

Cross-training: Learning probabilistic mappings between topics

Sunita Sarawagi

Soumen Chakrabarti
IIT Bombay

Shantanu Godbole

SIGKDD'04

Presented by Gabriel Ripoche
CS591CXZ – Text Mining
October, 13th 2004

Problem – Why cross-training?

Given a set of documents annotated using different taxonomies:

- How can these taxonomies be **reconciled**?
- How can these taxonomies lead to **more accurate classification**?

Why is this an interesting problem?

- **No global standards**
- **Evolving taxonomies**

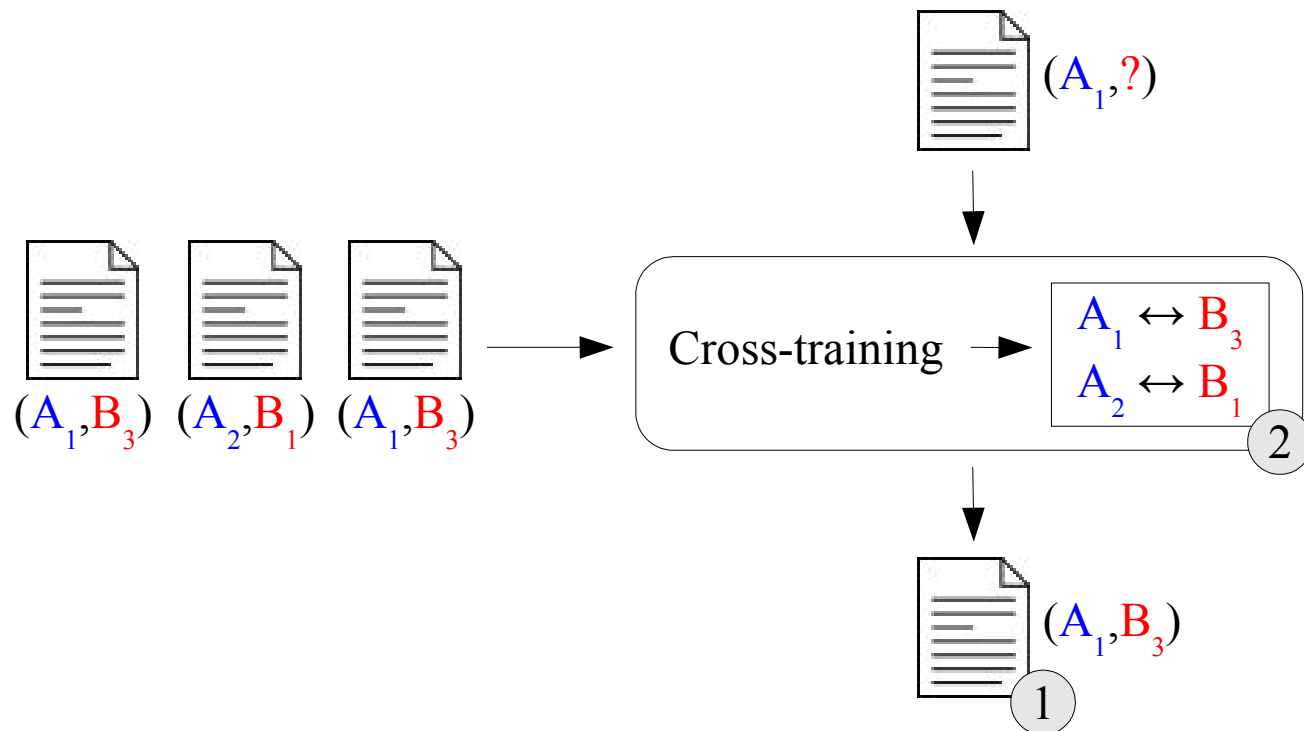
Applications:

- Semantic web
- Directories, catalogues, etc.

Concept – What is cross-training?

Cross-training:

- ① Use label assignments from taxonomy A to **make better inferences** about label assignments for taxonomy B
- ② Use inferred relationships between A and B to **map** A and B



Method 1: EM2D – Cross-trained naive bayes with EM

Principle:

- Use **expectation maximization** (EM) for semi-supervised learning. (small annotated set + large training set of unlabeled documents)
- Learning two label taxonomies (A, B) = learning over A x B matrix

How to do it?

