous tasks.
  - like a human brain, ideally we want a network (like a brain) to learn many tasks with little interference (or little forgetting of the past).
- CL mainly aims to solve CF. Unlike lifelong learning, CL usually does not emphasize leveraging the past knowledge.

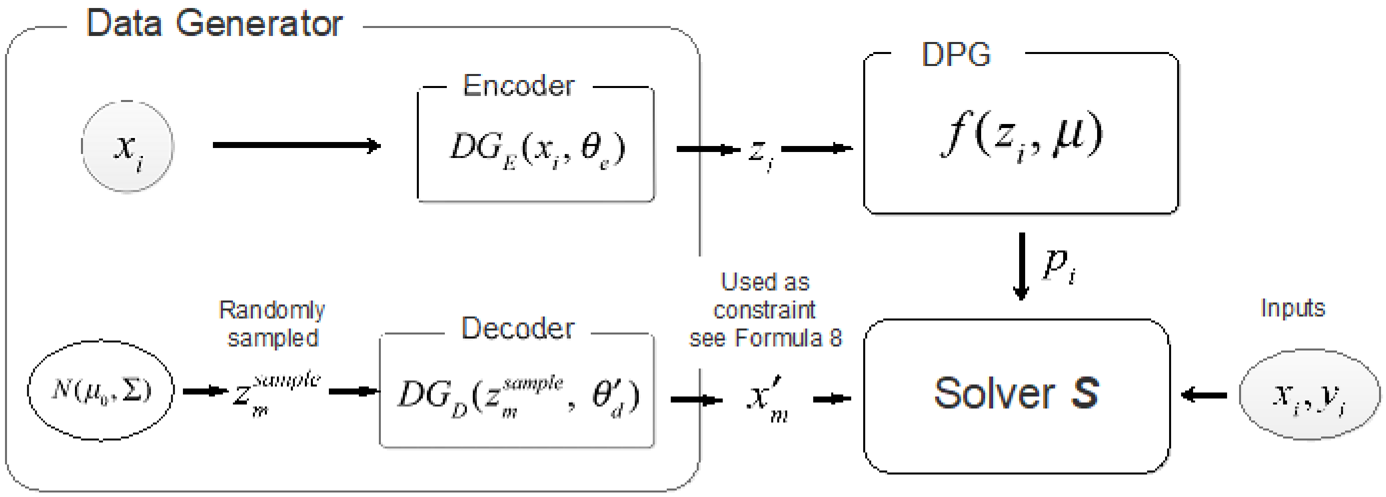
Continual learning (CL) – many approaches

- Kirkpatrick et al. (2017) - EWC: quantify the importance of weights based on their impact on previous tasks' performance.
  - In learning a new task, try not to disturb those important weights.
- Rebuffi et al. (2017) retain an exemplar set best approximating the previous tasks, and
  - use them and the new task data to learn the new task
- Shin et al. [2017] use GAN to learn generators for previous tasks
  - learn the new task parameters using new task data and the replayed data generated from GAN of previous tasks.
- ……

Overcoming CF via model adaptation (Hu et al., 2019)

- Learn a model (called Solver) with two sets of parameters.
  - The first set is shared by all tasks learned so far
  - The second set is dynamically generated to adapt the Solver to suit each test instance in order to classify it.

Data Generator (DG) ➢ Solver ➢ Dynamic Parameter Generator (DPG)



Meta-learning

- Meta-learning, also called *learning to learn*, is often used in one-shot or few-shot learning.
- It trains a meta-model with a large set of tasks. Each task has a set of labeled examples.
  - The learned model can quickly adapt to a new task using only a few examples (few-shot learning).
  - Meta-learning treats all these tasks as training "instances."
- Assume: training tasks and test tasks - same distribution
  - But in real life, new tasks fundamentally different in some aspects.

