Toward Annotation Efficiency in Biased Learning Settings for Natural Language Processing

Doctoral Thesis Defense

Tom Effland
Department of Computer Science, Columbia University
10/27/22
Toward Annotation Efficiency in Biased Learning Settings for Natural Language Processing
Natural Language Processing (NLP)

Turning collections of unstructured text into structured information.

Enormous potential value across diverse and niche applications.

Goal of this thesis: make building NLP models for new applications faster and cheaper.
Rare-Event Classification for Public Health *(Ch. 2)*

Food Poisoning?  
- Yes  
- No
Extracting Mentions of Entities (Ch. 3)

Find mentions of domain-specific concepts such as **People**, **Places**, **Organizations**, **Dates**

**Early life** [edit]

Adams was born in Cambridge on 11 March 1952 to Christopher Douglas Adams (1927–1985), a management consultant and computer salesman, former probation officer and lecturer on probationary group therapy techniques, and nurse Janet (1927–2016), née Donovan.[4][5] The family moved a few months after his birth to the East End of London, where his sister, Susan, was born three years later.[6] His parents divorced in 1957; Douglas, Susan and their mother moved then to an RSPCA animal shelter in Brentwood, Essex, run by his maternal grandparents.[7] Each remarried, giving Adams four half-siblings. A great-grandfather was the playwright Benjamin Franklin Wedekind.[8]
Parsing Syntax in Low-Resource Languages (Ch. 4)
Three Foundational Types of Models

Classification (Ch. 2)

Tagging (Ch. 3)

Tagging + Parsing (Ch. 4)
Toward Annotation Efficiency in Biased Learning Settings for Natural Language Processing
Challenge: The Annotation Bottleneck

- Supervised learning is the dominant paradigm, but:
  - For complex tasks, supervised expert annotation is time consuming and expensive
  - Annotation protocols that generate unbiased data are often inefficient
  - Shifting task definitions render old annotations obsolete

- Unsupervised pretraining + supervised fine-tuning is not efficient enough
Alternatives to Unbiased Manual Annotation

- **Weak supervision**
  - Abstract expert knowledge such as rules or weak distributional constraints
  - Automatically scales with corpus size

- **Imperfectly labeled (biased) data**
  - Use whatever we have at hand, often much cheaper
  - Allow experts to focus time on the important aspects of data

- **Some (but less) supervision is still typically necessary**
  - Generally we advocate for unsupervised + weakly-supervised + supervised
Toward Annotation Efficiency in Biased Learning Settings for Natural Language Processing
Three Biased Learning Settings

We develop annotation-efficient approaches to building better models in 3 biased settings.

Selection Bias (Ch. 2)
Incomplete Labels (Ch. 3)
Out-of-domain Data (Ch. 4)
This Thesis

Goal: Make new and diverse NLP applications cheaper to build by improving annotation efficiency. We approach this by developing techniques that:

1. Use other forms of expert knowledge besides supervised data labeling.
2. Embrace biased data that is much cheaper to collect.

Generally, we intervene at the loss function.
Contributions

We develop annotation-efficient approaches to building better models in 3 biased settings.

- Rare-Event Classification with Selection Bias *(Ch. 2)*
- NER with Low-Recall Partial Annotations *(Ch. 3)*
- Cross-Lingual Syntax Parsing in Low-Resource Languages *(Ch. 4)*
Annotation-Efficient Approaches to Building Better Models in Three Biased Settings

- Rare-Event Classification with Selection Bias (Ch. 2) [JAMIA ‘18]
- NER with Low-Recall Partial Annotations (Ch. 3) [TACL ‘22]
- Cross-Lingual Syntax Parsing in Low-Resource Languages (Ch. 4) [TACL ‘23]
Rare-Event Classification for Public Health

Food Poisoning?

Review by Epidemiologists

Yes $T(x) = 1$

No $T(x) = 0$

Extremely rare: $p(\text{Sick} = \text{True}) < 1\%$
Time to upgrade

- System used for 4 years, it needs an upgrade
- 13k labeled reviews, gathered through incidental feedback, we can use to improve the system
- BUT this data is **heavily biased by selection criteria** \( T(x) = 1 \)
- Annotating a large unbiased sample is impractical (recall \( p(\text{Sick} = \text{True}) < 1\% \))

*How to obtain an unbiased model using only the biased data?*
Debiasing the model with only biased data

- Label large sample of reviews filtered out by the system \( T(x) = 0 \)
- Estimate the likelihood a review is chosen by the system, \( P(T(x) = 1) \)
- Employ importance weights to retrain [Shimodiera, ‘00]
- Positive class very rare for this set, \( P(\text{Sick} = \text{True}|T(x) = 0) \ll 1\% \)
  - Assume labels are negative, increases training data size without extra expert labels
Experimental Setup

- **Train/Dev:** ~11K biased reviews from before Jan 1st, 2017
  - Biased: no additional biased-complement reviews
  - Gold: + 1K manually labeled biased-complement reviews
  - Silver: + 10K automatically labeled biased-complement reviews

- **Test:** ~2K biased reviews from after Jan 1st, 2017
  - + 1K manually labeled complement reviews
Silver Data Improves Precision at High Recall

