Using Wikipedia for Co-clustering Based Cross-domain Text Classification

Pu Wang and **Carlotta Domeniconi**
George Mason University
Jian Hu
Microsoft Research Asia

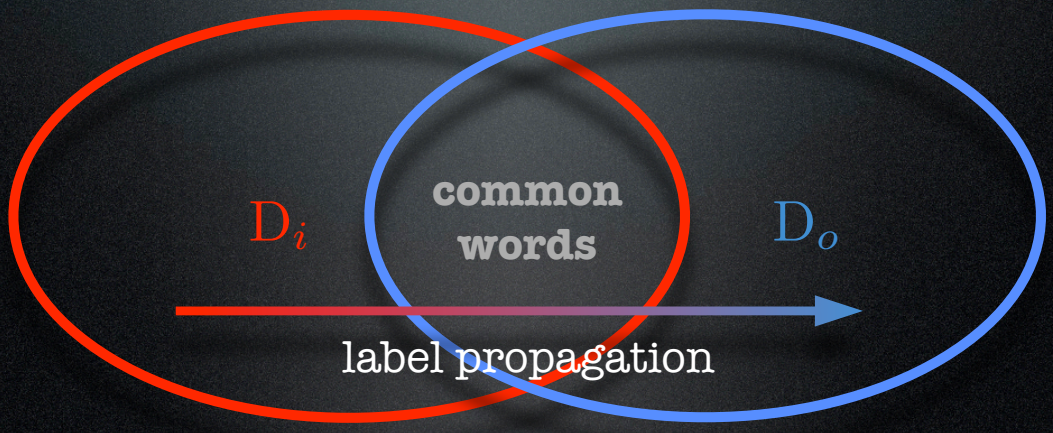
ICDM – Pisa, December 16, 2008

Motivation

- Labeled data are seldom available, and often too expensive to obtain.
- Abundant labeled data may exist for a different but related domain.
- **Goal**: Use the labeled data as auxiliary information to accomplish the task (classification) in the target domain.

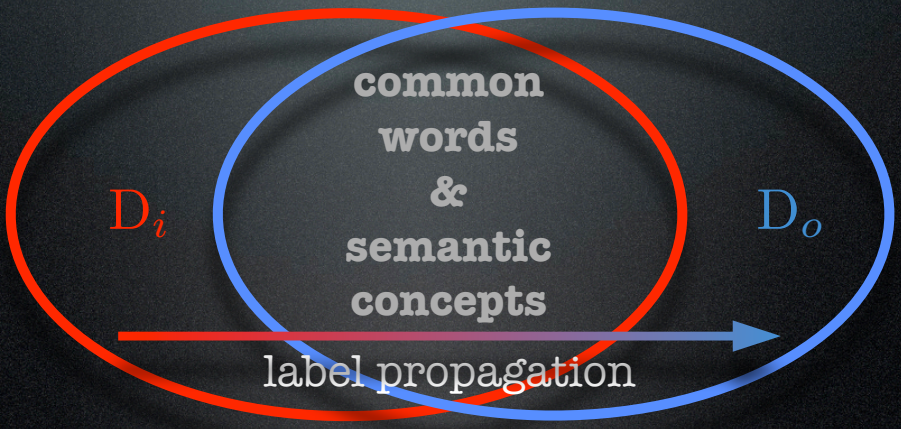
Main Idea

- Leverage the shared dictionary across the in-domain and out-of-domain (target) documents to propagate label information.



Main Idea

- Enrich document representation to fill the semantic gap.



Co-clustering based Classification (CoCC) [Dai et al., KDD 07]

- D_i: in-domain documents
- D_o: out-of-domain documents
- C : set of class labels
- W : dictionary of all the words

