Beyond Word2Vec: Using embeddings to chart out the ebb and flow of tech skills

Maryam Jahanshahi Ph.D.
Research Scientist
TapRecruit.co

Senior Finance Analyst
TapRecruit - Los Angeles

TapRecruit is looking for a smart, detail-oriented person to serve as a senior financial analyst. This person will be responsible for supporting the company’s FP&A requirements. Responsibilities will include working on TapRecruit Entertainment Group’s FP&A model, supporting analysis for long-term planning options, tracking key business operational metrics and producing monthly financial/operational reports. In addition to FP&A needs, this role will require strong organizational skills to help manage the new processes. The candidate will interface with senior managers across the department and evaluate/implement ideas to improve company projects for top-line growth. This position reports to a Senior Manager of Finance and will routinely interface with TapRecruit’s top management.

This is an ideal position for an individual who has gained strong analytical skills at a large investment bank or accounting firm and now seeks to apply those skills to a fast-growing entrepreneurial company. Strong quantitative and excel financial modeling skills are a must. The ideal candidate must be comfortable in a dynamic start-up environment, will bring energy and passion to everything he/she does, and will not be afraid to roll up his/her sleeves to tackle challenging analytical assignments.

This job is full-time, based in Los Angeles. We offer competitive compensation and stock option program.
Skills and qualifications matter in job descriptions.

<table>
<thead>
<tr>
<th>Finance Manager</th>
<th>Finance Manager</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Kraft Foods</strong></td>
<td><strong>Roche</strong></td>
</tr>
<tr>
<td>Junior (3 Years)</td>
<td>Senior (6-8 Years)</td>
</tr>
<tr>
<td>No Managerial Experience</td>
<td>Division Level Controller</td>
</tr>
<tr>
<td></td>
<td>Strategic Finance Role</td>
</tr>
<tr>
<td></td>
<td>MBA / CPA</td>
</tr>
</tbody>
</table>

- Same title, Different job
- Required Experience
- Required Responsibility
- Preferred Skill
- Preferred Education

- Same Title
- Required Experience
- Required Responsibility
- Preferred Skill
- Required Education
Research at TapRecruit
Helping companies make fairer and more efficient recruiting decisions

### NLP and Data Science:
- What are distinguishing characteristics of successful career documents?
- What skills are increasingly important for different industries?

### Decision Science:
- How do candidates make decisions about which jobs to apply to?
- How do hiring teams make decisions about candidate qualifications?
How have tech skills changed over time?
Strategies to identify changes among corpora

Manual Feature Extraction
Require selection of key attributes, therefore difficult to discover new attributes

Dynamic Topic Models
Require experimentation with topic number

Adapted from Blei and Lafferty, ICML 2006.
Embeddings use context to extract meaning

Window sizes capture semantic similarity vs semantic relatedness

Statistical modeling through software (e.g. SPSS) or programming language (e.g. **Python**)

Experience in **Python**, Java or other object-oriented programming languages

Proficiency programming in **Python**, Java or C++.
A simplified representation of word vectors

Dimension reduction is key to all types of embeddings models
Embeddings capture entity relationships
Dimensionality enables comparison between word pairs along many axes

Hierarchies

Comparatives and Superlatives

Man :: King as Woman :: ?

Adapted from Stanford NLP GLoVE Project
Embeddings reflect cultural bias in corpora
High dimensionality enables some bias reduction

Adapted from Bolukbasi et al., arXiv: 1607.06520.
Pretrained embeddings facilitate fast prototyping

Embeddings training should match corpus that is being tested on

<table>
<thead>
<tr>
<th>Corpus Generation</th>
<th>Corpus Tokens</th>
<th>Twitter</th>
<th>Common Crawl</th>
<th>GoogleNews</th>
<th>Wikipedia</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>27 B</td>
<td>42-840 B</td>
<td>100 B</td>
<td>6 B</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Corpus Processing</th>
<th>Vocabulary Size</th>
<th>1.2 M</th>
<th>1.9-2.2 M</th>
<th>3 M</th>
<th>400 k</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Language Model Generation</th>
<th>Algorithm</th>
<th>GLoVE</th>
<th>GLoVE</th>
<th>word2vec</th>
<th>GLoVE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vector Length</td>
<td></td>
<td>25 - 200 d</td>
<td>300 d</td>
<td>300 d</td>
<td>50 - 300 d</td>
</tr>
</tbody>
</table>

