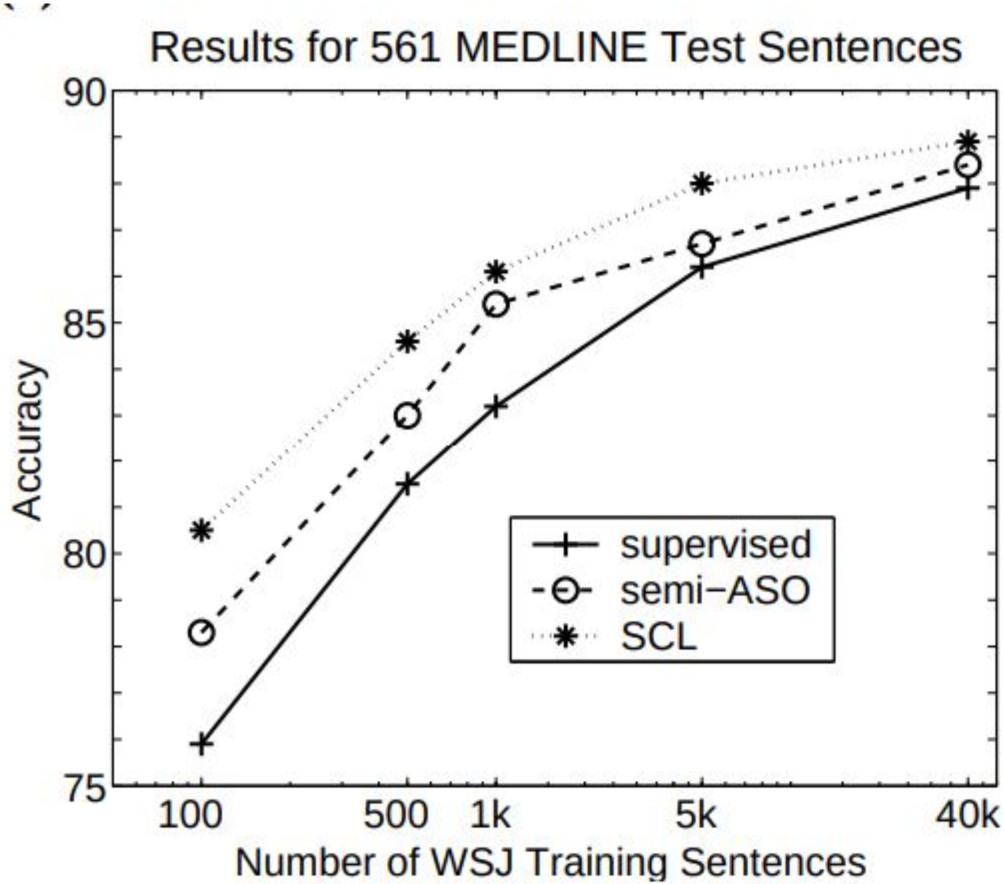
Domain Adaptation and Transfer Learning in NLP

Krishnapriya V

The need for DA

- Part of speech tagging
- Trained on: WSJ
- Tested on: Other domains

Blitzer et al, 2006



Ruder et al, 2018

Model	**Target domains test sets**					**Avg on**	
	Answers	**Emails**	**Newsgroups**	**Reviews**	**Weblogs**	**targets**	**WSJ**
TnT*	89.36	87.38	90.85	89.67	91.37	89.73	96.57
Stanford*	89.74	87.77	91.25	90.30	92.32	90.28	97.43
Src (+glove)	90.43 $\pm$.13	87.95 $\pm$.18	91.83 $\pm$.20	90.04 $\pm$.11	92.44 $\pm$.14	90.54 $\pm$.15	**97.50** $\pm$.03
Tri	**91.21** $\pm$.06	**88.30** $\pm$.19	**92.18** $\pm$.19	**90.06** $\pm$.10	**92.85** $\pm$.02	**90.92** $\pm$.11	97.45 $\pm$.03
Asym	90.62 $\pm$.26	87.71 $\pm$.07	91.40 $\pm$.05	89.89 $\pm$.22	92.37 $\pm$.27	90.39 $\pm$.17	97.19 $\pm$.03
MT-Tri	90.53 $\pm$.15	87.90 $\pm$.07	91.45 $\pm$.19	89.77 $\pm$.26	92.35 $\pm$.09	90.40 $\pm$.15	97.37 $\pm$.07
FLORS*	91.17	88.67	92.41	92.25	93.14	91.53	97.11

- Models assume train data and test data are from the same distribution
- Consider a discriminative model
- Model family: $P(Y|X, \theta)$
- Find:

$$argmax_\theta \int_\chi \sum_\gamma p(x, y) log p(y|x; \theta) dx$$

- We don't know $p(x,y)$
- Approximate it with observed $p'(x,y)$
- Problems when this observed data is not representative of real world $p_t(x,y)$

Notations

- Domain : (\chi, P(X))
- Task : (\gamma, P(y|x))
-
- Different domains -
    - P(X) - domain adaptation
    - X - multilingual setting
-
- Different tasks -
    - P(Y|X) - class imbalance
    - P(Y) - Multitask learning