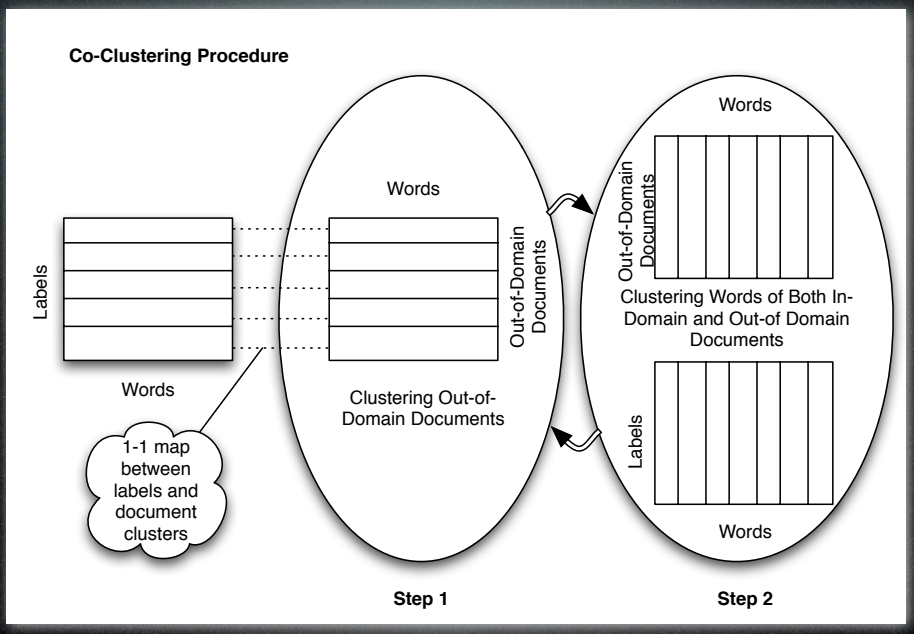
Co-clustering based Classification (CoCC) [Dai et al., KDD 07]

- Co-clustering of D_o:

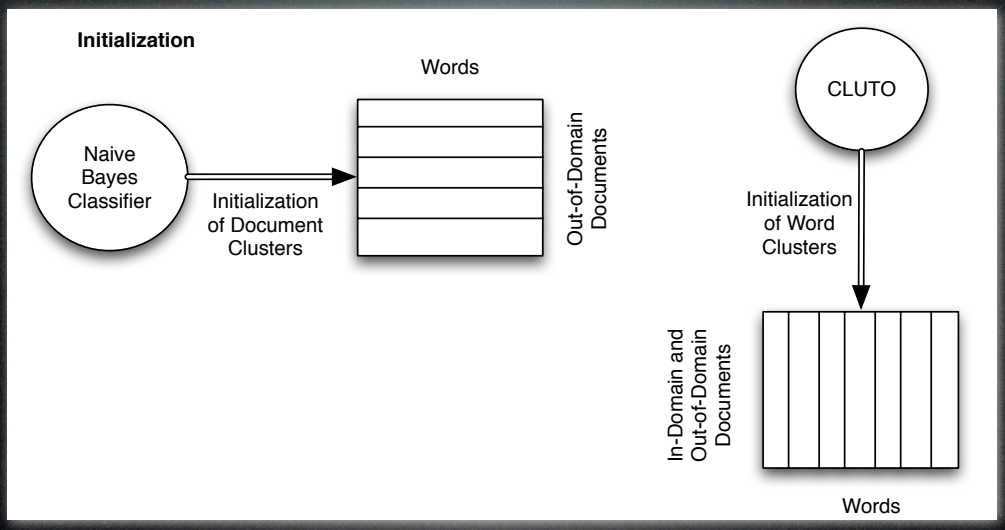
$$C_{D_o} : \{d_1, ..., d_m\} \rightarrow \{ \hat{d}_1, \hat{d}_2, ..., \hat{d}_{|C|}\} = \hat{D}_o$$

$$C_{\mathcal{W}} : \{w_1, ..., w_n\} \rightarrow \{ \hat{w}_1, \hat{w}_2, ..., \hat{w}_k\} = \hat{W}$$

Co-clustering based Classification (CoCC) [Dai et al., KDD 07]



Co-clustering based Classification (CoCC) [Dai et al., KDD 07]



Co-clustering based Classification (CoCC) [Dai et al., KDD 07]

- Iterative algorithm that achieves

$$\min_{\hat{\mathcal{D}}_o, \hat{\mathcal{W}}} \{ I(\mathcal{D}_o; \mathcal{W}) - I(\hat{\mathcal{D}}_o; \hat{\mathcal{W}}) + \lambda (I(\mathcal{C}; \mathcal{W}) - I(\mathcal{C}; \hat{\mathcal{W}})) \}$$

loss in mutual information between documents and words

loss in mutual information between class labels and words

Information Theoretic Co-clustering [Dhillion et al., KDD 03]

