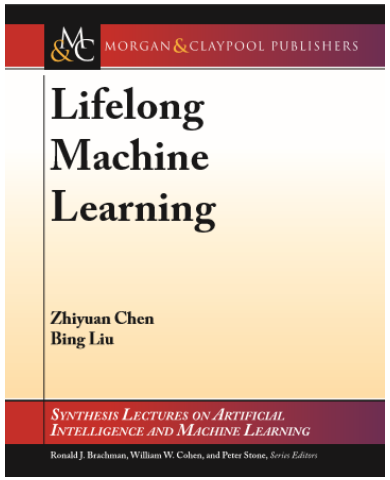
Bring NLP to the Next Level:
Lifelong Language Learning

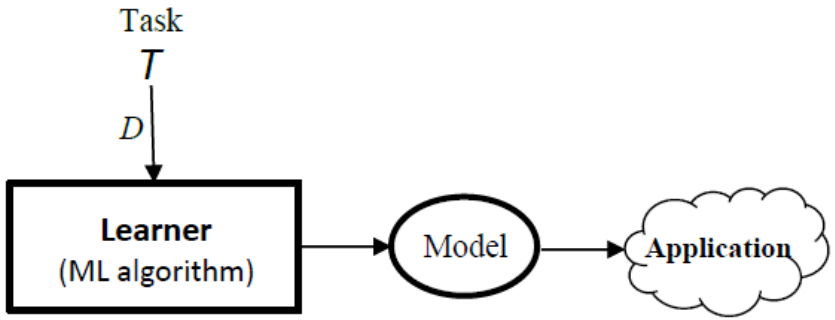


Book, Nov 2016

Bing Liu
Department of Computer Science
University of Illinois at Chicago

Introduction: classic learning paradigm (ML 1.0)

(Chen and Liu, 2014, 2016-book)



- Isolated single-task learning
  - No consideration any previously learned knowledge
- Weaknesses of "isolated learning"
  - Knowledge learned not accumulated, i.e., *no long term memory*.
    - Needs a large number of training examples
    - Suitable for well-defined & narrow tasks in restricted env.
    - No self-motivation, no self-learning
- ML 1.0 will never make a system truly intelligent

Introduction: Small Sample Learning

- Obtain big labeled training data is very difficult
  - cannot expect humans to label for every task
- Solution: Small sample learning or one-shot learning
- But these methods all need **prior knowledge**
- BIG question:
  - Where does the **prior knowledge** come from?
  - From human users? --- NOT an ideal solution
    - **Cannot learn by itself**

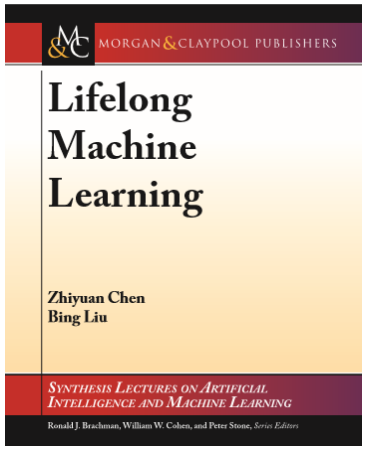
Introduction: humans don't learn in isolation

- Nobody has ever given me 1000 positive and 1000 negative reviews and ask me
  - to build a sentiment classifier
  - I can do it without any learning as
    - I have learned how people praise and criticize things
- If I don't have the accumulated knowledge, I cannot do it.
  - E.g., if someone gives me 2000 training +/- Arabic reviews,
  - I cannot build a classifier.

Machine learning: ML 2.0
(Chen and Liu, 2016-book)

- Humans never learn in isolation from scratch
  - We learn continuously and
    - accumulate knowledge learned in the past and use it to learn more & better.
    - Learn effectively from a few examples and self-motivated
- *Lifelong Machine Learning* (LML) (or *lifelong learning*):
  - mimics this human learning capability
- **Goal:** Create a machine that learns like humans
  - I believe we will never achieve human-like intelligence without it
  - Artificial General Intelligence (AGI)

LML - Getting Attentions!



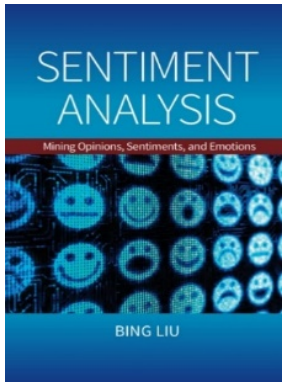
Book, Nov 2016

- First ever funding programs from two major agencies
- **DARPA:** Defense Advanced Research Projects Agency
  - **Lifelong Learning Machines (L2M)**
    - Abstract deadline: May 3, 2017
    - Proposal deadline: June 30, 2017 (was June 21, 2017)
- **CHIST-ERA**: European Coordinated Research on Long-term Challenges
  - **Lifelong Learning for Intelligent Systems (LLIS)**
    - Proposal deadline: January 17, 2017

Outline

- A motivating example
- Definition of lifelong learning
- Lifelong supervised learning
- Lifelong unsupervised learning
- Learning in the open world - discovering new problems
- Learning in model testing/application
- Summary

Motivating Example
(Liu, 2012; 2015)



- Interest in LML stemmed from extensive experiences on *sentiment analysis* in a company years ago.
- Sentiment analysis (SA)
  - Sentiment and target aspect: "*The screen is great, but the voice quality is poor.*"
    - Positive about screen but negative about voice quality
  - Extensive knowledge sharing across tasks/domains
    - Sentiment expressions & aspects

Knowledge Shared Across Domains

- After working with many SA problems for clients, I realized
  - a lot of sharing of concepts across domains
    - as we see more and more, fewer and fewer things are new.