From (Pan and Yang, 2010): A Survey on Transfer Learning

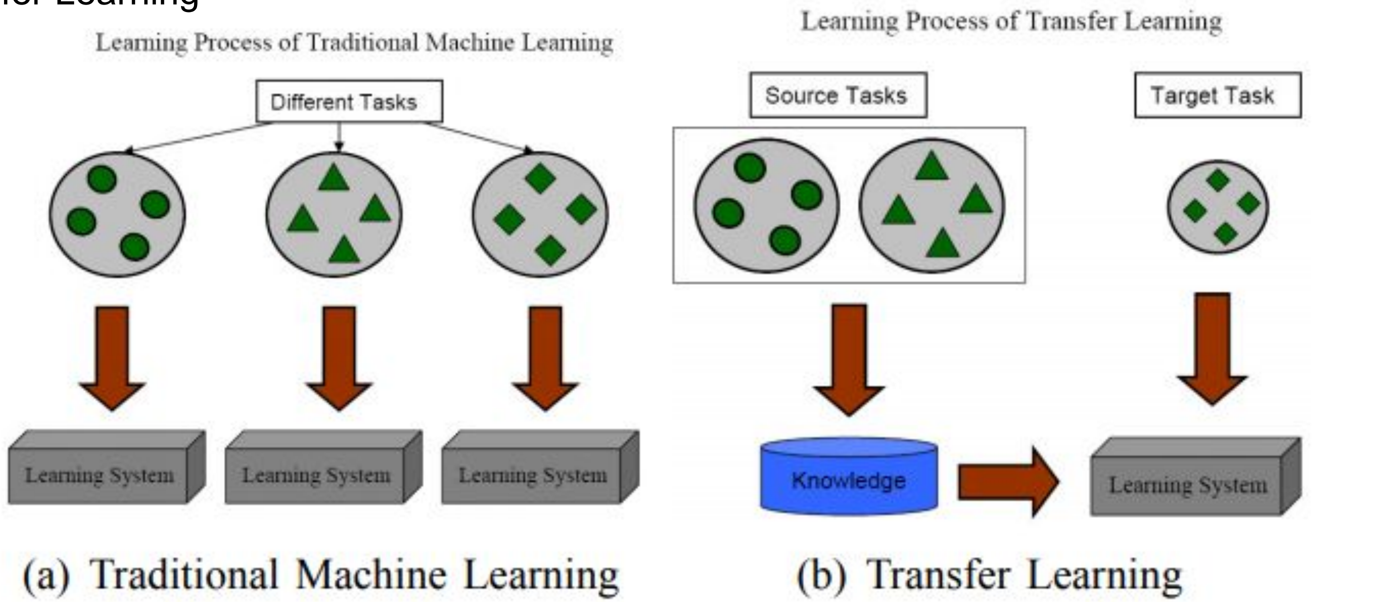


Fig. 1. Different Learning Processes between Traditional Machine Learning and Transfer Learning

Main approaches

- Feature representation
  - SCL, EasyAdapt, distributional representations
  - Finding common features/feature space
- Prior based
  - Bayesian models
  - Regularisation
- Instance based methods
  - Weight instances to make distributions similar
  - Instance selection, weighting, ruder et al

Feature representation

- Identifying common informative features
- Transforming the feature spaces of both source and target into a latent space
- Daume et al, EasyAdapt
- Structural correspondence learning (SCL)
- Distributed representations
  - Currently in focus.
- Neural SCL
- Stacked Denoising Autoencoder based methods - SDA, mSDA

Prior based methods

- Explicit P(y) term in generative models
- Regularisation in discriminative models
- Pereira
- Hierarchical
- Regularisation in NN
- Highly model dependant

Instance based methods

- Select and/or weight instances to make P(y|x) or P(x) in both domains similar
- Connected to semi-supervised learning methods
- Preventing negative transfer - important
- Ruder et al, 2018 - Showed classic tri-training beat most methods in PoS tagging under domain shift.

Recent (neural) methods

- Domain independent representations
- Language Modeling as a proxy task
- Fine tuning
- Neural adaptations of previous methods
- Concept of pivot and non-pivot features has been highly influential
- Autoencoder-based methods
- Domain adversarial training

NLP Tasks

- Most work has focussed on sentiment analysis for different domains
- (Blitzer et al) Amazon reviews dataset with 4 domains -
  - 12 adaptation tasks
  - Transfer across these domains is uneven
- PoS tagging
  - SANCL 2012 workshop - WSJ, Weblogs, Emails etc
  - Tweets
- Word Sense Disambiguation - WSJ to Medical data
- Named Entity Recognition
- Parsing etc...

Related

- Multitask learning
- Transfer learning
- Semi-supervised learning
- Few-shot learning

Lot of work (and success) in computer vision tasks.

(Recent literature for NLP is very scattered - no comprehensive evaluation of all these different methods under domain shift.)

Thanks!