$$I(\mathcal{D}_o; \mathcal{W}) - I(\hat{\mathcal{D}}_o; \hat{\mathcal{W}})$$

$$I(X;Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

$$I(\mathcal{C}; \mathcal{W}) - I(\mathcal{C}; \hat{\mathcal{W}})$$

$$f(w) = \sum_{d \in \mathcal{D}_o} p(d, w), f(d|w) = p(d|w) = \frac{f(d, w)}{f(w)},$$

$$f(d) = \sum_{w \in \mathcal{W}} p(d, w), f(w|d) = p(w|d) = \frac{p(d, w)}{f(d)},$$

$$\hat{f}(\hat{w}|\hat{d}) = p(\hat{w}|\hat{d}), \hat{f}(\hat{d}|\hat{w}) = p(\hat{d}|\hat{w}),$$

$$\hat{f}(d|\hat{d}) = p(d|\hat{d}), \hat{f}(w|\hat{w}) = p(w|\hat{w}),$$

$$\hat{f}(d|\hat{w}) = \hat{f}(d|\hat{d})\hat{f}(\hat{d}|\hat{w}) = p(d|\hat{d})p(\hat{d}|\hat{w})$$

$$\hat{f}(w|\hat{d}) = \hat{f}(w|\hat{w})\hat{f}(\hat{w}|\hat{d}) = p(w|\hat{w})p(\hat{w}|\hat{d})$$

$$g(c, w) = p(c, \hat{w})p(w|\hat{w}) = p(c, \hat{w})\frac{p(w)}{p(\hat{w})}$$

$$g(w) = \sum_{c \in \mathcal{C}} p(c, w), g(c|w) = p(c|w) = \frac{g(c, w)}{g(w)},$$

$$\hat{g}(c|\hat{w}) = \frac{\sum_{w \in \hat{w}} p(c|w)p(w)}{p(\hat{w})} = \frac{\sum_{w \in \hat{w}} p(c|w)p(w)}{\sum_{w \in \hat{w}} p(w)}.$$

Co-clustering based Classification (CoCC) [Dai et al., KDD 07]

$$I(\mathcal{D}_o;\mathcal{W}) - I(\hat{\mathcal{D}}_o;\hat{\mathcal{W}}) + \lambda I(\mathcal{C};\mathcal{W}) - I(\mathcal{C};\hat{\mathcal{W}})$$
$$= D(f(\mathcal{D}_o;\mathcal{W})||\hat{f}(\mathcal{D}_o;\mathcal{W})) + \lambda D(g(\mathcal{C},\mathcal{W})||\hat{g}(\mathcal{C},\mathcal{W}))$$

$$D(p(x)||q(x)) = \sum_x p(x)\log\frac{p(x)}{q(x)}$$

Co-clustering based Classification (CoCC) [Dai et al., KDD 07]

$$D(f(\mathcal{D}_o,\mathcal{W})||\hat{f}(\mathcal{D}_o,\mathcal{W}))$$
$$= \sum_{\hat{d}\in\hat{\mathcal{D}}_o}\sum_{d\in\hat{d}} f(d)D(f(\mathcal{W}|d)||\hat{f}(\mathcal{W}|\hat{d}))$$
$$D(f(\mathcal{D}_o,\mathcal{W})||\hat{f}(\mathcal{D}_o,\mathcal{W}))$$
$$= \sum_{\hat{w}\in\hat{\mathcal{W}}}\sum_{w\in\hat{w}} f(w)D(f(\mathcal{D}_o|w)||\hat{f}(\mathcal{D}_o|\hat{w}))$$
$$D(g(\mathcal{C},\mathcal{W})||\hat{g}(\mathcal{C},\mathcal{W}))$$
$$= \sum_{\hat{w}\in\hat{\mathcal{W}}}\sum_{w\in\hat{w}} g(w)D(g((\mathcal{C}|w)||\hat{g}(\mathcal{C}|\hat{w})))$$

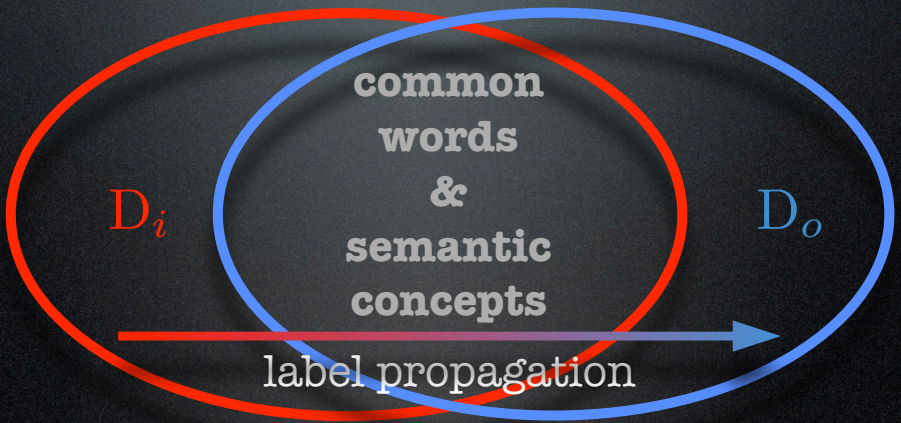
Co-clustering based Classification (CoCC) [Dai et al., KDD 07]

