Proceedings of the Twenty-Seventh Conference on Innovative Applications of Artificial Intelligence

SKILL: A System for Skill Identification and Normalization

Meng Zhao, Faizan Javed, Ferosh Jacob, Matt McNair
5550-A Peachtree Parkway, Norcross, GA 30092, USA
{Meng.Zhao, Faizan.Javed, Ferosh.Jacob, Matt.McNair}@CareerBuilder.com

Abstract

Named Entity Recognition (NER) and Named Entity Normalization (NEN) refer to the recognition and normalization of raw texts to known entities. From the perspective of recruitment innovation, professional skill characterization and normalization render human capital data more meaningful both commercially and socially. Accurate and detailed normalization of skills is the key for the predictive analysis of labor market dynamics. Such analytics help bridge the skills gap between employers and candidate workers by matching the right talent for the right job and identifying in-demand skills for workforce training programs. This can also work towards the social goal of providing more job opportunities to the community. In this paper we propose an automated approach for skill entity recognition and optimal normalization. The proposed system has two components: 1) Skills taxonomy generation, which employs vocational skill related sections of resumes and Wikipedia categories to define and develop a taxonomy of professional skills; 2) Skills tagging, which leverages properties of semantic word vectors to recognize and normalize relevant skills in input text. By sampling based end-user evaluation, the current system attains 91% accuracy on the taxonomy generation and 82% accuracy on the skills tagging tasks. The beta version of the system is currently applied in various big data and business intelligence applications for workforce analytics and career track projections at CareerBuilder.

1 Introduction

In 2012 the *Wall Street Journal* published an article (Peters and Wessel 2012) that stated that despite plenty of job openings in the city, over 14,000 people in Fort Wayne, Indiana were still looking for work. For STEM (science, technology, engineering, and math) fields, the US Bureau of Labor Statistics estimates that over 1.2 million jobs will be unfulfilled by 2018[1]. These examples demonstrate a growing supply and demand gap in the labor market. As a major player and contributor in the global recruitment industry, it is both entrepreneurially beneficial and socially responsible for CareerBuilder to seek solutions for such socio-economic challenges. Analyzing and working towards closing the skill gap will be a significant starting point because skills play a key role in the competitiveness of individuals, enterprises and even nations[2]. This drives a critical business and social need for a high quality skills taxonomy generation and tagging system.

There are two main challenges in identifying and realizing a skill from resumes or job descriptions: 1) the same skill represented in different ways (e.g., C# and C sharp); 2) the same term representing different skills in different contexts (e.g., Java in Java coffee and Java programming language). This naturally escalates the need for a more comprehensive and accurate system of Named Entity Recognition (NER) and Named Entity Normalization (NEN) to identify and map similar references to entities of skills. NER refers to the recognition of phrases of interest from text (commonly referred to as surface forms) and NEN refers to appropriate associations of these surface forms with a formal entity. There have been many applications of NER and NEN across various domains using rule-based as well as inference-based approaches in published literatures (see Section 4 for a complete review).

In this paper we present SKILL, an automated system for taxonomy generation and skills recognition and normalization. We assume that recruitment related documents are composed and edited in a professional manner. This premise rationalizes the foundation of the SKILL system: extracting surface forms from resumes, job postings and beyond, then properly normalizing them to eligible skill entities. Moreover, due to specialized and concentrated input data sources, we are able to attain and maintain a highly accurate and sizable taxonomy in an automated fashion.

The remainder of the paper is organized as follows. We describe the design and implementation of skills taxonomy generation and tagging in Section 2. A thorough evaluation of the SKILL system is presented in Section 3. Related

[1] http://www.bls.gov/opub/mlr/2009/11/art5full.pdf

[2] http://www.hbs.edu/competitiveness/research/

work is discussed in Section 4 and we conclude in Section 5.

2 Methodology

CareerBuilder has over 100 million resumes in English as of March 2014. By descriptive analysis, around 50% of the resumes contain special sections for vocational skills, headlined by a handful of typical headers, resulting in approximately 46 million unique phrases that describe or indicate skills. In addition, with further analysis of half a billion job offers in English, we found by sampling based approaches that approximately 80% of employment histories of resumes and over 90% of job requirements use skills (in terms of tools, certificates and/or domain-specific expertise) to summarize qualifications. This observation induces a critical demand for high quality occupational skills taxonomy, by which resumes and job postings could be analyzed and evaluated for placement matching. Moreover, such observation also supports our decision to use resumes and job requirements as signals for professional skills extraction. Figure 1 gives an overview of the SKILL system architecture.

