Co-Training for Topic Classification of Scholarly Data

Cornelia Caragea[1], Florin Bulgarov[1], and Rada Mihalcea[2]

[1]Computer Science and Engineering, University of North Texas
[2]Computer Science and Engineering, University of Michigan

EMNLP 2015

Scholarly Big Data

- Large number of scholarly documents on the Web

- PubMed currently has over 24 million documents

- Google Scholar is estimated to have 160 million documents

– Hence, *effective* and *efficient* methods for topic classification of research articles that can facilitate the retrieval of content that is tailored to the interests of specific individuals or groups are highly needed.

Previous Approaches to Topic Classification

- Many supervised approaches have been developed
  - Caragea et al. (2011) used the textual content of the target document and clustered words in an abstraction hierarchy in order to learn more robust model parameters.
  - Lu and Getoor (2003) proposed a model that incorporates both content and the citation relation between research articles.

- *However, to be successful, these supervised approaches require large amounts of labeled data.*
- *Hence, our question: Can we make an effective use of the large amounts of unlabeled data that, together with small amounts of labeled data, would result in accurate topic classification of research articles?*

Co-Training for Topic Classification

- We propose to explore an extension of co-training (Blum and Mitchell, 1998) for topic classification
  - In co-training, two classifiers trained on two different views of the data teach one another by re-training each classifier on data enriched with predicted examples that the other classifier is most confident about.

- *What can be the two views that describe the data in our domain?*

From Data to Knowledge

A typical scientific research paper:

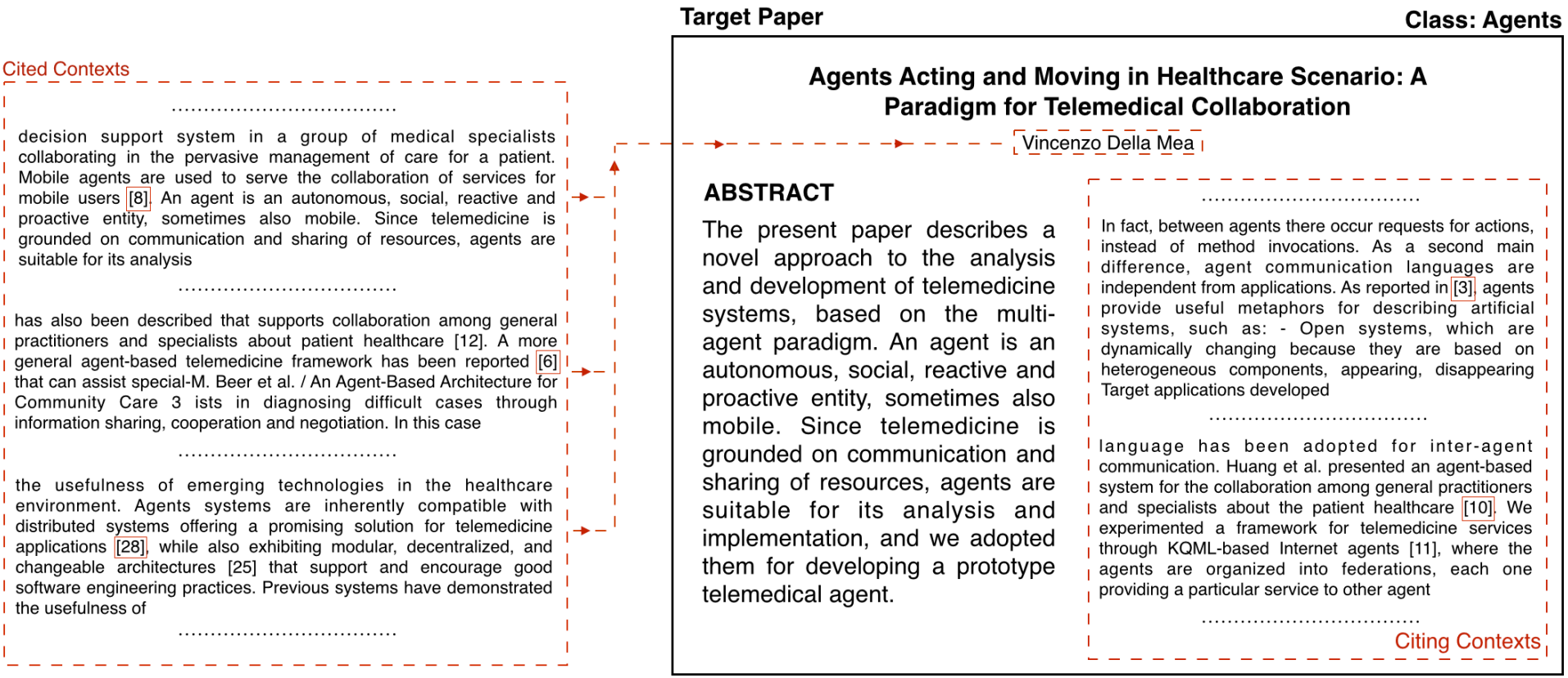
- Proposes new problems or extends the state-of-the-art for existing research problems
- Cites relevant, previously-published research papers in appropriate *contexts*.