- Easy to see this about sentiment words,
  - e.g., *good, bad, poor, terrible*, etc.

- There is also a great of sharing of aspects (product features)

Sharing of Product Features

- Observation: A great deal of product features (or aspects) overlapping across domains
  - Every product review domain has the aspect *price*
  - Most electronic products share the aspect *battery*
  - Many also share the aspect of *screen*.
  - Many also share sound quality
  - ….
- It is rather "silly" not to exploit such sharing in learning or extraction.

What does that mean for learning?

- How to systematically exploit such sharing?
  - Retain/accumulate knowledge learned in the past.
  - Leverage the knowledge for new task learning
- That is: *lifelong machine learning* (LML)
- This leads to our own work
  - Lifelong topic modeling (Chen and Liu 2014a, b)
  - Lifelong sentiment classification (Chen et al. 2015)
  - Several others …

LML is very suitable for NLP

- Knowledge, is shared easily across domains
  - Words and phrases almost have the same meaning in different domains.
  - Sentences in all domains follow the same syntax
  - We do not learn a new language whenever we go to a different domain
- NL knowledge is naturally accumulative and compositional
  - What you learned earlier is always useful later
- Knowledge is useful in different types of tasks.
  - NLP problems are closely related to each other
    - POS tagging, coreference resolution, entity recognition, …

Outline

- A motivating example
- Definition of lifelong learning
- Lifelong supervised learning
- Lifelong unsupervised learning
- Learning in the open world - discovering new problems
- Learning in model testing/application
- Summary

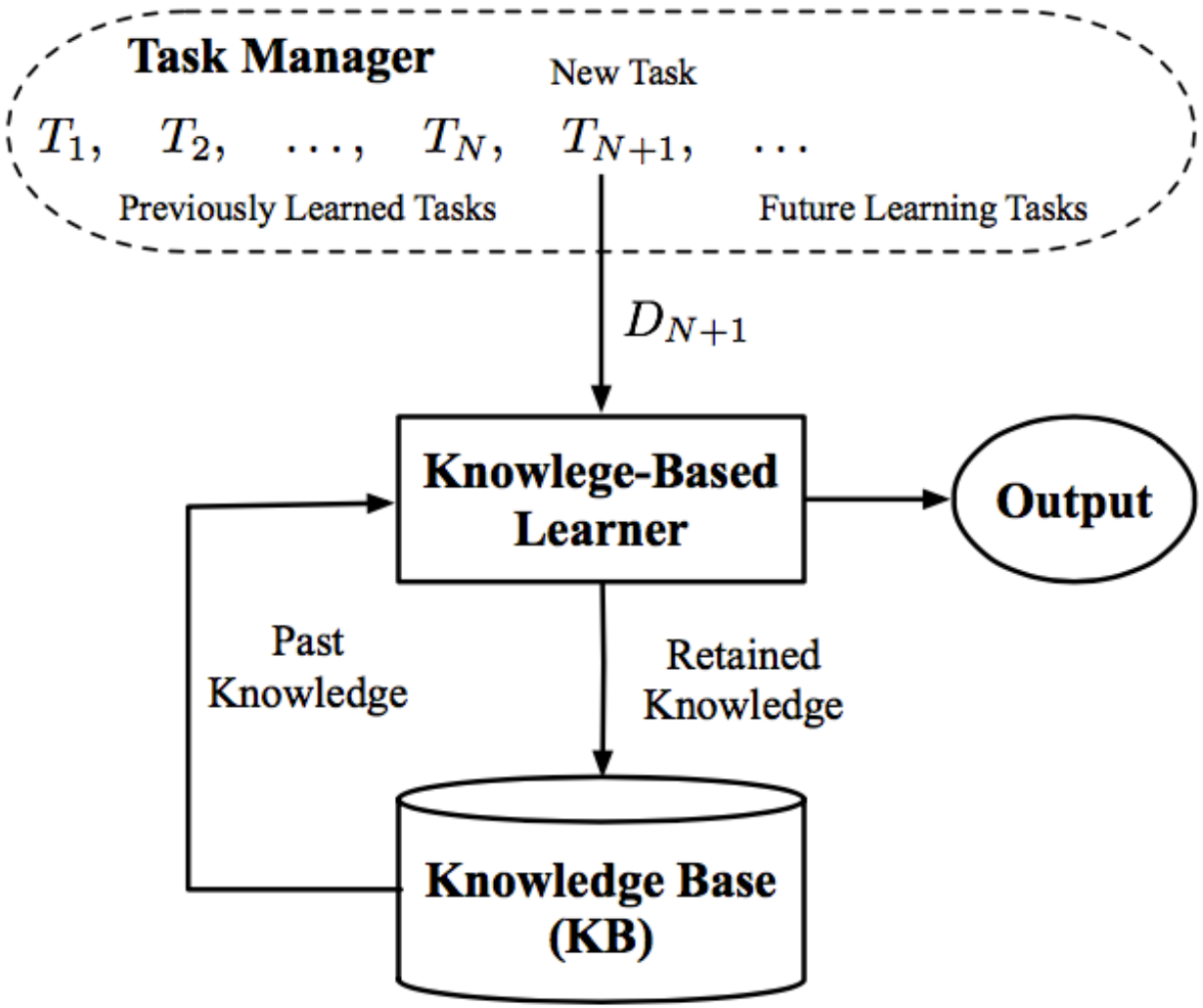
Definition of LML (existing)

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

- 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 learning for the (*N*+1)th task.
- After learning (*N*+1)th task, KB is updated with learned results from (*N*+1)th task.

