Predicting Hierarchical Relationship in Job Title Taxonomy

Shuang Jin sjin1@stanford.edu https://youtu.be/yb6EjHsbnJ8

Problem Statement

Goal

• predict the relationship between job titles in the taxonomy.

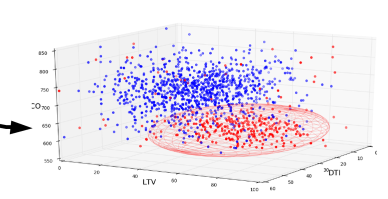
Motivation

• build a well-structured taxonomy to organize job market knowledge.

Problem Definition

• given a job title entity pair (x_{source}, x_{target}) predict their relationship y .

x_source: **Machine Learning Engineer**
x_target: **Computer Software Engineer**



x_target is a **Broader Term** of x_source.

Data

Source

• an established title taxonomy curated by a dedicated taxonomy team.

Train	290693	93%
Validation	18723	6%
Test	3106	1%

labels & Size

Label	Meaning	Data Size
BT	Broader Term	100K
NT	Narrower Term	100K
PT	Preferred Term	20K
NPT	Non-Preferred Term	20K
UKN	Unknown	60K

Features

	Raw	Tokenized
Source	supply chain specialist	0 0 0 0 0 65 73 10
Target	supplier quality specialist	0 0 0 0 0 1097 79 10
Label	NT	0 1 0 0 0

Models & Results

$$J = -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{C}\lambda_j \cdot y_j^i \cdot \log \hat{y}_j^i$$

Model	Precision	Recall	F1
Base	77.52%	72.11%	73.66%
Simple LSTM 64d	94.63%	91.90%	93.14%
Simple LSTM 128d	93.96%	93.04%	93.44%
Glove	77.33%	71.75%	73.27%
Glove LSTM 64d	95.84%	94.70%	95.21%
Glove LSTM 128d	95.31%	94.56%	94.90%

Challenges

Embedding trained from this project outperformed Glove embedding?

- Large percentage of spelling errors that cannot be recognized by Glove
- High repetitiveness of words in training data making embedding training less difficult
- **Solution**: Removing spelling mistakes

Large fluctuation on validation set performance after a few epochs?

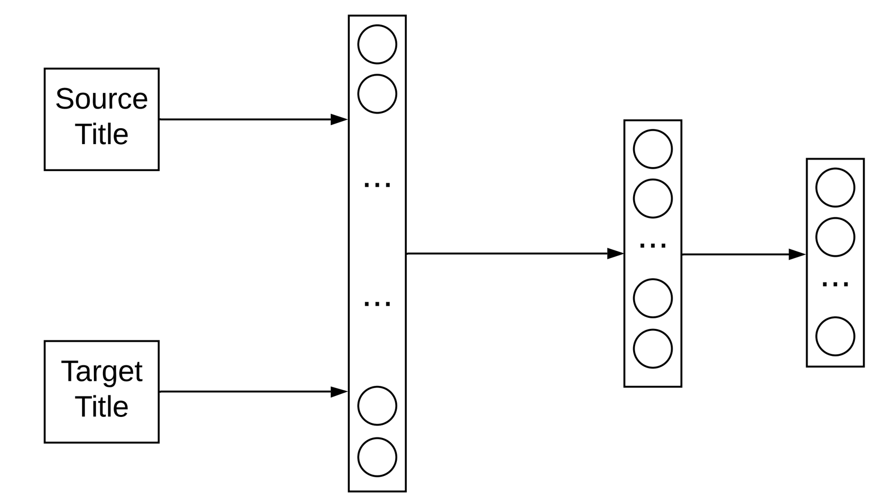
- Learning rate set too large
- **Solution**: Calibrating on the learning rate

One label performance significantly worse than others?