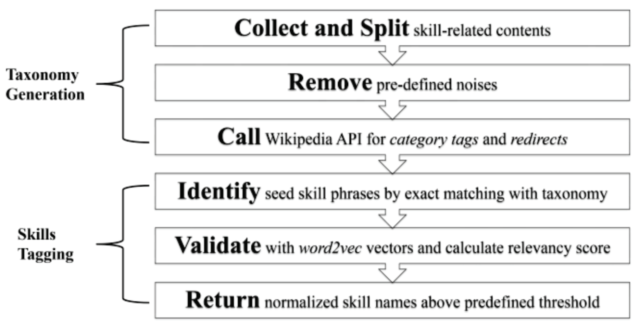

Figure 1: Overview of the SKILL system

In what follows, we will thoroughly describe the generation of skills taxonomy and the process of skills tagging in two subsections, respectively.

2.1 Skills Taxonomy Generation

2.1.1 Skills Definition and Wikipedia

We *define* skills as the content of a known section (e.g., Skills, Technical Skills, Technical Proficiency and etc.) in a resume, or the requirement section in a job posting. Our focus on building the skills taxonomy is based on sufficient retention and appropriate normalization of skill phrases. We assume *all* professional skills expressed in English are inclusive given the variety and veracity of the human capital data we have for training.

Given the presence of an explicit skill section in a resume or a requirement section in a job posting, we model the reading pattern of humans to capture the text blob exhibited in between the headers of skill section and the subsequent section or bottom text. We do not infer the content or meaning of the extracted text blob using typical natural language processing (NLP) techniques at this step, but try to capture as much information (data) as possible and split the text by punctuations, if any. Note that if any input source does not feature sections, we will split the whole context by punctuations. To remove noise, we have a filter dictionary of stopwords[3], adverbs, adjectives, country and city names, common rhetorical phrases and miscellaneous terms predefined on the basis of domain expertise. This dictionary capacitates 90% of noise exhibited in resume skills sections.

After seed skill phrases are collected and properly cleaned, we utilize Wikipedia for both de-duplication and normalization. Wikipedia is one of the largest knowledge repositories on the Web and has been successfully used in numerous large-scale entity normalization and recognition projects (see Section 4). We claim that if the task of skills normalization is given to us as a crowdsourcing project, we will most probably start with a web search query for the seed skill phrase. This web search will most likely lead to a Wikipedia page indicating whether the query is an actual skill or not. We automated this process using MediaWiki[4] API calls and created decision rules based on the *category* tags of Wikipedia documents.

An *open search* action is initiated first for the input query (a.k.a., seed skill phrases from resumes) for proper casing, followed by a *query* action for, if available, associated Wikipedia documents. We created a rule-based system for selecting and analyzing returned Wikipedia documents by *category* tags. The rules are based on the online instruction and definition of the Standard Occupational Classification (SOC) system[5] of the Bureau of Labor Statistics (Cosca and Emmel 2010; Emmel and Cosca 2010; Watson 2013). We used all the keywords from the SOC system documentation to justify Wikipedia *category* tags and make selections accordingly. Note that de-duplication and normalization of queries (surface forms) can be inferred by *redirections* of Wikipedia pages. The input query will be treated as a surface form if the resulting Wikipedia document title qualifies for a skill entity after keyword screening on *category* tags.

By using selected keywords pursuant to the SOC system specifications as screening factors, we are able to retain most phrases directly related to occupational skills. For instance, the term "programming language" is employed to justify some SOC 15 (IT and Math) skills; likewise, the term "accounting management" will indicate a skill in

[3] http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop
[4] http://en.wikipedia.org/w/api.php
[5] http://www.bls.gov/soc/major_groups.htm

business occupations. Another selected group of keywords is used to exclude common noises, which includes "city", "country", "cities", "journal", "conference" and etc., to name a few. In total, out of 46 million unique seed skill phrases, we created a taxonomy of 50K surface forms that is 91% accurate. All surface forms in the current taxonomy are considered independent from each other. Hierarchical structures of skills will be analyzed in future work.