The citations between research papers gives rise to an interlinked document network, commonly referred to as the *citation network*.

Citation Networks

- In a citation network, information flows from one paper to another via the citation relation (Shi et al, 2010)
- Citation contexts capture the influence of one paper on another as well as the flow of information
- Citation contexts or the short text segments surrounding a paper's mention serve as “micro summaries” of a cited paper!

A Small Citation Network



- Citation contexts capture how one paper influences another along various aspects such as topicality, domain of study, algorithms, etc.

Our Proposal for the Second View in Co-Training

- *Citation contexts* are very informative and can be used as an additional view in Co-Training for topic classification!

Co-Training for Topic Classification

Algorithm 1 Co-Training

Input: L, U, 's'

$L_1 \leftarrow L$, $L_2 \leftarrow L$

while $U \neq \varnothing$ **do**

Train classifier C_1 on L_1

Train classifier C_2 on L_2

$S \leftarrow \varnothing$

Move 's' examples from U to S

$U \leftarrow U \backslash S$

$S_1,\ S_2 \leftarrow$ *GetMostConfidentExamples*(S, C_1, C_2)

$L_1 \leftarrow L_1 \cup S_1$, $L_2 \leftarrow L_2 \cup S_2$

$U \leftarrow U \cup [S \backslash (S_1 \cup S_2)]$

end while

Ouput: The combined classifier C of C_1 and C_2

- In our co-training approach, all examples that are classified with a confidence higher than a certain threshold are moved into the labeled training set:
  - This is different from Blum and Mitchell's approach which moves 2(p+n) examples at each iteration (p:n is the ratio of positive to negative in the original labeled data)

Dataset

- The dataset used in our experiments is a subset sampled from the CiteSeer^x digital library, labeled by Dr. Lise Getoor's research group at the University of Maryland
- We obtained the citation contexts directly from the CiteSeerx digital library
  - Consists of 3,186 labeled papers
  - Each paper is categorized into one of six classes

Number of papers in each class						
Agents	AI	IR	ML	HCI	DB	Total
562	239	641	569	490	685	3186