Final Application
# Problems with pretrained embedding models

<table>
<thead>
<tr>
<th>Casing</th>
<th>Abbreviations vs Words</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>e.g. IT vs it</td>
</tr>
<tr>
<td><strong>Out of Vocabulary Words</strong></td>
<td>Domain Specific Words &amp; Acronyms</td>
</tr>
<tr>
<td><strong>Polysemy</strong></td>
<td>Words with multiple meanings</td>
</tr>
<tr>
<td></td>
<td>e.g. drive (a car) vs drive (results)</td>
</tr>
<tr>
<td></td>
<td>e.g. Chef (the job) vs Chef (the language)</td>
</tr>
<tr>
<td><strong>Multi-word Expressions</strong></td>
<td>Phrases that have new meanings</td>
</tr>
<tr>
<td></td>
<td>e.g. Front-end vs front + end</td>
</tr>
</tbody>
</table>
Custom language models tools
Modularized for different data and modeling requirements

- spaCy
- OpenNLP
- CoreNLP
- SyntaxNet
- gensim
- PyTorch
- TensorFlow
- Amazon SageMaker

**Corpus Processing**
- Tokenization, POS tagging, Sentence Segmentation, Dependency Parsing

**Language Modeling**
- Different word embedding models (GLoVE, word2vec, fastText)
Career language embedding model

Identified equal opportunity and perks language
Career language embedding model
Identified 'soft' skills and language around experience
I’ve got 300 dimensions...
but time ain’t one
Two approaches to connect embeddings

Static embeddings
stitched together

- Data hungry
- Requires alignment

2015
2016
2017
2018

Dynamic embeddings
trained together

- Data efficient
- Does not require alignment

2015
2016
2017
2018


Balmer and Mandt, arXiv: 1702:08359
Yao, Sun, Ding, Rao and Xiong, arXiv: 1703:00607
Rudolph and Blei, arXiv: 1703:08052
Dynamic Bernoulli embeddings
Outputs facilitate quick analysis of trends

Absolute drift
Identifies top words whose usage changes over time course

<table>
<thead>
<tr>
<th>words with largest drift (Senate)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IRAQ</td>
</tr>
<tr>
<td>tax cuts</td>
</tr>
<tr>
<td>health care</td>
</tr>
<tr>
<td>energy</td>
</tr>
<tr>
<td>medicare</td>
</tr>
<tr>
<td>DISCIPLINE</td>
</tr>
<tr>
<td>text</td>
</tr>
<tr>
<td>VALUES</td>
</tr>
</tbody>
</table>

Embedding neighborhoods
Extract semantic changes by nearest neighbors of drifting words

<table>
<thead>
<tr>
<th>1858</th>
<th>1940</th>
<th>2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>unemployement</td>
<td>depression</td>
<td>unemployement</td>
</tr>
<tr>
<td>unemployed</td>
<td>alleviating</td>
<td>unemployement</td>
</tr>
<tr>
<td>acute</td>
<td>destitution</td>
<td>jobless</td>
</tr>
<tr>
<td>deplorable</td>
<td>acute</td>
<td>rate</td>
</tr>
<tr>
<td>alleviating</td>
<td>reemployment</td>
<td>depression</td>
</tr>
<tr>
<td>destitution</td>
<td>deplorable</td>
<td>forecasts</td>
</tr>
<tr>
<td>urban</td>
<td>employment</td>
<td>crate</td>
</tr>
<tr>
<td>employment</td>
<td>distress</td>
<td>upward</td>
</tr>
<tr>
<td>distressing</td>
<td></td>
<td>lag</td>
</tr>
<tr>
<td>predict</td>
<td></td>
<td>economists</td>
</tr>
</tbody>
</table>
Experiments with dynamic embeddings