2.1.2 Word Sense Disambiguation

By the nature of natural languages, ambiguity in word senses is a common corollary. We categorize the ambiguity of skill entities into two types: terms with single skill sense (Type-I) and terms with multiple skill senses (Type-II). Again, it is reasonable to assume that all extracted seed phrases can be mapped to meaningful skills, and thus Type-I surface forms are always mapped to corresponding skill senses. For instance, *Python* will always be regarded as a programming language instead of a snake. Likewise, *BI* will be interpreted as an acronym of Business Intelligence.

We define and detect Type-II ambiguity primarily according to the disambiguation page of Wikipedia, if a seed skill phrase is linked to multiple qualified Wikipedia documents. For these terms, disambiguation is conducted through the Google Search API[6]. We employed Google search rankings (by relevancy) for this task because over 95% of the time they are approved by our domain experts. The idea is similar to the *maximum likelihood estimation* in statistical inferences (see, e.g., van der Vaart 2000). Some examples of skill search results from Google Search API: *SVM* is mapped to "Support Vector Machine" and *ZooKeeper* always refers to "Apache ZooKeeper". We are working on a more robust and reliable approach toward the task and will publish more technical details in the near future.

2.2 Skills Tagging

It is important for our skill NER system to avoid misrepresenting an organization (e.g., "Bank Indonesia" abbreviated as BI for "Business Intelligence"), or tagging the resume of an actual zookeeper with some big data skills such as Apache Zookeeper. To implement and produce highly precise and relevant skills recognition system, we utilize *word2vec* (Mikolov et al. 2013). *Word2vec* is an efficient implementation of the continuous *bag-of-words* and *skip-gram* models for computing vector representations of words. These representations can be used in natural language processing applications as equivalent to words represented.

The motivation is to combine a given surface form with corresponding *neighboring* (either semantically or syntactically) phrases to improve relevancy, under the assumption that related skills are likely to appear closely in a recruitment document. We further confirmed from various cross-domain samples that frequency based descriptive approaches are inferior to the inference based neural network language models (results not shown). In fact, a recent systematic comparison also confirms the current state-of-the-art that context predicting vector representations are more accurate than context counting ones (Baroni, Dinu, Kruszewski 2014). However, the performance of *word2vec* is highly dependent on the quality of training data.

In order to properly create relevant skill vectors, we experimented with three approaches to train the data.

- **Original vectors**: vectors trained using all the seed skill phrases extracted from input sources. This is the same data set used for Wikipedia API call for taxonomy building.
- **Surface form filtered vectors**: original vectors filtered by surface forms. Namely, elements of original vectors are retained only if they are contained in the taxonomy. Note that the size of original vectors could reduce dramatically following the screening.
- **Surface form trained vectors**: vectors trained solely by surface forms. Instead of using all the seed skill phrases as for original vectors, we only employ surface forms for training. This approach significantly increases the relevancy of skill vectors, but reduces coverage and complicates taxonomy expansion. If any new surface form is added to the taxonomy, the entire vector space has to be retrained.

We chose surface form trained vectors for production for significant semantic applicability and business congruity.

When training the vector, we have found the n-gram support of *word2vec* a bit limited. Instead of using the built-in n-gram training feature of *word2vec*, we converted all surface forms into uni-grams by replacing spaces with underscores. In doing so, we have observed significant improvement in the semantic relevancy of the trained vectors. Moreover, we choose the *skip-gram* model with *hierarchical softmax* (Mikolov et al. 2011). We have noticed the *skip-gram* model combined with *softmax* boosts the relevancy of vectors for rare skill phrases (e.g., GARCH, a time series model). We set *min-count* as 1 for **surface form trained vectors** to recruit all surface forms while default for training **original vectors**. Note that the *min-count* controls the level of noise allowed, of which the default value is observed to be optimal by our experiments when noise level is moderate. We increased the vector size to 200 from default of 100, because we have found the size most comprehensive in representation. It is important to mention that the assignment of a surface form with a related *word2vec* vector is a *bijective mapping*. Based on our

[6] https://developers.google.com/web-search/docs/

experiments, these settings produce the sufficient vector representation and relevancy for downstream analysis, given our input data of about 100M converted phrases.

The tagging process is controlled by relevancy scores of surface forms. Given a surface form matched from an input text, the relevancy score is the percentage of co-existing (from the input text) surface forms reflected in the vector of the surface form out of all co-existing surface forms matched.