Avg. Cited Contexts	Avg. Citing Contexts
45.59	20.77

Dataset summary

Experimental Setting

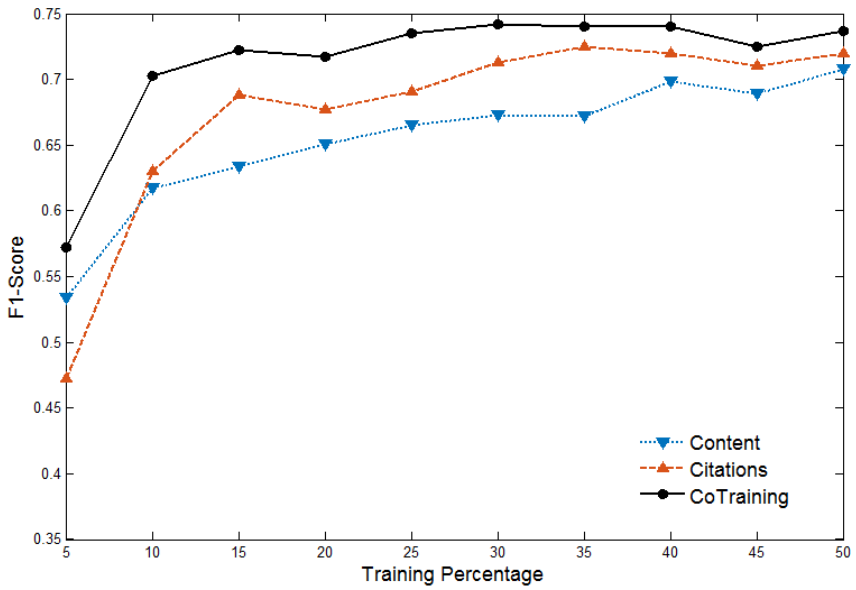
- Our experiments are designed to explore the following questions:
  - How does our co-training algorithm that uses content and citation contexts as two independent views of the data compare with supervised learning?
  - How does our co-training algorithm compare with semi-supervised learning, self-training and expectation maximization with Naïve Bayes?
  - How does co-training work in the absence of either citing or cited contexts?
    - A *cited context* for a document *d* is defined as a context in which *d is cited* by some paper d_i in the citation network.
    - A *citing context* for *d* is defined as a context in which *d is citing* some paper d_j in the citation network.

Experimental Setting

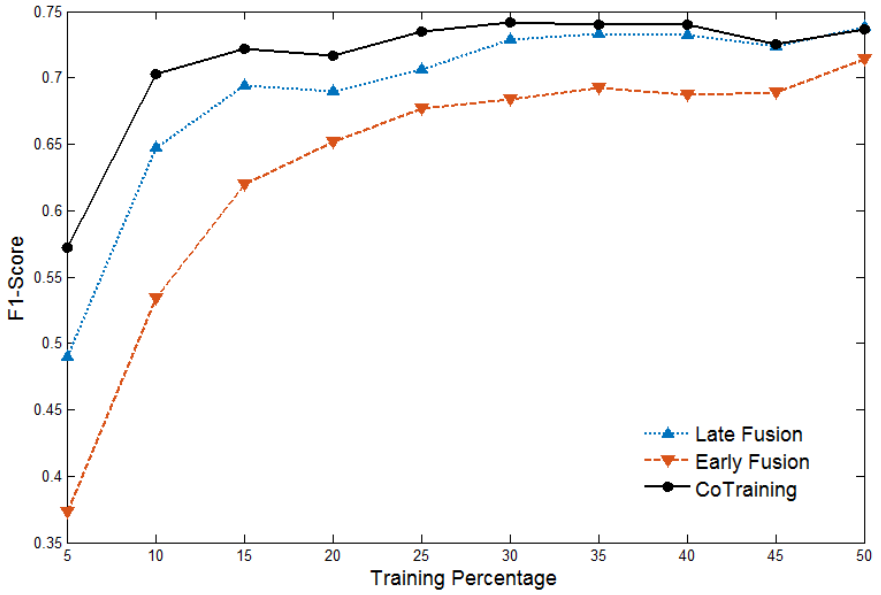
- In experiments, we split the CiteSeer^x sample of 3,186 papers in train *T*, validation *V*, and test *T*.
  - The validation and test sets have about 200 papers each.
  - We sample a set of papers from train with a fixed size of 2000 papers and use them as unlabeled data.
  - The remaining 786 papers are used as labeled training data.
- Each experiment is repeated 10 times with 10 different random splits of the data and the results are averaged.
- We used the Naïve Bayes Multinomial model on the "bag-of-words" representation of the data.