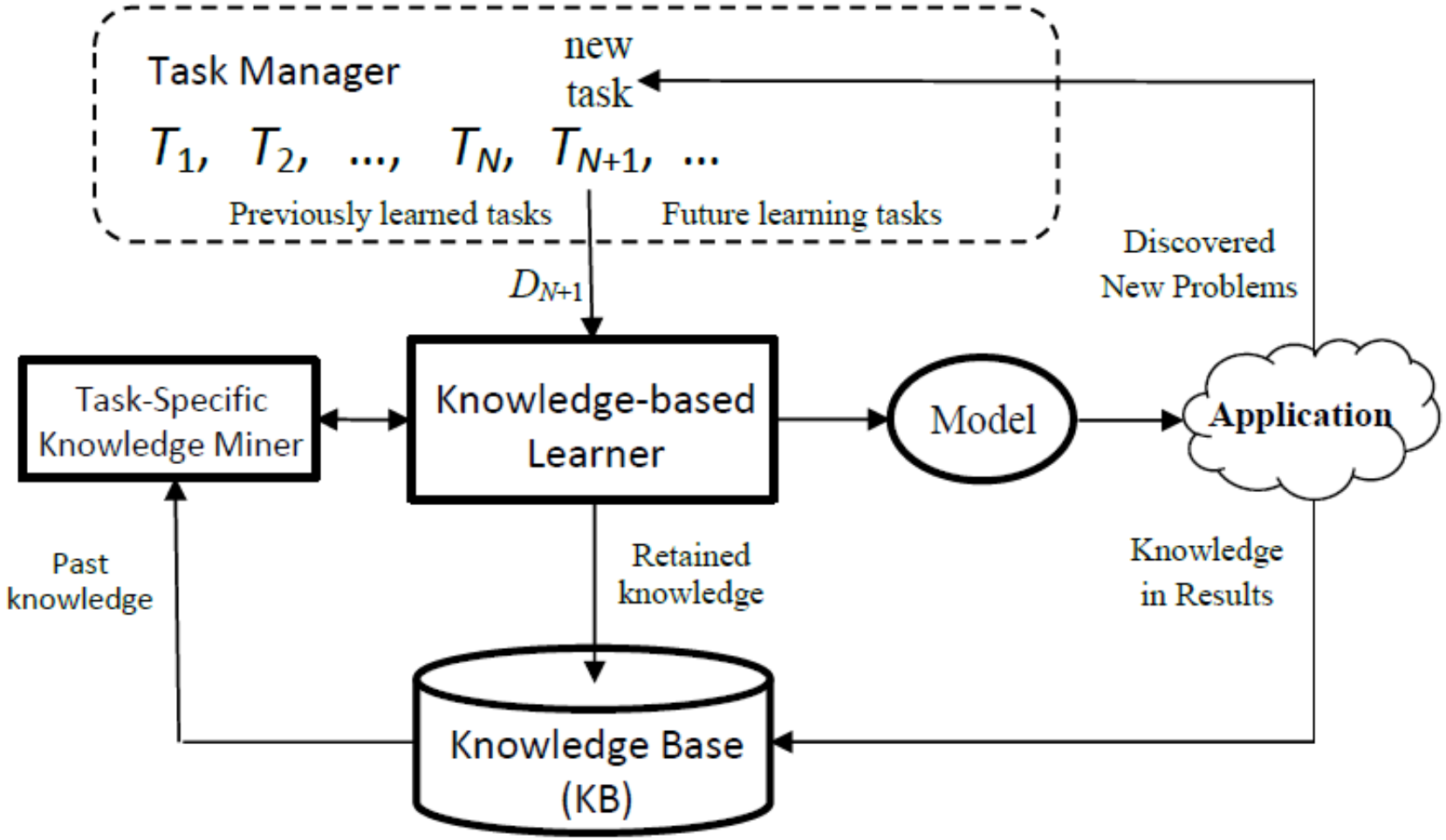
LML System Architecture

(Chen and Liu, 2016-book)



A Better LML Definition (new definition)

(Fei et al 2016, Shu et al 2017a, 2017b)



Key Characteristics of LML

(Chen and Liu, 2016-book)

- Continuous learning process
  - Learning not just in training, but also in testing
- Knowledge accumulation in KB (long-term memory)
- Use and **adapt** the past learned knowledge to help future learning and problem solving

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

- A motivating example
- Definition of lifelong learning
- Lifelong supervised learning
- Lifelong unsupervised learning
- Learning in the open world - discovering new problems
- Learning in model testing/application
- Summary

Lifelong Sentiment Classification

(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 LML 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"

Objective Function

- 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 $X^0_{-,w} = N^t_{-,w} + N^{KB}_{-,w}$

Use words not appeared in training

- Interestingly, use words for classification in testing that have not appeared in training

  $P(w_i \mid c_j)$

- w_i did not appear in the training data, but appears in the test data.
  - Traditional classifier cannot use it
- But in lifelong learning, we can because
  - w_i may have appeared in past learned knowledge.