$$\mathcal{C}_{\mathcal{D}_o}^{(t)}(d) = \operatorname*{argmin}_{\hat{d}} D(f(\mathcal{W}|d)||\hat{f}^{(t-1)}(\mathcal{W}|\hat{d}))$$

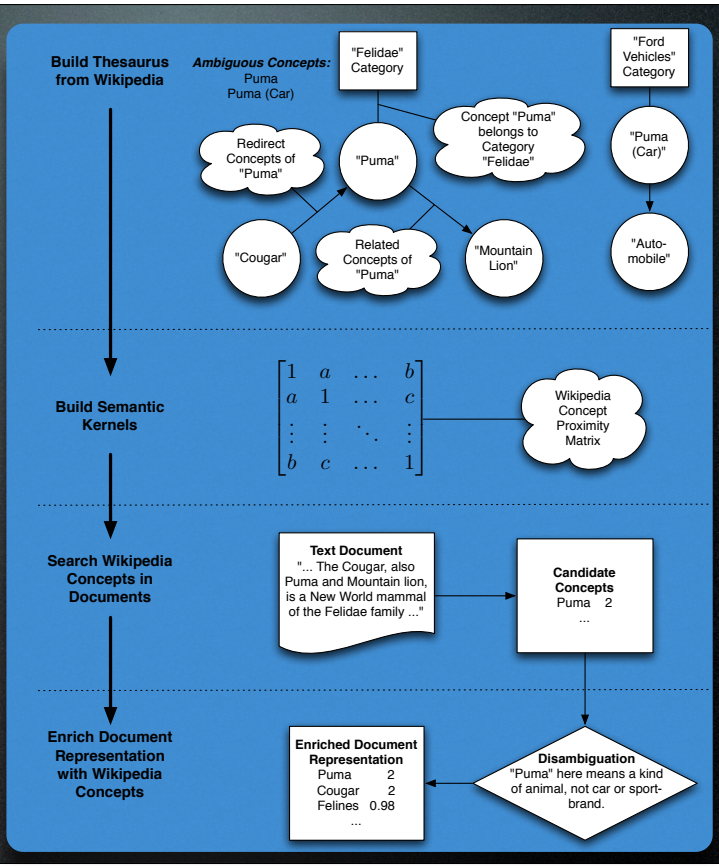
$$\mathcal{C}_{\mathcal{W}}^{(t+1)}(d) = \operatorname*{argmin}_{\hat{w}} f(w)D(f(\mathcal{D}_o|w)||\hat{f}(\mathcal{D}_o|\hat{w})) + \lambda g(w)D(g((\mathcal{C}|w)||\hat{g}(\mathcal{C}|\hat{w})))$$

Main Idea

- Enrich document representation to fill the semantic gap.



Building Semantic Kernels from Wikipedia: Overall Approach



Proximity Matrix

	Terms				Concepts			
Terms	1	0	⋯	0	0	0	⋯	0
	0	1	⋯	0	0	0	⋯	0
	⋮	⋮	⋱	⋮	⋮	⋮	⋱	⋮
	0	0	⋯	1	0	0	⋯	0
Concepts	0	0	⋯	0	1	a	⋯	b
	0	0	⋯	0	a	1	⋯	c
	⋮	⋮	⋱	⋮	⋮	⋮	⋱	⋮
	0	0	⋯	0	b	c	⋯	1

$$S = \lambda_1 S_{BOW} + \lambda_2 S_{OLC} + (1 - \lambda_1 - \lambda_2)(1 - D_{cat})$$