Results: Co-training vs. Supervised Learning

- How does co-training compare with supervised learning?



Co-Training vs. Each View

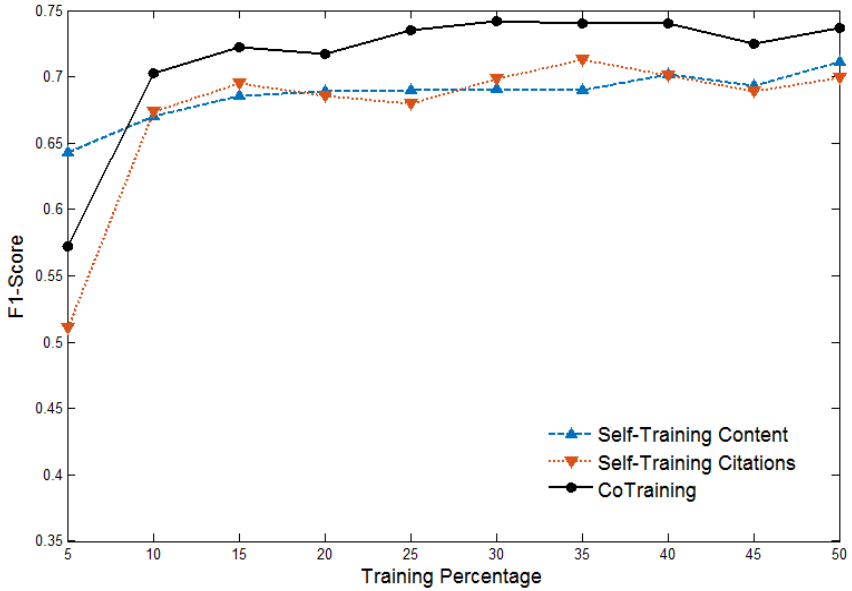


Co-Training vs. Early and Late Fusion

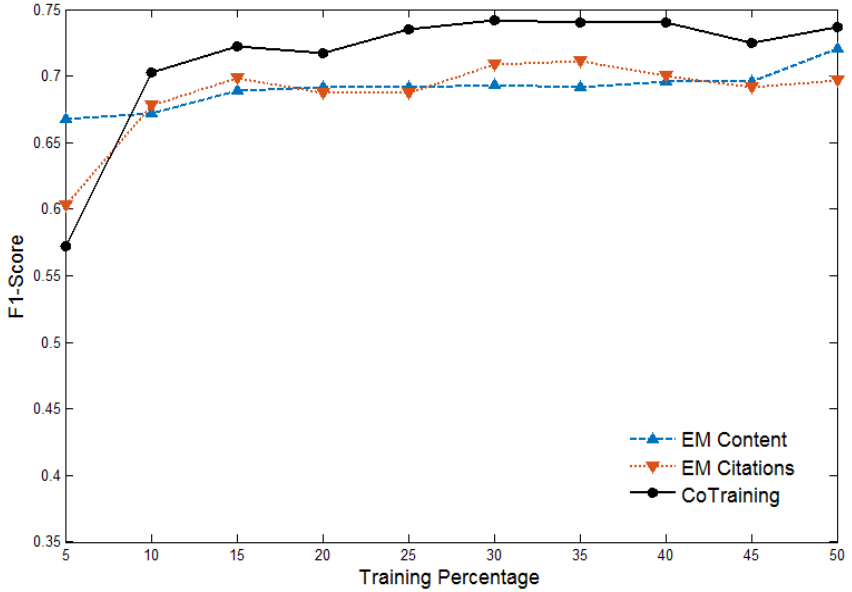
- Co-Training that uses citation contexts and content as two different views significantly outperforms supervised approaches that use either citation contexts or content, or their combination (as early or late fusion).

Results: Co-training vs. Semi-Supervised Learning

- How does co-training compare with semi-supervised learning?