Silver: 69% precision @ 98% recall

Prototype: 48% precision @ 98% recall

= + 21% precision @ 98% recall
Takeaways

● Novel approach for improving a deployed rare-event classifier using only biased incidental feedback from domain experts

● Considerable improvements to deployed real-world system with immediate impact

● Chapter has additional positive results showing with additional models, tasks, and detailed error analysis

“Discovering Foodborne Illness in Online Restaurant Reviews”, Thomas Effland et al. (JAMIA 2018)
Annotation-Efficient Approaches to Building Better Models in Three Biased Settings

- Rare-Event Classification with Selection Bias (Ch. 2)
- NER with Low-Recall Partial Annotations (Ch. 3)
- Cross-Lingual Syntax Parsing in Low-Resource Languages (Ch. 4)

[JAMIA ‘18] [TACL ‘22] [TACL ‘23]
The General Problem: Low Recall Annotations

Raw supervised training results in \textit{low recall} models
Proposal: A Latent View of Missing Annotations

Latent

\[
\text{known tags} = \{(U - \text{ORG}, 7)\}
\]

\[
L_{\text{Labels}} = \sum_{\text{sentence, known tags}} -\log p_\theta(\text{known tags}|\text{sentence})
\]

Caveat: minimizing \(L_{\text{Labels}}\) by itself results in **low precision** models because the O tags aren’t observed
Proposed Solution: Expected Entity Ratio Loss (EER)

\[ \text{LOSS} = L_{\text{Labels}} + \lambda L_{\text{EER}} \]

- \( L_{\text{Labels}} \): Encourage high entity recall
- \( L_{\text{EER}} \): Encourage the rate of entity tags under model to be in a certain range
- \( \lambda \): Balance loss scales

Theorem 2 in chapter \( \Rightarrow \) Minimizer of loss is recovers true parameters with infinite data (under reasonable conditions)
Expected Entity Ratio Loss (EER) Details

Inputs:

\[ \rho : \text{Expected proportion of entity tags} \]

\[ \gamma : \text{Margin of uncertainty} \]

\[ p_\theta(\text{some tag} \neq 0) : \text{Marginal probability of predicting an entity tag} \]

\[ L_{\text{EER}} = \max\{0, |\rho - p_\theta(\text{some tag} \neq 0)| - \gamma\} \]

Penalize model if probability of predicting an entity tag is outside \( \rho \pm \gamma \)
Experimental Setup: Datasets and Preprocessing

- 7 datasets in 6 languages from CoNLL 2003 and Ontonotes 5
- Each downsampld to 1K entity annotations using sampler ([0.8%,8%] recall). Observed annotations clustered at top of document.
- Three ways of preprocessing data to reduce false negative annotations
  - **all**: use full dataset as is
  - **short**: drop all unlabeled documents
  - **shortest**: drop all unlabeled sentences
- Each reduces number of false negatives but also reduces size of training set
- Ideally, approaches work well across all of these no matter how many false negatives
Experimental Setup: Approaches

- **Raw:**
  - Normal supervised training on data with missing annotations
  - Results in low-recall models

- **SNS:** [Li ‘21]
  - Span-based model, each possible span an independent classification
  - Sample unobserved spans, assuming they are negative
  - Does not scale to longer texts and uses ad-hoc decoding
  - Weaker theoretical guarantees

- **EER:** our approach

All methods use same underlying BERT contextual representations
Results: 7 Datasets, 1K Annotations Each

<table>
<thead>
<tr>
<th></th>
<th>Raw</th>
<th>SNS [Li '21]</th>
<th>EER</th>
</tr>
</thead>
<tbody>
<tr>
<td>best of (all, short, shortest)</td>
<td>65</td>
<td>69.6</td>
<td>71.7*</td>
</tr>
<tr>
<td>average of (all, short, shortest)</td>
<td>31.8</td>
<td>54.5</td>
<td>70.2*</td>
</tr>
</tbody>
</table>

* statistically significant over others
Takeaways

● Simple, effective way to relax the “high-recall” requirement of labeled NER data

● Theoretically sound and state-of-the-art performance

● Much more robust to varying numbers of false negatives

● Chapter has additional positive results with other low-recall annotation settings, and shows sparse annotation plus EER is as good as exhaustive annotation for modest data (<10K entities), along with other analysis

“Partially Supervised Named Entity Recognition via the Expected Entity Ratio Loss”, Thomas Effland and Michael Collins. (TACL 2022)
Annotation-Efficient Approaches to Building Better Models in Three Biased Settings

- Rare-Event Classification with Selection Bias (*Ch. 2*)
- NER with Low-Recall Partial Annotations (*Ch. 3*)
- Cross-Lingual Syntax Parsing in Low-Resource Languages (*Ch. 4*)

[JAMIA ‘18] [TACL ‘22] [TACL ‘23]
Improving Low-Resource Cross-Lingual Parsing

- Model predicts **part-of-speech (POS) tags** and **labeled dependency tree**
- **Labeled data** exists for ~100 **languages** (Universal Dependencies)
- Want to **improve parsing for “low-resource” languages** with few labeled data
Previous State-of-the-Art Approach