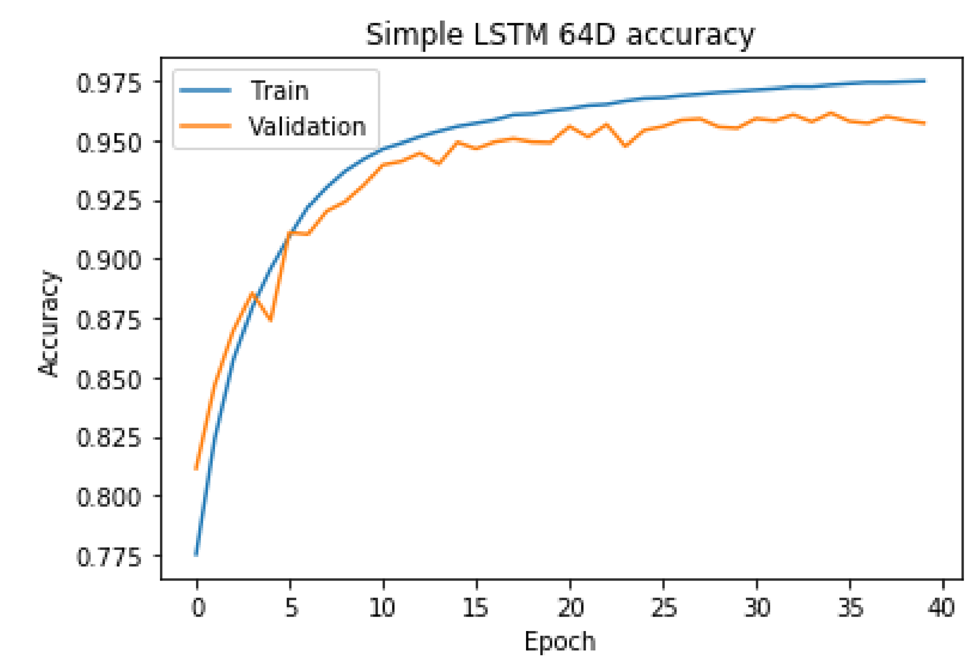
- Data imbalance
- **Solution**: adding class weights to loss

Selected Model

Glove LSTM 64d



	Input	Embedding w/ Glove	LSTM	Softmax
Dim	(16)	(16, 100)	(64)	(5)
Param #	0	351700	42240	325



	BT	NT	NPT	PT	UKN
BT	0.984	0.001	0.005	0.005	0.004
NT	0	0.988	0.001	0.005	0.006
NPT	0.005	0.015	0.913	0.068	0
PT	0	0.010	0.029	0.951	0.010
UKN	0.012	0.012	0.005	0.015	0.956

Normalized confusion matrix

Prediction Examples

Source	Target	Prediction	Score
vice president of construction	consultant sap security	URT	.9999
director training development	head - hr & administration	BT	.9997
information technology manager	senior manager human resources information system	NT	.9577

Discussion

Observation

Label Difficulty for *Machine* vs. *Human*:

- Machine: NPT > PT > UKN > BT > NT
- Human: PT ~ NPT > BT ~ NT > UKN

Why is UKN easy for human but not so easy for machine?

- Humans use knowledge such as related industries, skills, etc. to make the judgement
- **Future Work**: Mine title-related data as additional features

What do humans do to get better on PT vs. NPT?

- Humans use popularity data to compare which title is used more often
- **Future Work**: adding counts as new features

What do humans do to get better on PT vs. NPT?

- Humans use popularity data to compare which title is used more often
- **Future Work**: adding counts as new features

What could be the pitfalls for the model in production?

- Labeled data has sector bias
- **Future Work**: balancing sector data

References

[1] Mamadou Diaby and Emmanuel Viennet. Taxonomy-based job recommender systems on facebook and linkedin profiles. 2014 IEEE Eighth International Conference on Research Challenges in Information Science (RCIS), pages 1–6, 2014.

[2] Alex Sherstinsky. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. CoRR, abs/1808.03314, 2018.

[3] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543, 2014.