Content-based

Outlink category-based

Distance-based

Proximity Matrix

	Terms				Concepts			
Terms	1	0	⋯	0	0	0	⋯	0
	0	1	⋯	0	0	0	⋯	0
	⋮	⋮	⋱	⋮	⋮	⋮	⋱	⋮
	0	0	⋯	1	0	0	⋯	0
Concepts	0	0	⋯	0	1	a	⋯	b
	0	0	⋯	0	a	1	⋯	c
	⋮	⋮	⋱	⋮	⋮	⋮	⋱	⋮
	0	0	⋯	0	b	c	⋯	1

$$P_{ij} = \begin{cases} 1 & \text{if } c_i \text{ and } c_j \text{ are synonyms;} \\ \mu^{-depth} & \text{if } c_i \text{ and } c_j \text{ are hyponyms;} \\ S & \text{if } c_i \text{ and } c_j \text{ are associative concepts;} \\ 0 & \text{otherwise.} \end{cases}$$

$$S = \lambda_1 S_{BOW} + \lambda_2 S_{OLC} + (1 - \lambda_1 - \lambda_2)(1 - D_{cat})$$

Content-based

Outlink category-based

Distance-based

Building Semantic Kernels

Machine learning, statistical learning and data mining are related subjects.

Original BOW Vector

<machine:1, statistical:1, learn:2, data:1, mine:1, relate:1, subject:1>

Find Wikipedia Concepts and Keep as it is

$\phi(d)$ → <relate:1, subject:1; **machine learning:1, statistical learning:1, data mining:1**; ... >

×

P →

...	Machine Learning	Statistical Learning	Data Mining	Artificial Intelligence	...
Machine Learning	1	0.6276	0.4044	0.1314	...
Statistical Learning	0.6276	1	0.2839	0.1146	...
Data Mining	0.4044	0.2839	1	0.0792	...
Artificial Intelligence	0.1314	0.1146	0.0792	1	...
...				...	

$\tilde{\phi}(d) = \phi(d)P$ →

||

Enriched Document Vector Representation

<relate:1, subject:1; **machine learning:1, statistical learning:1, data mining:1**; *artificial intelligence:0.3252*>

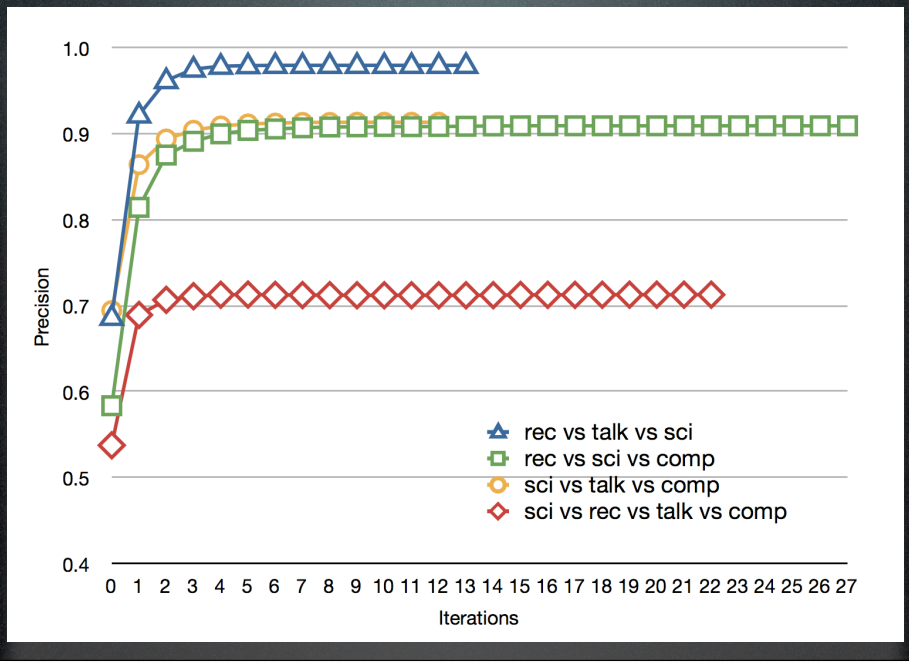
Empirical Evaluation

- **Data sets**: 20Newsgroups and SRAA
- **Methods**:
  - CoCC w/ and w/out enrichment
  - NB w/ and w/out enrichment