One Result

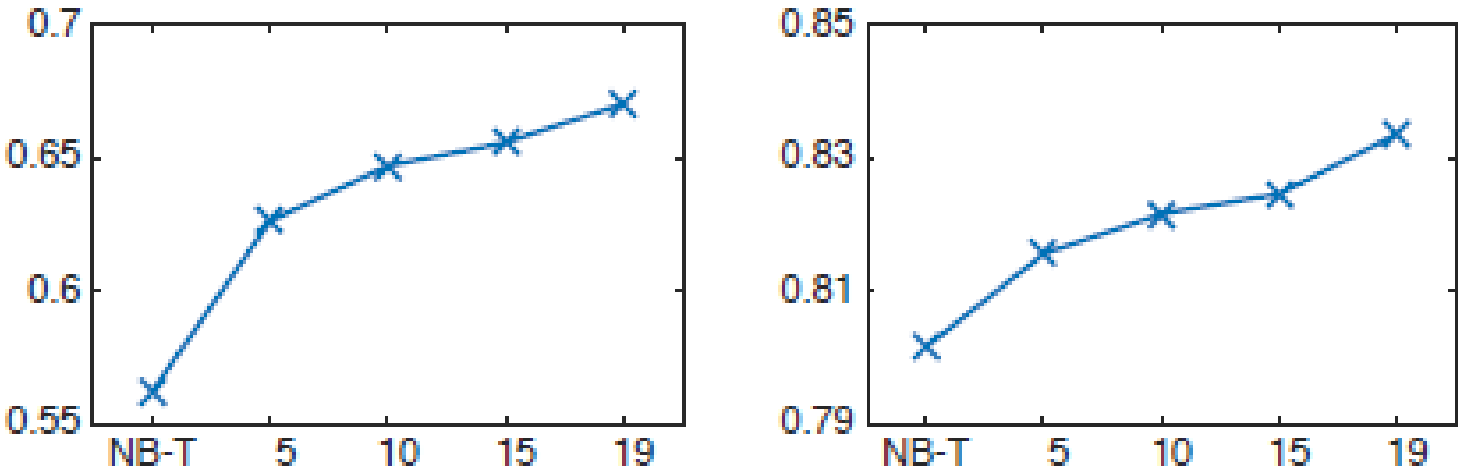


Figure 1: Figure 1. (Left): Negative class F1-score of LSC with #past domains in natural class distribution. (Right): Accuracy of LSC with #past domains in balanced class distribution.

Outline

- A motivating example
- Definition of lifelong learning
- Lifelong supervised learning
- Lifelong unsupervised learning
- Learning in the open world - discovering new problems
- Learning in model testing/application
- Summary

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

- *"The battery is great, but pictures are poor."*
  - Topic terms: battery, picture
- Topic extraction actually has two tasks:

  (1) extract topical terms
  - "picture," "photo," "battery," "power"

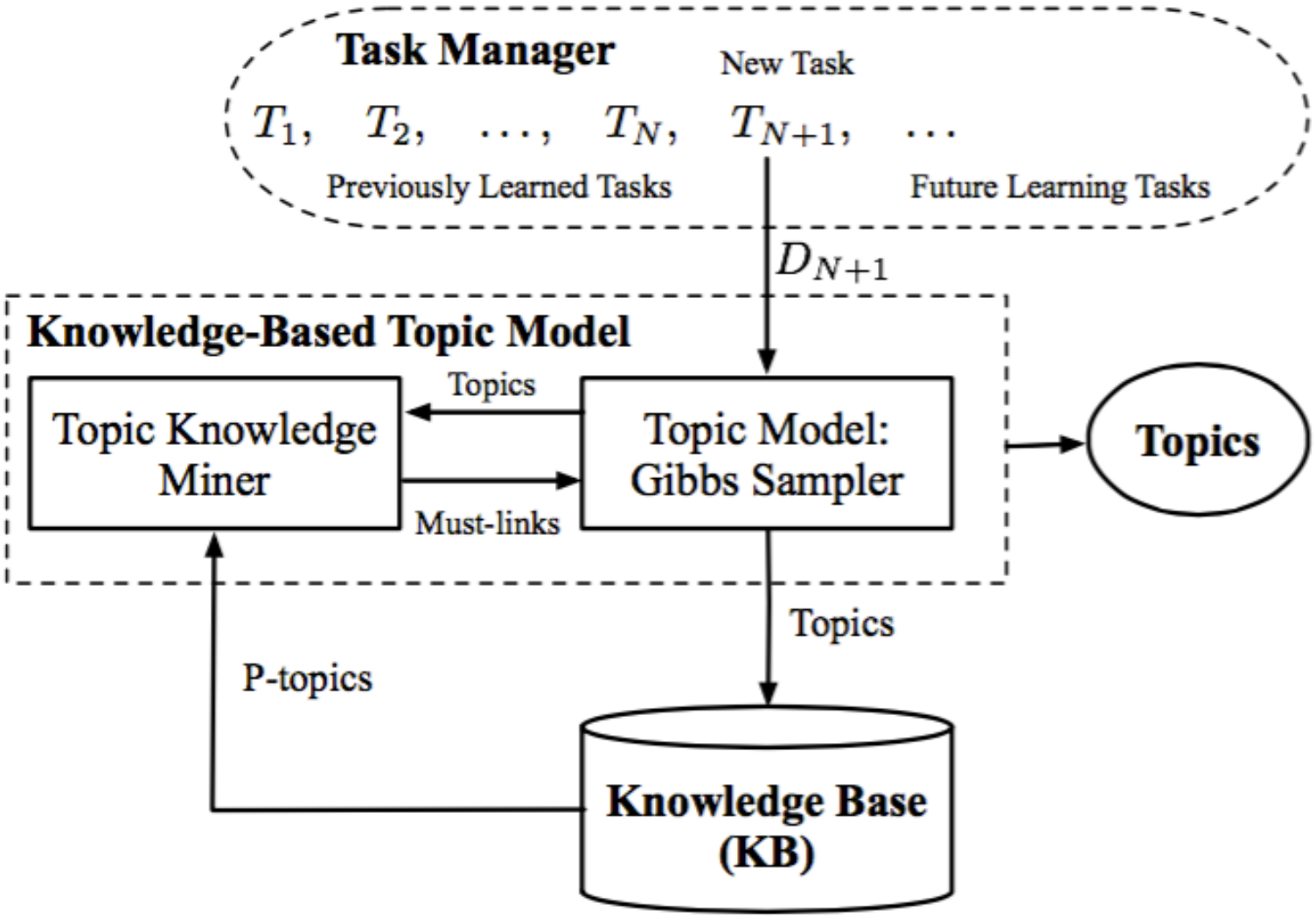
  (2) cluster them (synonym grouping).
  - {"picture," "photo"}, {"battery," "power"}
- Top modeling performs both tasks at the same time.
  - E.g., {*price*, *cost*, *cheap*, *expensive*, …}