Formally, let Γ be the set of all candidate surface forms matched from a document. Then for any surface form $x_i \in \Gamma$ and its *bijective* vector v_i, the relevancy score λ_i, is defined as:

$$\lambda_i = \frac{\sum_{j,x_j \in \Gamma} I_{v_i}(x_j)}{\sum_{j,x_j \in \Gamma} I_{\Gamma}(x_j)}, \tag{1}$$

where for set A, $I_A(x)$ is the indicator function, such that $I_A(x) = 1$ if $x \in A$, and 0 otherwise. Note that we control the vector settings such that $x_i \notin v_i, \forall i$, i.e., a vector is not inclusive of its surface form.

Currently, we keep the threshold at 2% justified by domain expertise on cross-SOC sample resumes. Any surface form with a relevancy score higher or equal will be recognized and normalized from a given document. We observe such threshold tend to generate the most balanced results between relevance and coverage. To ensure output of skill NER for short input text, we automatically turn this feature off when input text consists of **150** words or less (chosen based on results discussed in Section 3.2).

3 Evaluation

In this section, we evaluate the SKILL system. More specifically, we evaluate the quality of both the taxonomy and entity recognition. We evaluate the developed taxonomy using both a sampling based data-driven approach as well as comparison to a gold standard taxonomy that we use as a baseline. We evaluate the tagging system through user survey by *random sampling in estimating for proportion* (Govindarajulu 1999). For this we use precision and recall as metrics. The taxonomy generation and skills tagging are evaluated separately using slightly different approaches.

3.1 Ontology Evaluation

Ontology evaluation is increasingly implemented in various applications in the field of semantic web and has been frequently discussed in recent literatures (Brewster et al. 2004; Dellschaft et al. 2006; Fernández et al. 2009; Ulanov et al. 2010). See (Brank, Grobelnik, and Mladenic 2005) for an overview of the topic.

We evaluated our skills taxonomy using a hybrid method of both data-driven and gold standard comparison. The European Centre for the Development of Vocational Training (Cedefop) and external stakeholders sponsored an ongoing project of a multilingual classification system titled *European Skills, Competences, Qualifications and Occupations* (ESCO)[7]. There are 5K skill/competence entities supported by the pilot version of the current ESCO system, compared to 50K in our taxonomy. Due to ontology imbalance in size, we only use ESCO as a gold standard for recall. We apply a sampling based data-driven method for estimating precision.

Formally, assuming the true precision of the system is P, and h is the error rate of estimated precision $\hat{p}$ at confidence level C that satisfies (Cochran 1977)

$$\Pr\left(\left|\frac{\hat{p} - P}{P}\right| < h\right) = C, \tag{2}$$

where given sample S, and its true positive set $S_{TP} \subseteq S$, the sample precision can be calculated as

$$\hat{p} = \frac{\sum_i I_{S_{TP}}(x_i)}{\sum_i I_S(x_i)}. \tag{3}$$

Therefore, the sample size n of the random sample S from a population of size N is determined such that

$$\begin{cases} n_0 = \dfrac{z_\alpha^2 (1-P)}{h^2 P} \\ n = \dfrac{n_0}{1 + (n_0 - 1)/N} \end{cases}, \tag{4}$$

where $\alpha = (1 - C)/2$, and z_α is the upper $\alpha/2$ quantile of the standard normal distribution such that

$$\Pr(Z < z_\alpha) = 1 - \frac{\alpha}{2}, \quad Z \sim N(0,1).$$

In our evaluation process, we set $h \equiv 0.05$ and $C \equiv 95\%$. This means the data-driven evaluation is implemented such that **95%** of the time the estimated sample precision is no more than **5%** different from the true population precision.

Furthermore, assuming the ESCO library as a gold standard and thus a set of true positives, we define recall as

$$R = \frac{\sum_i I_{ESCO \cap SKILL}(x_i)}{\sum_i I_{ESCO}(x_i)}. \tag{5}$$

Note that the recall here is a population parameter such that the entire taxonomy of the current SKILL system is compared against the gold standard ESCO library by *direct matching*.

In summary, the estimated precision of the SKILL taxonomy generation is **91%** and recall is **76%**.

[7] https://ec.europa.eu/esco/home

3.2 Tagging Evaluation

The second component of the proposed SKILL system is the skills entity recognition and normalization. We evaluate the NER capacity of our system through sampling based user survey. While an automatic approach is also applicable and available, we believe who knows best about their skills are the users themselves.