Co-Training vs. Self-Training

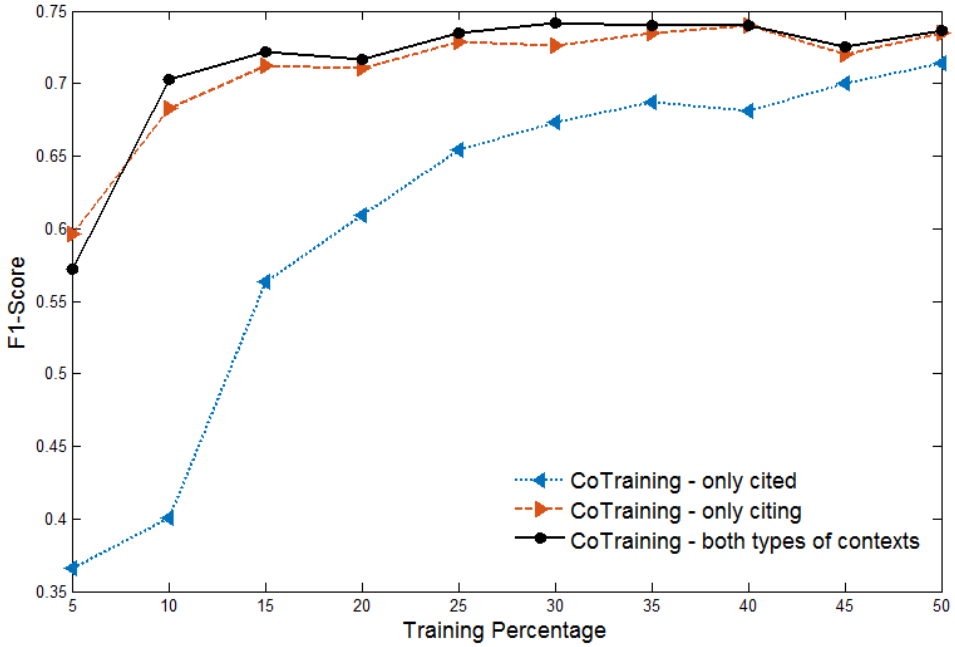


Co-Training vs. Expectation Maximization

- Co-Training that uses citation contexts and content as two different views significantly outperforms semi-supervised approaches: Self-Training and Expectation Maximization.

Results: Using Different Citation Context Types

- How does co-training work in the absence of either citing or cited context?



- Co-Training that uses only citing contexts and content as two different views performs similarly with the Co-Training that uses both citing and cited contexts and content.

Results: Co-Training Results on the Test Set

Method	Labeled docs. (%)	Precision	Recall	F1-Score
Co-Training	**30**	**0.749**	**0.743**	**0.742**
Co-Training - only citing	40	0.747	0.740	0.740
Co-Training - only cited	50	0.724	0.717	0.714
Self-Training - Content	50	0.723	0.711	0.711
Self-Training - Citations	35	0.730	0.710	0.713
EM - Content	50	0.738	0.714	0.721
EM - Citations	35	0.729	0.707	0.711
Early Fusion	50	0.718	0.710	0.714
Late Fusion	50	0.748	0.734	0.738
Content - Fully Supervised	100	0.730	0.728	0.720
Citations - Fully Supervised	100	0.745	0.740	0.738

- Results on the test set show that the proposed co-training method outperforms all compared models, reaching the highest F1-score of 0.742, while using the smallest amount of labeled documents, i.e. 30%.

Conclusions

- We proposed the use of citation contexts and content as two independent views in co-training for topic classification of research articles.
- Our results showed that co-training outperforms:
  - Supervised classifiers that use either content or citation contexts
  - Semi-supervised classifiers, trained on the same fractions of labeled and unlabeled data as co-training.
- The results also showed that, using citation contexts with content in co-training, the human effort involved in data labeling can be largely reduced.
- Future directions:
  - Explore Co-Training with contexts and content for other domains.
  - Investigate Co-Training with multiple views, e.g., citation contexts, content and link information.

Thank you!