- In EM update rule, use known label to **restrict contribution** of training document to possible subset of labels
- E.g.: If A_2 is known, only $P(B_x|A_2)$ should be updated, not all $P(B_x|A_y)$

		Taxonomy A			
		A ₁	A ₂	A ₃	A ₄
Tax. B	B ₁				
	B ₂				
	B ₃				

Note: Initialization is important! Paper discusses ways to initialize EM

Method 2: SVM-CT – Cross-trained SVM

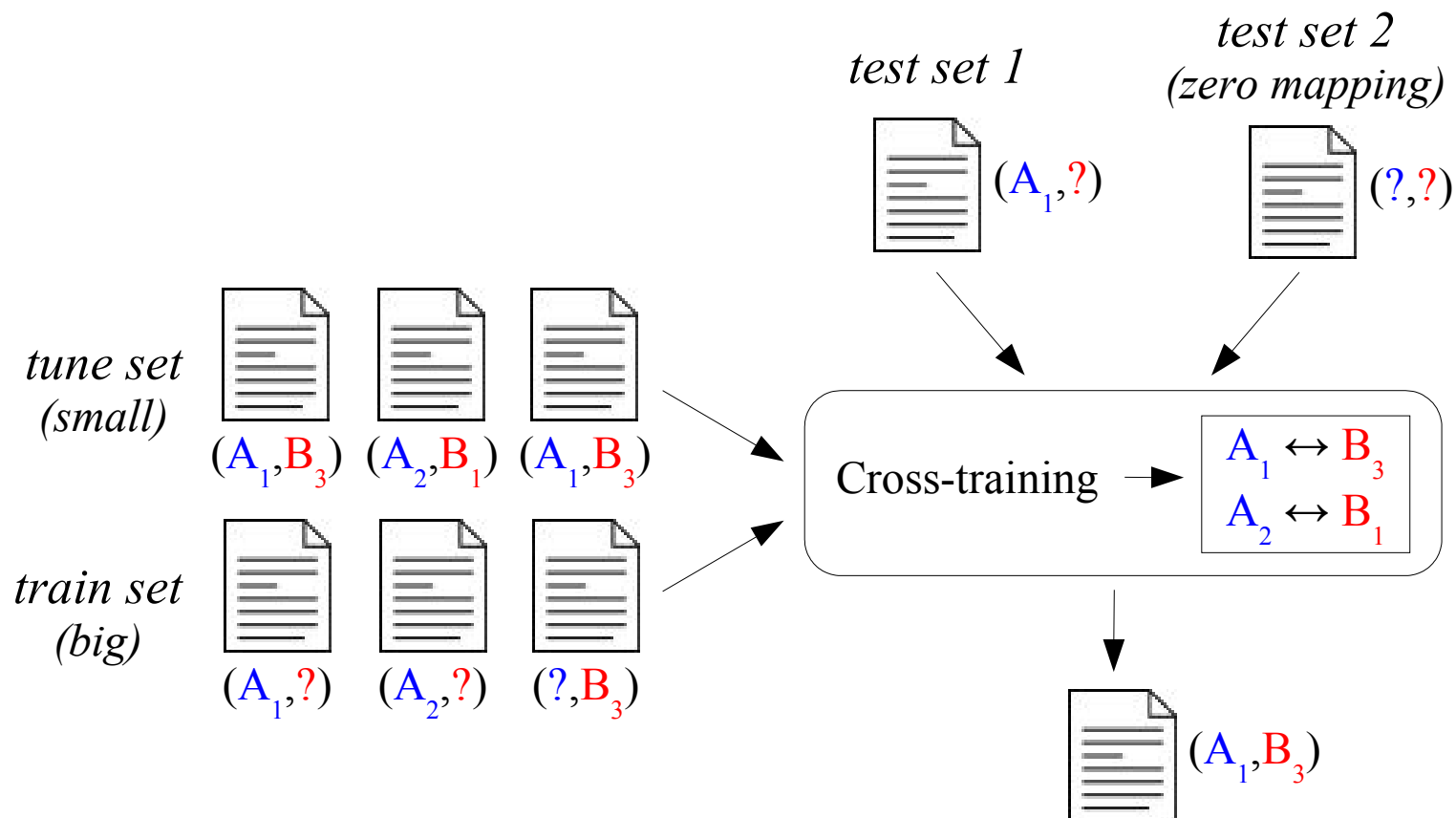
Principle:

- For each taxonomy, learn SVM classifiers over both document terms + additional features representing labels of taxonomy.
- Iterate + alternate
(learn SVM-A over B features, then SVM-B over A features, ...)