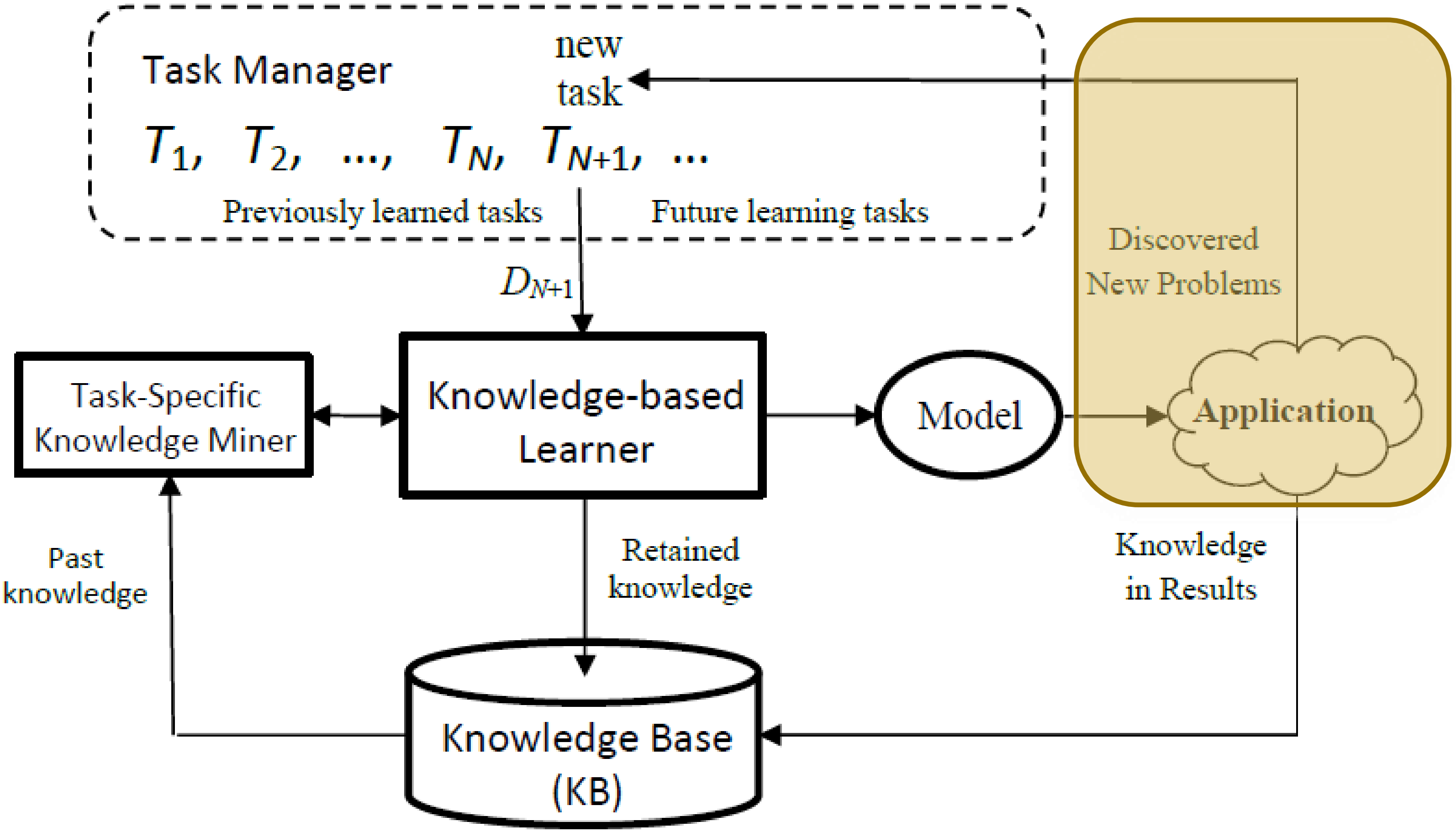
Open-world Learning (OWL)

(Fei et al, 2016)

- Traditional learning makes the closed world assumption:
  - Classes in testing have been seen in training, no new class in testing
- Learning in the open world
  - Training data: $D^t = \{D_1, D_2, \ldots, D_t\}$ of classes $Y^t = \{l_1, l_2, \ldots, l_t\}$.
  - Test data: $D_{t+1}, Y^{t+1} \in \{l_1, l_2, \ldots, l_t, \boldsymbol{l_0}\}$
  - **Tasks**
    - **1:** learn a classifier to classify test instances from $Y^t = \{l_1, l_2, \ldots, l_t\}$, and detect/reject instances not from Y^t.
    - **2:** identify unseen classes in rejected instances (Shu et al. 2018).
    - **3**: incrementally learn the new/unseen classes