We selected active users (those who submitted job applications) from Feb 2014 to Aug 2014. Note that monthly active users of Careerbuilder.com are relatively stable and steady ($\pm$5% in size for the past 4 years insofar) across all industries (categorized by SOC), so any monthly data can represent the overall user profiles without confounding from seasonality or industry variation. The only reason that we selected 6 months of active users is for the purpose of collecting sufficient sample data in a timely manner. Usually, our expectation of the response rate is exceedingly low.

We tag the most recent resume of a selected user with our trained taxonomy and present the top 10 skills by relevancy scores to the user for validation. While the actual number of skills recognized is up to 14 times higher, we restrict the number of skills to be displayed in the survey to facilitate a simple and friendly survey experience. To measure recall, we request the user to add up to 5 skills missing from the presented list.

We received over 3K responses within 2 weeks of the experiment. By formula (2) – (5), this is more than sufficient to conclude the evaluation metrics. More specifically, the overall NER precision is **82%**, and recall is **70%**. It is worth mentioning that more than **80%** of the participated users approved at least **70%** of identified skills. See Table 1 for a breakdown of user responses.

% of Approved Skills	# of Responses	Cumulative %
100%	902 (28%)	28%
90% - 99%	661 (21%)	49%
80% - 89%	618 (19%)	68%
70% - 79%	432 (13%)	81%
60% - 69%	251 (8%)	89%
50% or less	352 (11%)	100%

Table 1: Breakdown of User Responses

4 Related Work

NEN and NER are well-studied tasks that have been applied to solve numerous domain-specific problems in the areas of information extraction and normalization. An early approach that leveraged Wikipedia for large-scale NER and disambiguation tasks is discussed in (Cucerzan, 2007). The approach employed Wikipedia to extract surface form-to-entity mappings and corresponding category tags and used co-occurring terms for disambiguation. NEMO, a system for extracting and normalizing organization names from biomedical publications is described in (Jonnalagadda and Tonham 2011). It utilizes a multi-level rule-based approach for NER, and an extended version of local alignment string similarity metric to determine whether two organization names are related.

(Magdy et al. 2007) tackle the problem of cross-document Arabic person name normalization by training a machine learning classifier on several features such as edit distance, shared name mentions and phonetic edit distance amongst others. The Wikilinks project (Singh et al. 2011) leverages Wikipedia and distributed inference on a graphical model for large-scale cross-document entity resolution. The approach also avoids local maxima for cross-document entity resolution by utilizing Markov Chain Monte Carlo (MCMC) inference and simulated annealing instead of more straightforward greedy approaches. ZenCrowd (Demartini et al. 2013) combines probabilistic automatic techniques for NEN with crowdsourcing within a framework that uses a three-stage blocking process for dealing with large data sets. While the system can't do real-time NEN due to its human intelligence component, the combined approach results in higher precision than their baseline fully automatic or fully crowdsourced systems. A system to extract job skills from text is discussed in (Kiyimaki et al. 2013). It uses the LinkedIn skills taxonomy and applies the spreading activation algorithm on the Wikipedia hyperlink graph to extract both explicitly stated as well as inferred skills. ESCO is a project by the European Commission and various stakeholders to identify and categorize skills, occupations and other relevant competencies. While we know what steps were taken to create and refine the various taxonomies, no information is available on the methodology and techniques used.

Similar to many approaches discussed, SKILL leverages Wikipedia for extracting entities and surface form to entity mappings. However to the best of our knowledge SKILL is the only system that leverages a context-predicting distributional semantic model for contextual skills tagging. SKILL is most similar to (Kiyimaki et al. 2013) as both approaches extract job skills from text. However SKILL is more extensive because it incorporates a mechanism to build a skill taxonomy and uses a deep learning-based approach for intelligent contextual skills tagging.

5 Conclusion

We developed, deployed, and evaluated SKILL, a system for skill entity recognition and normalization. SKILL is designed to meet the increasing business need of workforce analytics implemented at CareerBuilder in attempt to close the skill-gap in the U.S. labor market. There are 50K surface forms mapped to 30K skill entities in the current system. We evaluated the SKILL system using random sampling based approaches and user surveys. We have achieved **91%** accuracy and **76%** recall on taxonomy building and **82%** accuracy on actual skill tagging with **70%** recall.

We have exposed the SKILL system as an API service, serving various teams within the organization and its subsidiaries. The SKILL system has been successfully implemented in multiple workforce analytics and business intelligence projects, revealing critical insights in labor market trends for both customers and project managers. We have completed a final evaluation of the system and rolled out to production.