How to do it?

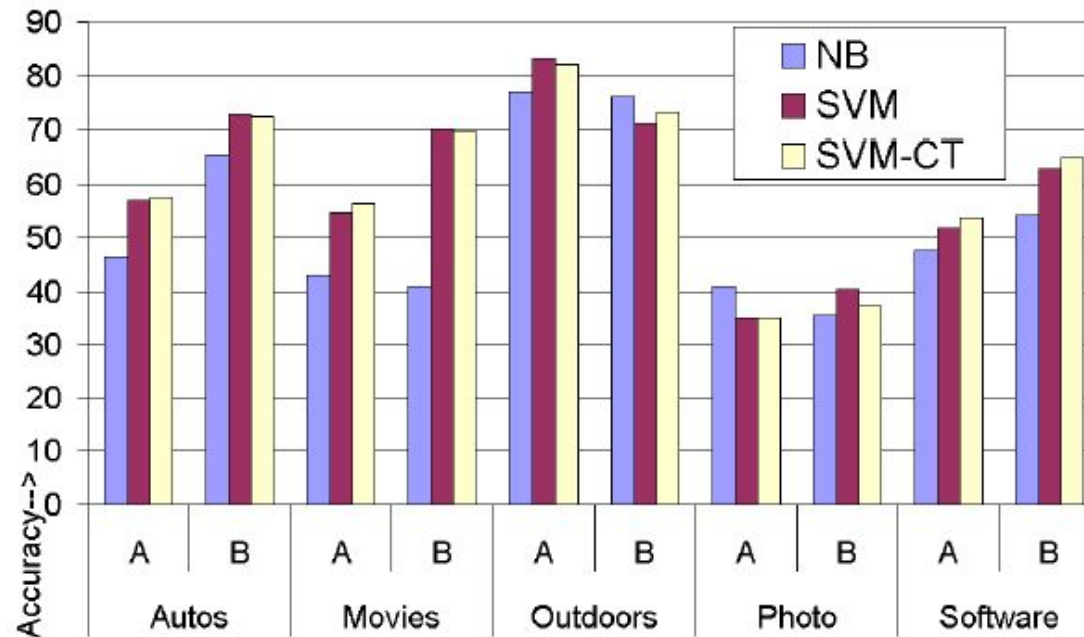
- Add N additional features corresponding to each possible label from other taxonomy
- E.g.: When learning SVM for taxonomy B, add |A| features, one for each label in $A = (A_1, A_2, A_3, A_4)$

Training and evaluation settings



Experiment 1: NB vs. SVM

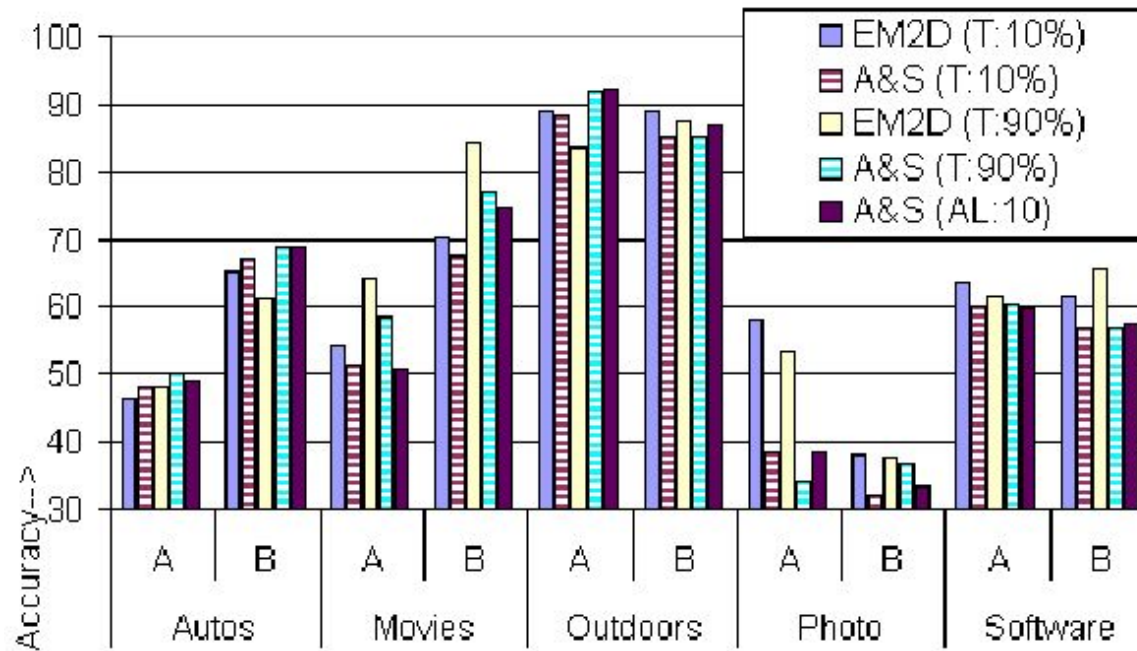
- Comparing baseline Naive Bayes, SVM, and SVM-CT