Lifelong learning extended

(Fei et al., 2016 (SVM); Shu et al., 2017 (DNN), Xu et al, 2019 (DNN))



DOC: Deep Open Classification

(Shu et al. 2017)

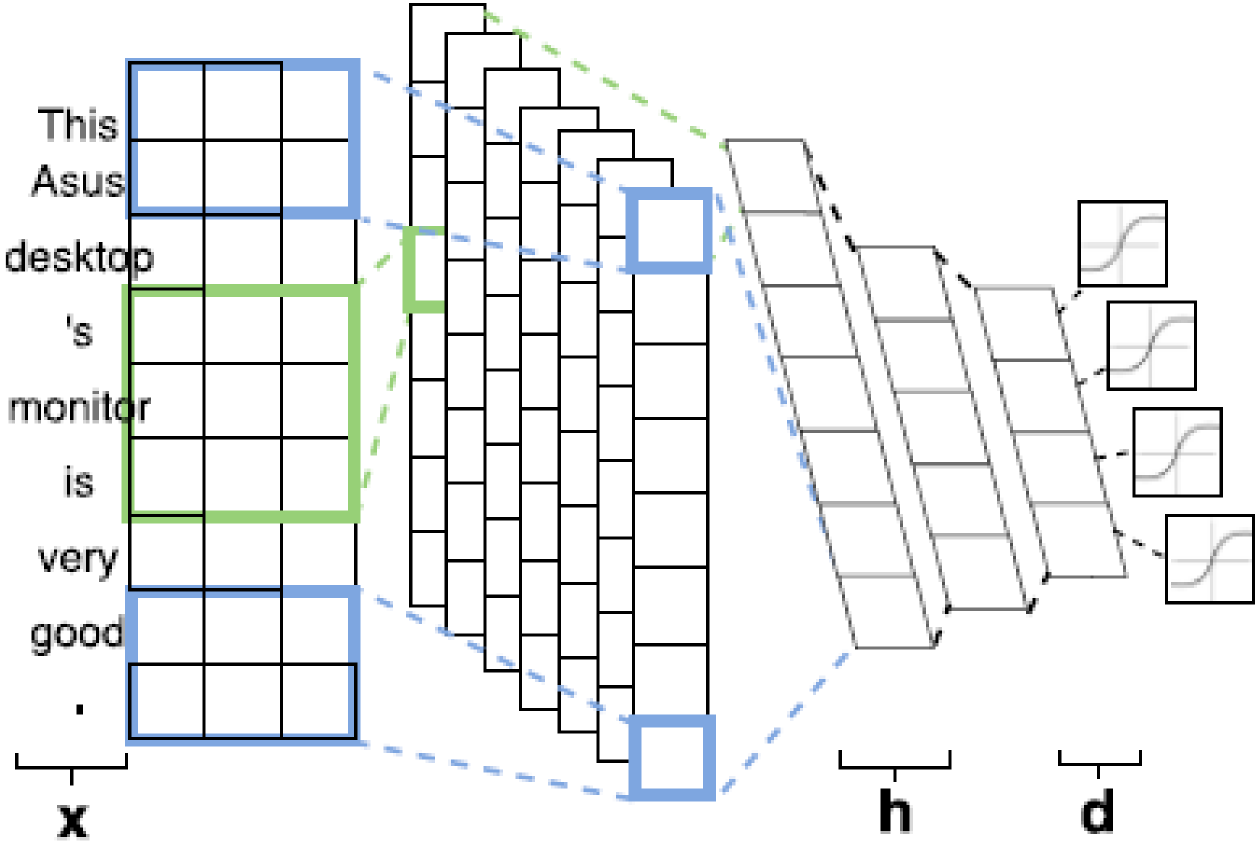


Figure 1: Overall Network of DOC

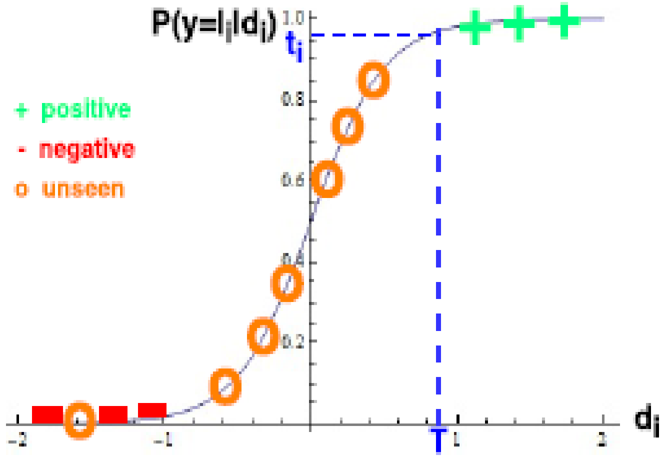
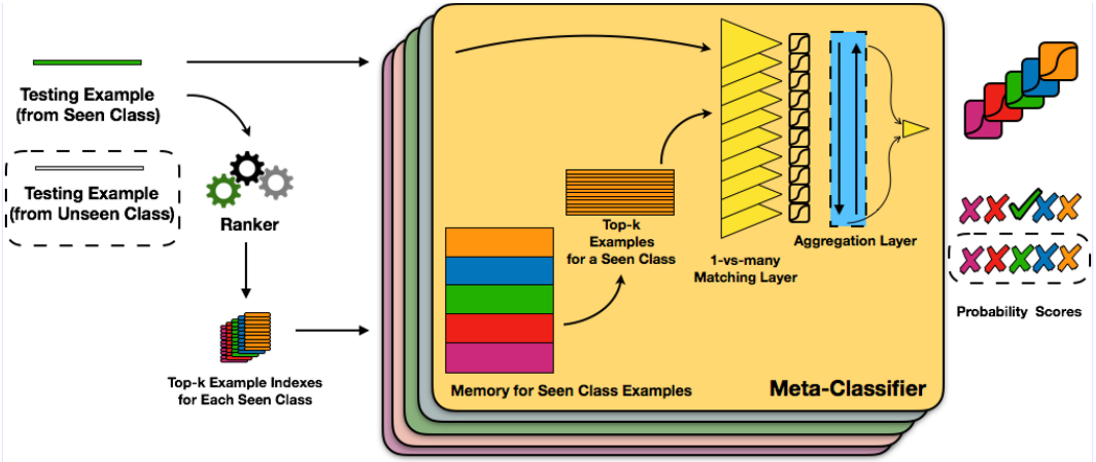


Figure 2: Open space risk of sigmoid function and desired decision boundary $d_i = T$ and probability threshold t_i.

Open-world learning with meta-learning

(Xu et al. 2019)



- **L2AC –meta-learning**
  - It maintains a dynamic set S of seen classes that allow new classes to be added or deleted without re-training.
    - Each class is represented by a small set of training examples.
  - In testing, the meta-classifier uses only the examples of the seen classes on-the-fly for classification and rejection

Outline

- What is lifelong learning?
- Sentiment analysis and scale-up challenges
- Lifelong sentiment classification
- Lifelong topic models for aspect extraction and grouping
- Lifelong label propagation for identifying entities and aspects
- Learning on the job for aspect extraction
- Continual learning, meta-learning and open world learning
- Summary

Summary

- Lifelong learning is *necessary* for future machine learning.
  - Isolated single task learning is insufficient for intelligence.
  - We are in an exciting time with lots of opportunities and challenges
    - Sentiment analysis is just one application.
- LL is still in its infancy with huge challenges (Chen & Liu, 2018):
  - Correctness and applicability of knowledge, self-supervision and interaction, knowledge representation & reasoning, composition, etc.
  - Going beyond
    - Is the current DNN suitable for lifelong learning?
    - What is the cognitive model for learning?

Thank You

Q&A

https://www.cs.uic.edu/~liub/lifelong-learning.html

https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html