- State of the art: highly multilingual language model pretraining and fine-tuning on as many languages as possible [Kondratyuk ‘19]
- Transferring to new languages is zero-shot or few-shot fine-tuning
- Models make erratic and syntactically implausible predictions on “underrepresented” languages
Proposal: Intuition

- Many egregious errors are simple in nature, such as:
  - Wildly over predicting less common tags, such as punctuation
  - Predicting dependency combinations that never occur in the training data of any language

- Generally, models fail to match many low-order statistics of the target syntactic structures

- Models don’t effectively learn these from small amounts of fine-tuning

- **Hypothesis:** enforcing model regularity w.r.t. estimates of the target low-order statistics **(using them as weak supervision)** is complementary to transfer learning
Proposed Statistics: Use Syntactic Typology

- We describe 7 families of marginal statistics based on syntactic typology

- Universally Impossible Arcs: rule out 93% of combinations of (head tag, dep label, child tag), such as NOUN ... PUNCT
Proposal: Expected Statistic Regularization (ESR)

- Generalization of NER EER loss to any statistic of the model and data

\[
\text{Loss} = L_{\text{Labels}}(\theta; D_L) + \lambda L_{\text{ESR}}(\theta; D_U)
\]

$L_{\text{Labels}}$: Encourage supervised accuracy on small labeled dataset

$L_{\text{ESR}}$: Encourage the statistics that quantify model behavior to be close to target values

$\lambda$: Balance loss scales
The ESR Term

- Define differentiable vectorized “statistic” function \( f \) that quantifies simple aspects of model behavior on samples of data.
- Given target values for those statistics and margins of uncertainty, penalize the model for statistics that deviate from the targets.

\[
L_{ESR} = \mathbb{E}_{D^k_U} [\ell(t, \sigma, f(D^k_U, p_{\theta}))]
\]

- Penalize deviation of current stats from targets, modulated by margins
- Describe the model with many quantities using the minibatch

Target statistic values and margins
A minibatch of unlabeled data
The model
ESR for Cross-Lingual Syntax Summary

1. Pretrain parser on as many languages/treebanks as possible [Kondratyuk ‘19]
2. Design statistics that quantify important aspects of model behavior
3. Estimate statistic targets and margins on small samples of labeled data in the target language using bootstrap sampling
4. Fine-tune the model on the target treebank with supervised loss and additional ESR loss for unlabeled batches.
Low-Resource Transfer Benchmark: Setup

Data: 44 languages that are not in the training set, downsamped to 50 labeled sentences. Avg over 3 dataset samples per language.

Approaches:

- **Baseline (FT):** supervised fine-tuning of multilingual parsing model
- **Ours (ESR-CLD):** supervised fine-tuning of multilingual parser plus ESR using the best target statistic from preliminary experiments
  - **Child-Label-Direction (CLD):** probability of an edge label $X$, with child tag $T$, headed in direction $D$. $T X$ vs. $X T$
Low-Resource Transfer Benchmark: Results

Improvement in Labeled Attachment Score of ESR (ours) over Fine-tuning

- 90% are modern European languages
- 20% are modern European languages

AVG +2.3

Treebank

Languages:
- Estonian
- Belarusian
- Norwegian
- Dutch
- Ukrainian
- Urdu-UDTB
- Croatian
- Slovenian
- Persian
- Czech-PDT
- Bulgarian
- Latin-ITB
- Lithuanian
- German
- Polish-PDB
- Romanian
- Slovak
- Serbian
- Catalan
- Spanish
- Armenian
- Danish
- Portuguese
- Old_Church_Old_French
- Uyghur
- Indonesian
- Tamil-ITB
- Greek-GDT
- Naja-NSC
- Marathi
- Galician
- Hungarian
- Afrikaans
- Turkish_Ge
- Vietnamese
- Telugu
- Western_Ar
- Welsh
- Gothic
- Faroese
- Scottish_G
- Maltese
- Wolof-WTB
Takeaways

- Proposed **novel and general “Expected Statistic Regularization”** for shaping models on unlabeled datasets with high-level summary information
- Proposed **method for estimating regularization targets and margins**
- Contributed **significant application** to improving state-of-the-art low-resource cross-lingual parsing
- Show **ESR is complementary to transfer learning and fine-tuning** in low-resource settings
- Chapter has additional positive results with more statistics, transfer settings, learning curves, baselines, and ablations

Contributions

We develop annotation-efficient approaches to building better models in 3 biased settings.

- Rare-Event Classification with Selection Bias \((Ch. 2)\)
- NER with Low-Recall Partial Annotations \((Ch. 3)\)
- Cross-Lingual Syntax Parsing in Low-Resource Languages \((Ch. 4)\)

[JAMIA ‘18] [TACL ‘22] [TACL ‘23]
Thanks!

Contact:  teffland@cs.columbia.edu

Code, results, and links to papers:

- (Ch. 2) https://github.com/teffland/FoodborneNYC
- (Ch. 3) https://github.com/teffland/ner-expected-entity-ratio
- (Ch. 4) https://github.com/teffland/expected-statistic-regularization