6 Acknowledgement

We would like to acknowledge Qinlong Luo and Tae Seung Kang for insightful comments on the editing and formatting of the paper. We would also like to acknowledge Chris Min, the Data Warehousing and Big Data Platform teams of CareerBuilder for helping with testing the SKILL system.

References

Baroni, M., Georgiana D., and Kruszewski, G. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 238-247. Baltimore, MD: ACL Press.

Brank, J., Grobelnik, M., and Mladenic, D. 2005. A survey of ontology evaluation techniques. In *Proceedings of the Third Conference on Data Mining and Data Warehouses, 166-170. Ljubljana, Slovenia: Conference on Data Mining and Data Warehouses.*

Brewster, C., Alani, H., Dasmahapatra, S., and Wilks, Y. 2004. Data Driven Ontology Evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation. Lisbon, Portugal: International Conference on Language Resources and Evaluation.*

Cochran, W. 1977. *Sampling Techniques, 3rd Edition*. John Wiley.

Cosca, T., and Emmel, A. 2010. Revising the Standard Occupational Classification system for 2010. *Monthly Labor Review* 8: 32-41.

Cucerzan, S. 2007. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Prague, Czech Republic, 708–716.

Dellschaft, K., and Staab, S. 2006. On How to Perform a Gold Standard based Evaluation of Ontology Learning. *In Proc. of ISWC-2006 International Semantic Web Conference*. Athens, GA: Springer, LNCS.

Demartini, G., Difallah, D. E., and Cudré-Mauroux. P. 2013. Large-scale linked data integration using probabilistic reasoning and crowdsourcing. *The VLDB Journal* 22, 5 (October 2013), 665-687. DOI=10.1007/s00778-013-0324-z

Emmel, A., and Cosca, T. 2010. The 2010 SOC: A classification system gets an update. *Occupational Outlook Quarterly* 54(2): 13-19.

Fernández, M., Overbeeke, C., Sabou, M., and Motta, E. 2009. What Makes a Good Ontology? A Case-Study in Fine-Grained Knowledge Reuse. In *Proceedings of the 4th Asian Conference on The Semantic Web*, 61-75. Shanghai, China: Springer, LNCS.

Govindarajulu, Z. 1999. *Elements of Sampling Theory and Methods*. Prentice Hall PTR.

Jonnalagadda, S. and Topham, P. 2011. NEMO: Extraction and normalization of organization names from PubMed affiliation strings. *Computing research repository,* vol. abs/1107.5743.

Kivimaki, I., Panchenko, A., Dessy, A., Verdegem, D., Francq, P., Fairon, C., Bersini, H., Saerens, M. 2013. A graph-based approach to skill extraction from text. In *Proceedings of TextGraphs-8 Workshop. In Empirical Methods for Natural Language Processing (EMNLP 2013)*. Seattle, USA.

Magdy, W., Darwish, K., Emam, O., and Hassan, H. 2007. Arabic cross document person name normalization. In *Proceedings of the Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*, Prague, Czech Republic.

Mikolov, T., Deoras, A., Povey, D., Burget, L., and Černocký, J. 2011. *Strategies for Training Large Scale Neural Network Language Models*. In Proceedings of the 7th Automatic Speech Recognition and Understanding Workshop, 196-201. Big Island, HI: IEEE Computer Society.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Peters, M., and Wessel, D. 2012. A Jobless Dilemma: What's Wrong With Fort Wayne? *The Wall Street Journal.* http://online.wsj.com/news/articles/SB10001424127887323316804578161141400688884.

Singh, S., Subramanya, A., Pereira, F., and McCallum, 2011. A. Large-scale cross-document coreference using distributed inference and hierarchical models. *In Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, Portland, Oregon, 793–803.

Ulanov, A., Shevlyakov, G., Lyubomishchenko, N., Mehra, P., and Polutin, V. 2010. Monte Carlo Study of Taxonomy Evaluation. In Proceedings of the 21st Workshop on Database and Expert Systems Applications, 164-168. Bilbao, Spain: IEEE Computer Society.

van der Vaart, W. 2000. *Asymptotic Statistics*. Cambridge, UK: Cambridge University Press.

Watson, A. 2013. Implementing the 2010 Standard Occupational Classification in the Occupational Employment Statistics program. *Monthly Labor Review*