Understanding the Tradeoff between Cost and Quality of Expert Annotations for Keyphrase Extraction

Hung Chau
University of Pittsburgh, Pittsburgh, PA, USA

Saeid Balaneeshin, Kai Liu, Ondrej Linda
Zillow Group, Seattle, WA, USA

Introduction and Motivation
- Domain-specific keyphrase extraction models are widely used in many real-world applications.
- Keyphrase extraction can be performed via pre-trained or unsupervised models. However, such models might suffer from a lack of domain-specific knowledge.
- It is still unclear whether the training set extracted from annotations with multiple annotation/review steps would increase the performance of supervised keyphrase extraction models significantly.
- Main contributions:
  ○ Experimentally analyzes the tradeoff between the cost of annotation strategies and their impacts on classification and sequence labeling keyphrase extraction models' performance.
  ○ Results demonstrate that different strategies should be selected depending on specific metrics (e.g., precision).

Annotation Procedure
Three experts receive a training and pass a test on the understanding of the task and the BRAT interface. They review all listing descriptions for testing and 50 for training data and record the labeling time.

Guidelines
- The annotation guidelines are created by experts through multiple annotation-discussion iterations.
- The experts may combine guidelines to select or reject a keyphrase.
- Coding Procedure for Training Set
  ○ Performing review after original annotations to improve the quality of the keyphrase labels.
  ○ Experts review their own annotations and make adjustments in a self-review process; while, in a peer-review process, they correct other annotations and make final decisions.

Statistics
- We focus on the real estate domain in English and create a dataset with 50 and 20 listing descriptions for training and evaluation. The average length of the sampled listing descriptions was 125 words, with some having as few as 50 and as many as 500.
- n-grams in Selected Keyphrases: Figure 3 shows that the majority of selected keyphrases are bigrams (around 40%), trigrams account for about 30% and unigrams are the least common.
- As seen from Figure 5, the number of added/removed keyphrases in the peer-review process is around 5 times more than that in self-review process and that the number of added keyphrases is higher than removed in both self- and peer-review processes.
- Results: the majority of selected keyphrases are bigrams (around 40%), trigrams account for about 30% and unigrams are the least common.

Model Performance Comparison
- We investigate three review types: (1) orig: original (2) self: self-review, or (3) peer: peer-review; and three voting rules: (1) unan: annotated by at least one expert, (2) majority: annotated by two or more experts, or (3) unanimity: annotated by all three experts. In total, we generate 18 different training data sets.
- Table 1: Example set of guidelines created by the experts.

Annotator Data Analysis
- Statistics: focus on the real estate domain in English and create a dataset with 50 and 20 listing descriptions for training and evaluation. The average length of the sampled listing descriptions was 125 words, with some having as few as 50 and as many as 500.
- n-grams in Selected Keyphrases: Figure 3 shows that the majority of selected keyphrases are bigrams (around 40%), trigrams account for about 30% and unigrams are the least common.
- As seen from Figure 5, the number of added/removed keyphrases in the peer-review process is around 5 times more than that in self-review process and that the number of added keyphrases is higher than removed in both self- and peer-review processes.
- Results: the majority of selected keyphrases are bigrams (around 40%), trigrams account for about 30% and unigrams are the least common.

Conclusion
- Different annotation strategies can be considered depending on a specific metric.
- Exploiting domain-specific knowledge and expert annotations is important for keyphrase extraction problems.
- This work is limited to one small dataset in one specific domain.
- Our priority is to investigate whether the approach is valid for other NLP tasks such as relation extraction.
- Investigating how useful the generated codebook for weak supervision modeling.