Transfer Learning on Stack Exchange Tags

Fanming Dong, Shifan Mao, Weiqiang Zhu *(advised by Danqi Chen)*

Introduction

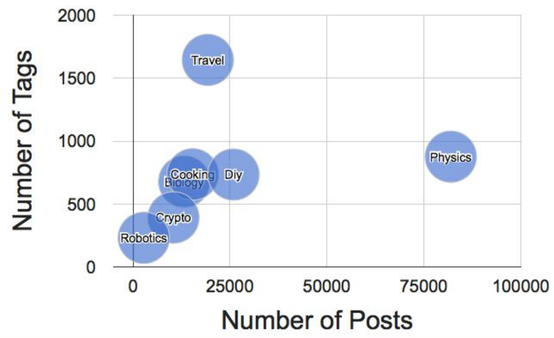
What can you learn about physics from studying biology?

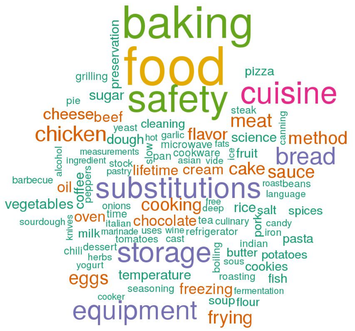
Stack Exchange is an online forum where people crowd source answers to "hot questions". For easy navigation of questions and answers, the challenge asks people to label questions with correct tags. But what if it is a question we have never seen before?

We use NLP methods to find a post's tags according to its title and content. More importantly, we explore a model's transfer learning ability to predict tags from unseen domains.

Data Exploration

Our dataset comes from Kaggle. Data includes post title, content, and tags from Stack Exchange on a variety of six topics - biology, cooking, cryptography, diy, robotics, and travel. The tags are words or phrases that describe the topic of questions. Test set contains about 81,000 physics posts with no tags.





Word frequency in category cooking

Methodology

We base transfer learning between categories on Glove word2vector model.

1. Unsupervised Learning

- topic discovery with latent-Dirichlet allocation
- tf-idf weighted embedding similarity

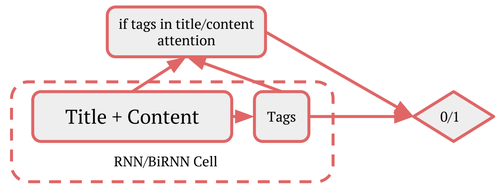
$$\vec{u}_{\text{post}} = \sum_{\text{word}} \text{tf-idf}(\text{word}, \text{corpus})\vec{u}_{\text{word}}$$

$$P(\text{tag}|\text{post}) \propto \cos(\vec{u}_{\text{tag}}, \vec{u}_{\text{post}})$$

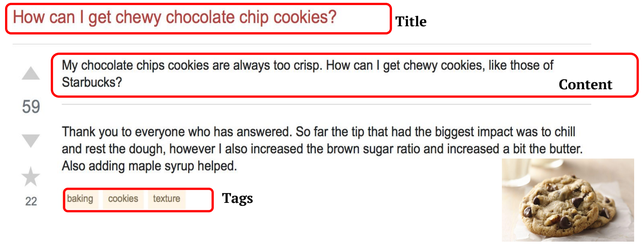
2. Logistic Regression

3. Recurrent Neural Network

We choose the GRU RNN and test both basic RNN and BiRNN cells. We implemented both a standard attention mechanism and a direct "tag-in-title" feature to capture the relationship between words in title/content and tags.