<table>
<thead>
<tr>
<th>Small Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Job Types</td>
</tr>
<tr>
<td>Time Slices</td>
</tr>
<tr>
<td>Number of Documents</td>
</tr>
<tr>
<td>Vocabulary Size</td>
</tr>
<tr>
<td>Data Preprocessing</td>
</tr>
<tr>
<td>Embedding Dimensions</td>
</tr>
</tbody>
</table>
Small corpus identified MBAs and PhDs
Reduced requirement for advanced degrees in many jobs

**Demand for MBAs is Falling**
- MBAs in All Jobs: -35%
- MBAs in DS Jobs: -15%
- MBAs in Tech Jobs: +30%

**Demand for PhDs is Falling**
- PhDs in All Jobs: -35%
- PhDs in DS Jobs: -20%
- PhDs in ML Jobs: -30%
Small corpus identified skill demands

Data Viz is up and Hadoop (but not Spark) is down

Demand for Data Visualization tools is up

Tableau: +20%

PowerBI: +100%

Demand for Hadoop is down in DS and ML roles

Hadoop: -30%

Spark: Steady

Blue boxes indicate phrases identified from top drifting words analysis. Grey and pink boxes indicate 'control' skills.
Battle of the Languages

Difference between supply vs demand of scripting languages

Demand for Perl is down

Perl -40%  Python Steady

Blue boxes indicate phrases identified from top drifting words analysis. Grey and pink boxes indicate 'control' skills.
Battle of the Languages

Difference between supply vs demand of scripting languages

**Demand for Python up in Tech roles**

- Python in Tech Jobs: +30%
- Python in DS Jobs: Steady

**Demand for Java is up**

- Java in Tech Jobs: +30%
- Java in DS Jobs: +35%

Blue boxes indicate phrases identified from top drifting words analysis. Grey and pink boxes indicate 'control' skills.
Experiments with dynamic embeddings

<table>
<thead>
<tr>
<th></th>
<th>Small Corpus</th>
<th>Large Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Job Types</strong></td>
<td>All</td>
<td>All</td>
</tr>
<tr>
<td><strong>Number of Documents</strong></td>
<td>50 k</td>
<td>500 k</td>
</tr>
<tr>
<td><strong>Vocabulary Size</strong></td>
<td>10 k</td>
<td>10 k</td>
</tr>
<tr>
<td><strong>Data Preprocessing</strong></td>
<td>Basic</td>
<td>Basic</td>
</tr>
<tr>
<td><strong>Embedding Dimensions</strong></td>
<td>100 d</td>
<td>100 d</td>
</tr>
</tbody>
</table>
SQL was a top drifting word

Large corpus identified role-type dependent shifts in requirements

SQL requirement increases in specific functions:

- FP&A Roles: +70%
- Sales Roles: Steady
- Marketing Roles: Steady
- FinTech Roles: Steady
- BizDev Roles: +50%
- HR Roles: +25%

Data Science & Tech Jobs:

<table>
<thead>
<tr>
<th>Year</th>
<th>2016</th>
<th>2017</th>
<th>2018</th>
<th>2019</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>37.5%</td>
<td>37.5%</td>
<td>50%</td>
<td>50%</td>
</tr>
</tbody>
</table>
Beyond word2vec
- Flavors of static word embeddings: The Corpus Issue
- Considerations for developing custom embedding models
- Dynamic Bernoulli embeddings are robust with small datasets

How have tech and data science skills changed?
- Demand for MBAs and PhDs is falling
- Core Skills: DataViz & Scripting Languages
- Commodification of distributed systems impacts demand for Hadoop
- Demand for SQL in a variety of core business functions
Thank you AI Conference!

Maryam Jahanshahi Ph.D.
Research Scientist
@mjahanshahi
maryam-j