Key observation in product reviews

(Chen and Liu, 2014)

- A fair amount of topic overlapping across reviews of different products or domains
  - Every product review domain has the aspect *price*,
  - Most electronic products share the aspect *battery*
  - Many also share the aspect of *screen*.
- This sharing of concepts / knowledge across domains is true in general.

LTM Architecture



An Example

- Given a newly discovered topic:

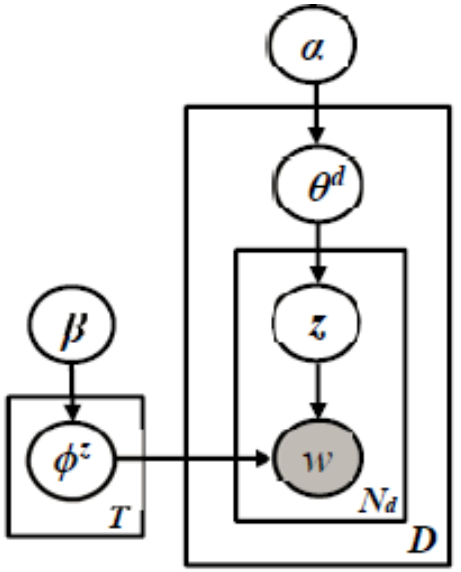
  {*price*, *book*, *cost*, *seller*, *money*}

  - 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*.

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

Outline

- A motivating example
- Definition of lifelong learning
- Lifelong supervised learning
- Lifelong unsupervised learning
- Learning in the open world - discovering new problems
- Learning in model testing/application
- Summary