Transfer Learning

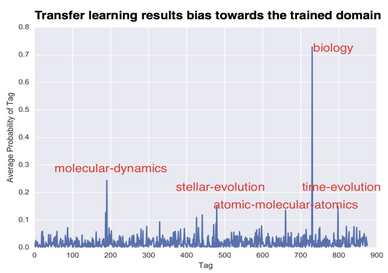


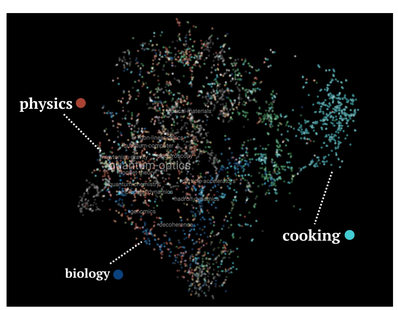
	RNN	tf-idf based search	latent-Dirichlet allocation
baking **cookies** texture	**cookies** chocolate	chips software tortilla-chips apple stock	chicken sushi tofu meat
oven cooking-time bacon	**oven bacon**	bacon lamb	frying chicken roasting meat
sauce pasta tomatoes italian-cuisine	acidity **sauce** substitutions	acidity texture alkalinity	learning raw basics mixing
substitutions herbs parsley	**substitutions** spices **herbs**	**herbs spices**	**herbs**
eggs basics poaching	**eggs** substitutions baking	sardines eggs scrambled-eggs mussels	water filling fresh raw
ice-cream	cream **ice-cream**	texture caramel meatloaf	**cream**
grilling salmon **cedar-plank**	salmon fish seafood	cedar-plank kettle	learning raw basics mixing
storage-method ripe avocados	avocados fruit	**avocados**	gas oil cost

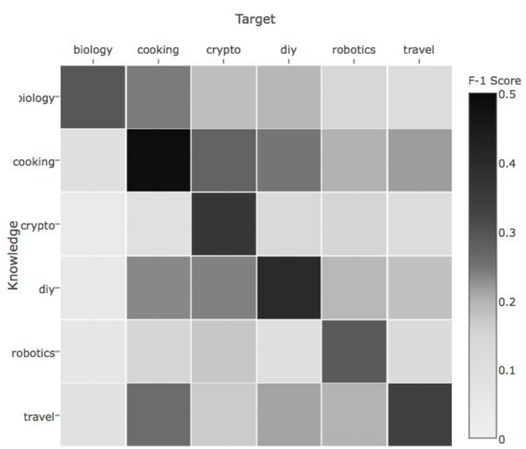
Model Evaluation

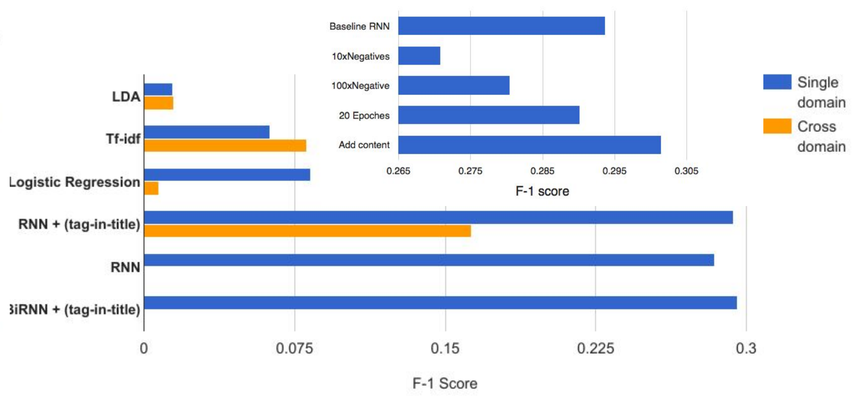
Precision = [cookies] / [cookies, chocolate] = 1/2
Recall = [cookies] / [baking, cookies, texture] = 2/3

$$\text{F-1 score} = \frac{2p \cdot r}{p + r} \approx 0.57$$









Discussions

Conclusion:

- Learning on a single domain can achieve reasonable F1 scores (0.48 for "cooking"). But F1 scores on transfer domains are much lower (0.25 at best).
- The effect of transfer learning based on fixed Glove model is limited. The prediction results are highly biased towards the domain of the training set.

Next Steps:

TopicRNN: Topic models focus on the global structure, while RNN models capture local structure. Next we can combine two models.