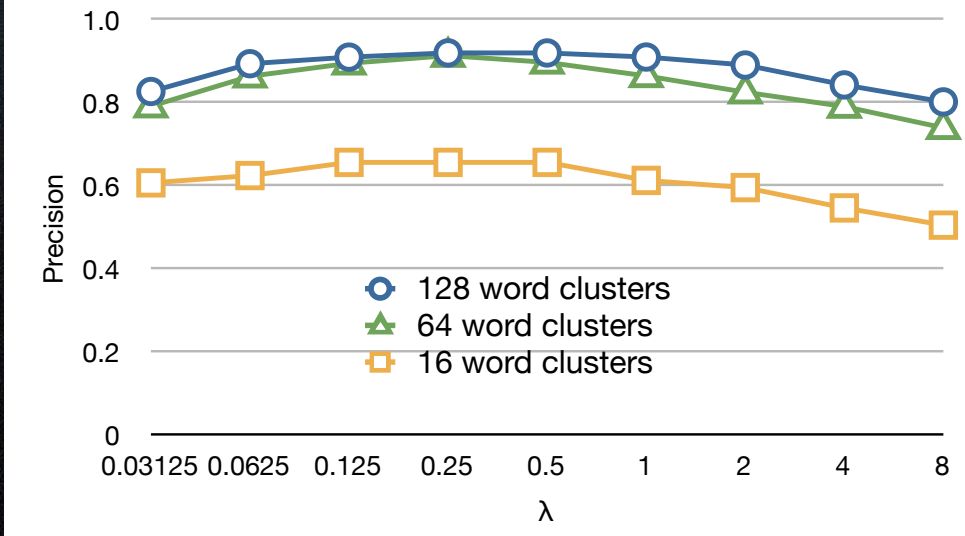
Cross-domain Classification Precision Rates

Data Set	w/o enrichment		w/ enrichment	
	NB	CoCC	NB	CoCC
rec vs talk	0.824	0.921	0.853	0.998
rec vs sci	0.809	0.954	0.828	0.984
comp vs talk	0.927	0.978	0.934	0.995
comp vs sci	0.552	0.898	0.673	0.987
comp vs rec	0.817	0.915	0.825	0.993
sci vs talk	0.804	0.947	0.877	0.988
rec vs sci vs comp	0.584	0.822	0.635	0.904
rec vs talk vs sci	0.687	0.881	0.739	0.979
sci vs talk vs comp	0.695	0.836	0.775	0.912
rec vs talk vs sci vs comp	0.487	0.624	0.538	0.713
real vs simulation	0.753	0.851	0.826	0.977
auto vs aviation	0.824	0.959	0.933	0.992

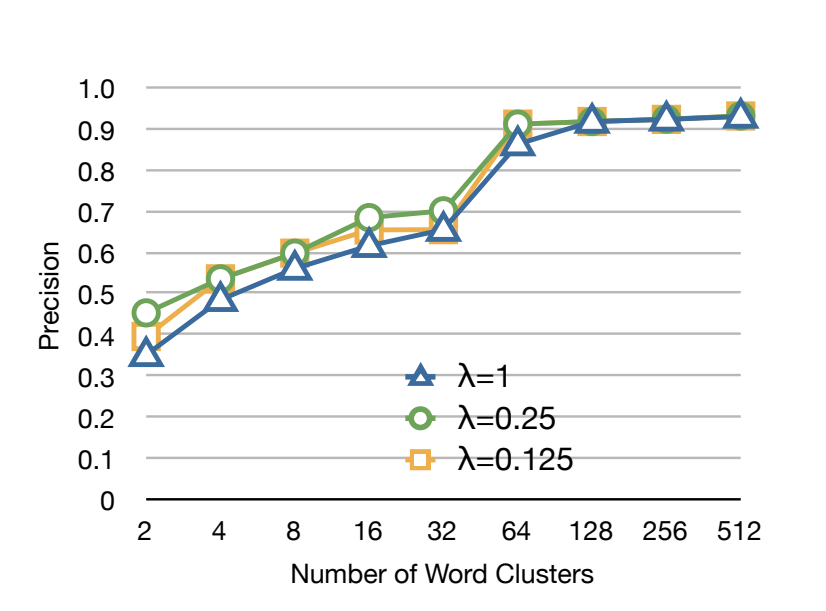
CoCC with enrichment: Precision as a function of the number of iterations



CoCC with enrichment: Precision as a function of λ (sci vs talk vs comp)



CoCC with enrichment:
Precision as a function of the number of word clusters (sci vs talk vs comp)



Conclusions

- Extended co-clustering approach for cross-domain text classification by embedding background knowledge using Wikipedia
- Future work:
  - Explore alternative representations for common language substrate
  - Cross-language text classification