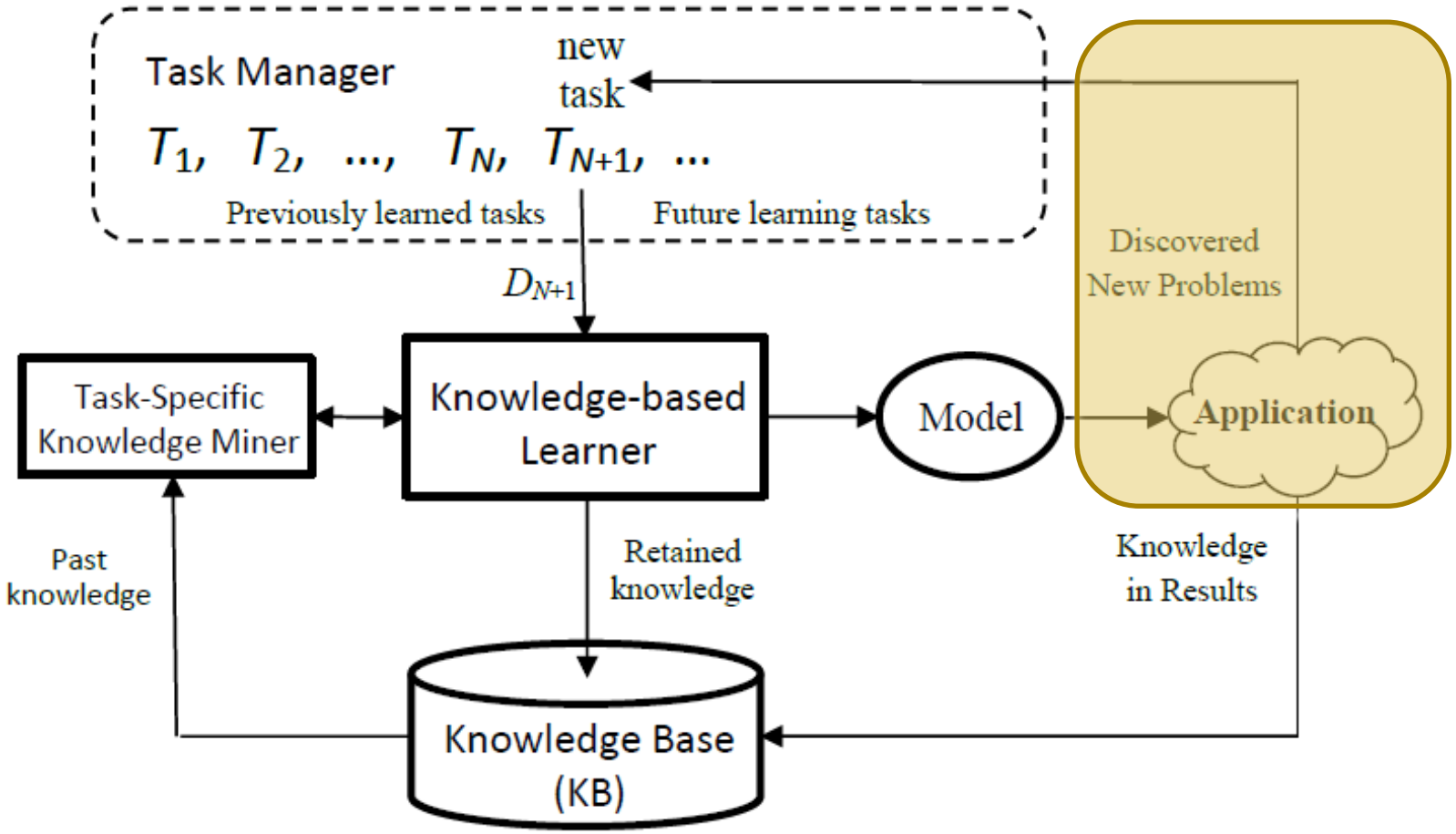
LML needs to identify unknowns

(Fei et al, 2016; Shu et al., 2017)

- If a system does not know what it knows & does not know, it
  - cannot function in a dynamic open world or learn new things by itself
    - E.g., self-driving cars
- 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}, $Y^{t+1} \in \{l_1, l_2, \ldots, l_t, l_0\}$
  - Incrementally learn new classes

Lifelong learning extended (1)

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

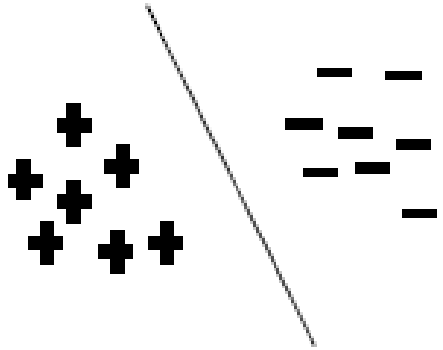


Open classification: CBS learning

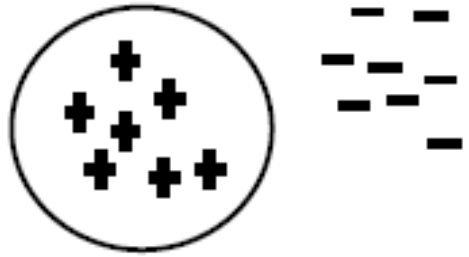
- To detect unseen classes, Fei and Liu (2016) proposed CBS learning:
  - Center-based similarity (CBS) space learning.
- It performs space transformation
  - Each document vector *d* is transformed to a CBS space vector
    (1) Compute centers c_i for the positive class
    (2) Compute the similarities of each document to c_i.
    - This gives us a new data set (in CSB space).

Traditional learning vs. CBS learning

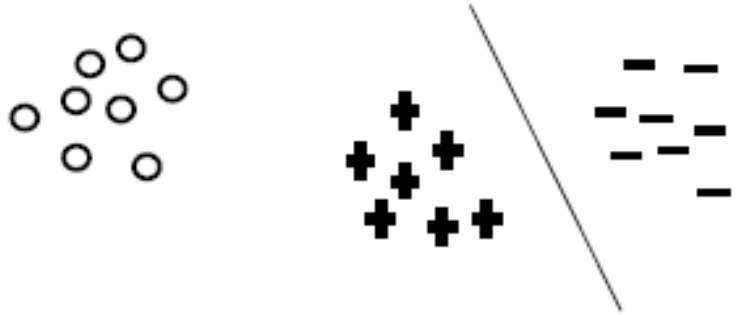
- Traditional learning (using SVM)

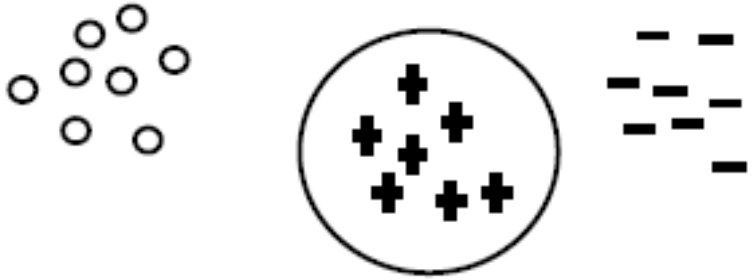


CBS learning



- Classification (testing)





DOC: Deep Open Classification of text documents

(Shu et al. 2017)

