Importance Weighting and Unsupervised Domain Adaptation of POS taggers:
A Negative Result

Written by Barbara Plank, Anders Johannsen and Anders Søgaard for EMNLP 2014
Presented by Xia Cui for NLP@UoL
Domain Adaptation

• Adapting a model trained on one domain to a different domain
Unsupervised Domain Adaptation

- Domain Adaptation without $T_L$
In POS Tagging

• Factorize sequences into emission and transition probabilities

• *Importance Weighting* can change emission probabilities and transition probabilities by assigning weights to sentences

• Example
  1) a/A b/A
  2) a/A b/B
  3) a/A b/B
Argument in this paper

• *Importance weighting* cannot help adapting POS tagger to new domains using only unlabelled target data

1) Negative results with the most obvious weight functions across various English datasets

2) Negative results with randomly sampled weights

3) An analysis of annotated data indicating that there is little variation in emission and transition probabilities across the various domains
Prior Work on Importance Weighting

• Use a *domain classifier*: Train a classifier to discriminate between source and target instances (y ∈ {s,t})
  o Søgaard and Haulrich, 2013: n-gram text classifier
  o Plank and Moschitti, 2013: tree-kernel based classifier

• $P(t|x) ≈ \frac{P_t(x)}{P_s(x)}$

• Select a subsample of the source data by subsequently setting the weight of all selected data points to 1, and 0 otherwise
  o Jiang and Zhai, 2007: rely on a sequential model trained on labelled target data
Experiment

• Data
  • SANCL 2012 Shared Task: Yahoo answers, user reviews, emails, weblogs and newsgroups

• Model
  • Weighted Perceptron: learning rate dependent on current instance $X_n$
    
    $$w^{i+1} \leftarrow w^i + \beta_n \alpha(y_n - \text{sign}(w^i \cdot x_n))x_n$$

  • $\beta_n$: weight associated with $X_n$
  • Extended to structured perceptron (Collins, 2002)
Importance Weighting

1) Use document classifiers to obtain weights for the source instances (Søgaard and Haulrich, 2011)

2) Train a text classifier that discriminates the two domains mark whether it come from source or target and train a binary classifier (for each instance)

3) Obtain sentence’s probability for the target domain by doing 5-fold cross-validation
Results

• No significant improvements on any of datasets by different setups
Random Weighting

• Weight function based on document classifiers may simply not characterize the relevant properties of the instances and hence lead to bad re-weighting of the data

• Random Strategies (500 sample for each)
  1) Sampling random uniforms
  2) Random exponentials
  3) Random Zipfians

• No weighting lead to significant improvements again!
Analysis

• Low ambiguity of word forms in the data
  • Make the room for improvement with importance weighting smaller

• KL divergences over POS bigrams are also very low
  • Transition probabilities are also relatively constant across domains suggesting limited room for improvement for importance weighting

• Much bigger differences in OOV rates
  • Explain most of the performance drop across domains
    o Implement structured perceptron tagger with type-constrained inference
      • Only improve performance on unseen words
Conclusion

• None of weight functions lead to significant improvements
• Most errors are due to unseen words, which cannot be captured by unsupervised weight functions