- SVM and SVM-CT are about the same
- SVM beats NB in almost all cases
- WHY use NB in rest of experiments if SVM is better?!?

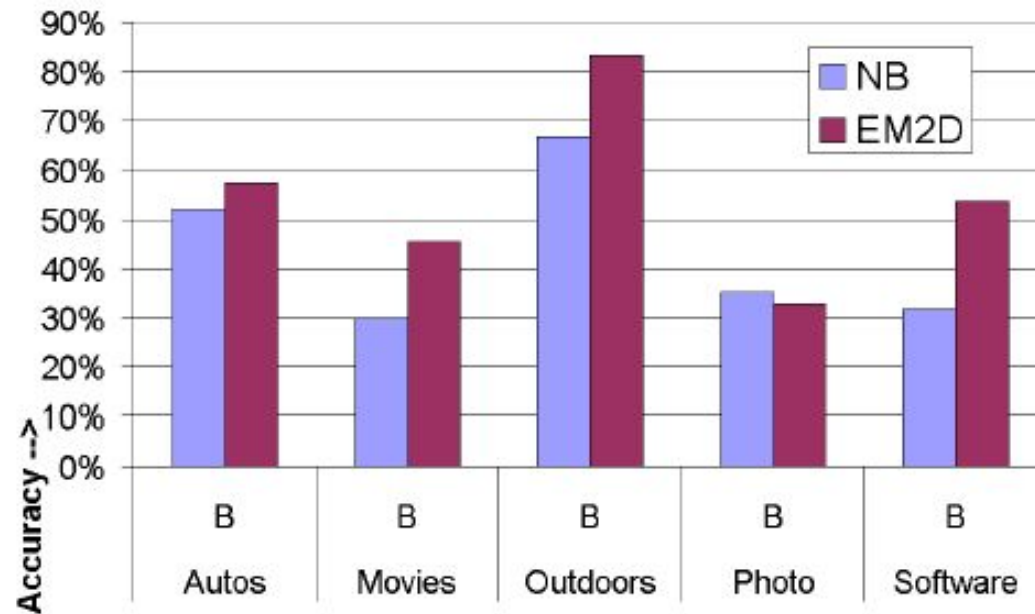
Experiment 2: EM2D vs. A&S

- How does EM2D compare to state-of-the-art?