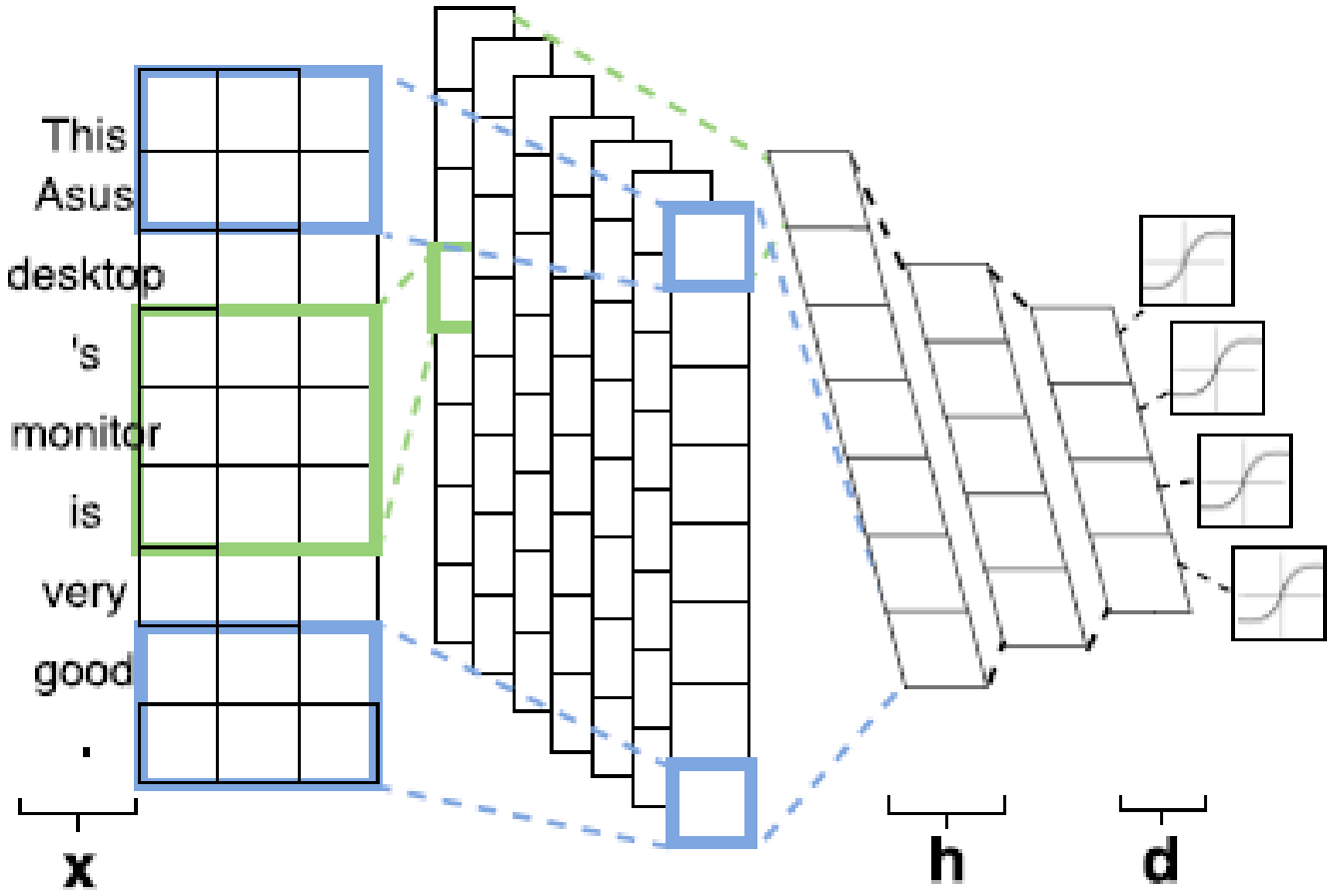


Figure 1: Overall Network of DOC

Finding the rejection threshold

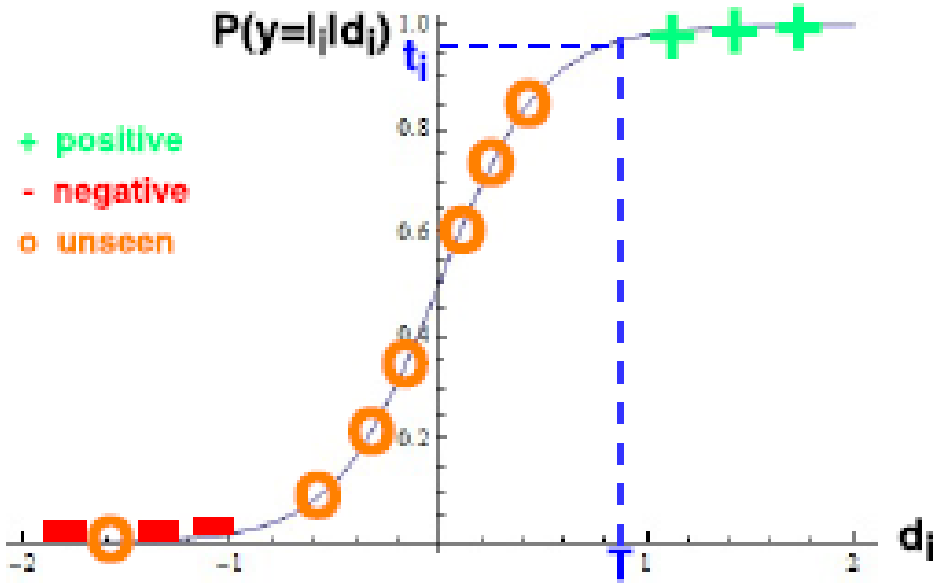


Figure 2: Open space risk of sigmoid function and desired decision boundary $d_i = T$ and probability threshold t_i.

Learning cumulatively

- Incrementally add a class without retraining from scratch

- "Human learning": uses the past knowledge F_t to help learn the new class I_{t+1}.
  - Find similar classes SC from known classes Y^t.
    - Old classes: Y^t = {movie, cat, politics, soccer}.
    - New class: I_{t+1}= basketball
    - SC = {soccer}
  - Building F_{t+1} by focusing on separating I_{t+1} and SC.

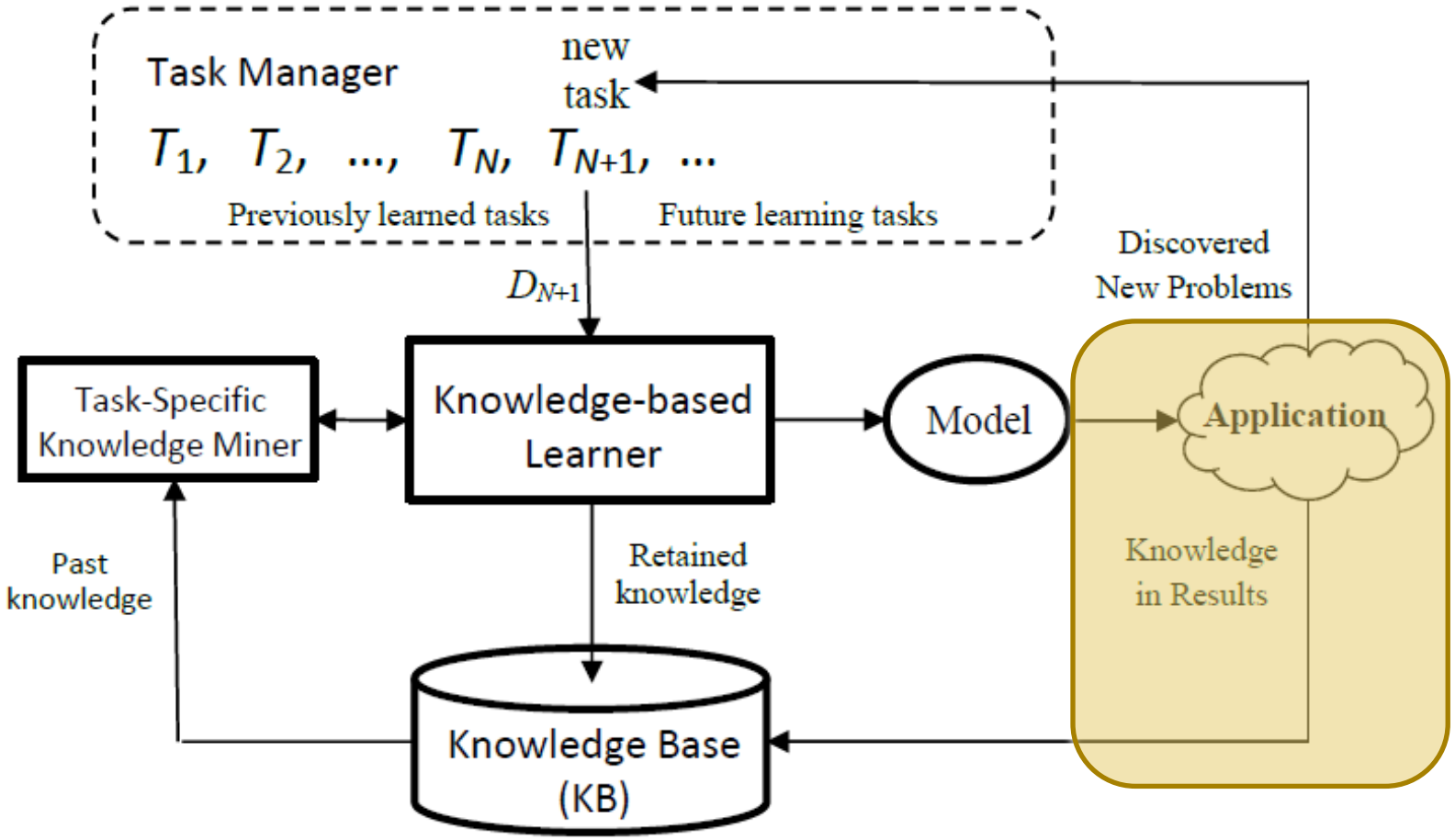
Outline

- A motivating example
- Definition of lifelong learning
- Lifelong supervised learning
- Lifelong unsupervised learning
- Learning in the open world - discovering new problems
- Learning in model testing/application
- Summary

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 → 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.

Lifelong learning extended (2)



Outline

- A motivating example
- Definition of lifelong learning
- Lifelong supervised learning
- Lifelong unsupervised learning
- Learning in the open world - discovering new problems
- Learning in model testing/application
- Summary

Summary

- This talk gave a brief introduction to LML and some NLP applications
- Existing LML research is still in its infancy
  - Understanding of LML is very limited
- There are huge challenges (Chen & Liu, 2016):
  - Correctness and applicability of knowledge, knowledge representation and reasoning, self-motivation, composition, learning in interaction, etc.
- It is a next natural step for ML and for NLU.

Lifelong Machine Learning

Download: https://www.cs.uic.edu/~liub/lifelong-machine-learning-draft.pdf

Thank you

Q&A