Experiment 3: EM2D asymmetric

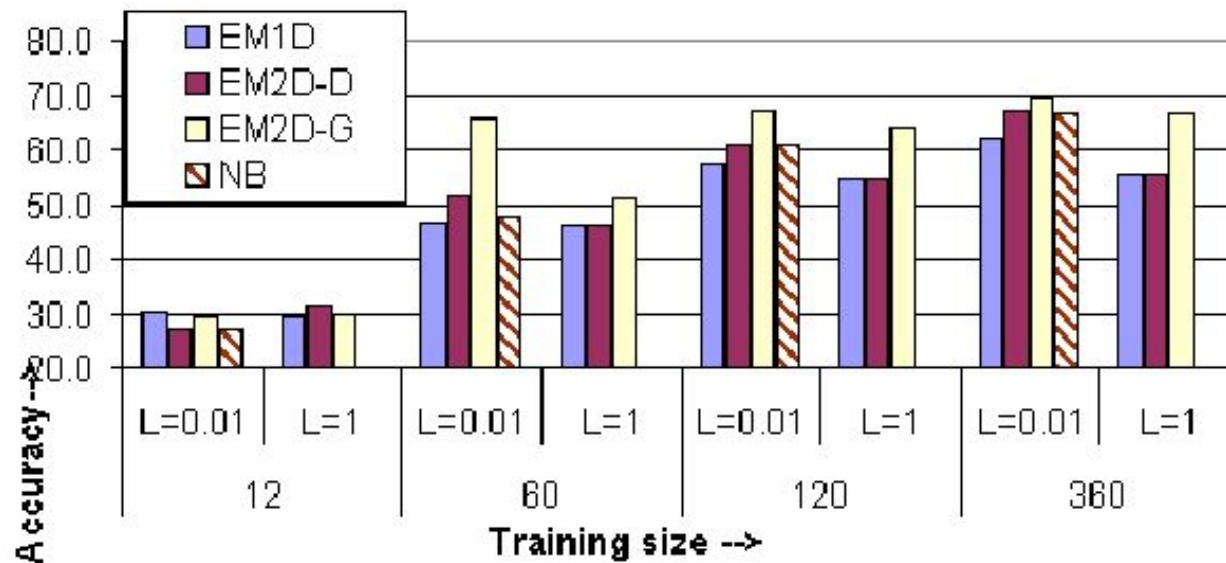
- Can cross-training work if one taxonomy has very few examples?



- Smearing in $D_A - D_B$ over all B labels (don't rely on B label assignments)
- Small damping factor for $D_A - D_B$ (docs in $D_A - D_B$ don't count for much)
- Tuneset used (use reliable data for the real cross-training)

Experiment 4: EM2D zero-label

- Can cross-training guess **both labels** better?



- EM1D = NB + EM (no cross-training)
- EM2D-D = EM2D + model aggregation (?)
- EM2D-G = EM2D + guess (first guess one label, then use it to guess other)

Experiment 5: SVM-CT mapping

- Feature weights in SVM-CT give interesting mappings

- ① Learns 1 to 1 mappings
- ② Learns 1 to N mappings (parent-child)
- ③ Learns non existent mappings

Dataset	Dmoz.	Maps to Yahoo.	Weight
Autos	News&Magazines	News&Media	0.147
		Volkswagen	-0.156
Movies	Genres/Western	Titles/Western	0.242
		Titles/Horror	-0.052
Outdoors	Scuba Diving	Scuba	5.878
		Snowmobiling	-0.647
Photo	Techs&Styles	Pinhole Ph'graphy	2.796
		3D	0.964
		Panoramic	0.921
Software	Accounting	Organizations	-1.184
		NOTA	0.156
		Screen Savers	0.103
		OS/Unix	-0.171
Dataset	Yahoo.	Maps to Dmoz.	Weight
Autos	Corvette	Chevrolet	0.981
		Parts&Accessories	-0.266
Movies	SciFi&Fantasy	Series/Star-Wars	1.123
		Reviews	-0.824
Outdoors	Scuba	Scuba Diving	4.822
		Wildlife	-0.437
Photo	Pinhole Ph'graphy	Techs&Styles	0.4842
		Photographers	-0.270
Software	OS/MSWindows	OS/MSWindows	0.018
		NOTA	-0.001
		OS/Unix	-0.008

①

②

③

②

Conclusion, questions, and critiques

Conclusions:

- Method presented for **cross-training** using multiple taxonomies
- Many experiments cover **different situations**
- **Direct applications** in ontology merging, semantic web, ...

Questions & Critiques:

- Compare SVM-CT and EM2D? SVM > NB (1), is **SVM-CT > EM2D** ?
- Comparing EM2D and A&S (2): A&S is crippled, is that **fair**?
- Asymmetric scenario (3): isn't their method a **hack**?
- EM2D-G is best in zero-label (4). Why is that so? Why is it surprising?
- Paper poorly organized