LEARNING DOMAIN SPECIFIC SEMANTIC WORD RELATIONSHIPS FROM MEDICAL DOCUMENTS MILESTONE REPORT

DAVID CALLENDER, RUI CHEN, ZHENSHUAI DING

1. Introduction

As stated in our project proposal we are exploring the use of unsupervised learning algorithms to analyze text data that is relevant to the bio-medical field. Specifically, we have been using the efficient and effective word2vec neural network to analyze the abstracts of journal articles in the medical field, in hopes of determining the ability of the abstract features it produces to predict word analogies representing medically relevant semantic concepts.[2]

We have been able to come up with promising initial results and refinements, and are on target with the goals the we perceived were essential for being on schedule with regard to the final project deadline.

2. Progress Summary

2.1. **Summary of proposed milestone goals.** As stated in the project proposal our milestone goals can be summarized as:

"At the time of milestone, we will setup our training data, figure out the useful vocabulary, implement tools to run basic tests on our data and improve their accuracy."

Specifically, we had two sets of objectives: primary goals that were necessary to meet to be on track for successfully completing the project, and secondary goals that we would work towards time permitting:

Primary goals:

(1) Create sets of word pairs with known semantic connection by extracting them from UMLS.
(2) Collect raw text data, the abstracts from medical journals or conference paper. (shoot for 1 billion words).
(3) Write tools to extract abstracts from text and format the data for learning software.
(4) Based on word frequencies in abstracts and on word pairs from UMLS, create our master vocabulary (list of words we care about).
(5) Get word2vec implementation functioning on our corpus of medical paper abstracts.

Secondary goals:

(1) Convert text data into bag-of-words format so that LDA model can also work on our dataset.

Date: October 28, 2014.

(2) Write tools using LDA model to work on the same function as word2vec model.
(3) Initial test for accuracy starting with smaller vocabulary and small corpus.
(4) Compare word2vec and LDA results on various size corpora and vocabulary to each other and to results from word2vec paper.

The progress we have made towards these goals can broadly be categorized as follows. First, obtaining training data for the unsupervised learning algorithm and developing tools to process and clean the data. Second, obtaining known (true) semantically related medical terms to evaluate the abstract vector representations of the medical terms. Finally, performing initial tests of the performance of the word2vec algorithm using the accuracy metric proposed by Mikolov et al. [4, 2]. In this section we will present our efforts and accomplishments for the first two categories. We will devote a separate section to a discussion of our initial results.

2.2. **Training data.** In our project proposal we suggested two sources of text corpora for training: journal abstracts obtained from PubMed[1] and the US government's online repository of Clinical trials.[2] Concurrent with the work done by Mikolov et al.[2] our final goal is to obtain a corpus on the order of 1 billion words. We estimated that if the average length of journal articles exceeded 375 words we could acquire a 1 billion word corpus by searching for articles using the keyword "medicine." We were able to download an approximately 30 GB XML file of citations from PubMed using this search. However, after extensively cleaning the data by removing punctuation, symbols, and numbers, as well as converting everything to lowercase; the resulting data set was more like half a billion words (521,282,646). Interestingly, we discovered that searching for the term "diseases" resulted in more citations for articles and indeed yielded a proportionately larger corpus of 656,346,892 clean words. We deemed this corpus perfectly sufficient for our initial tests. We have yet to evaluate our proposed alternate/additional data set of clinical trial descriptions. This file in XML format is only 1 GB, so it's contribution after cleaning may be negligible.

Related to the goal of obtaining our training data, was the necessity of developing tools for processing and cleaning the data. As already mentioned, we made much progress in this regard. Working with such large files necessitated utilizing more sophisticated XML processing tools than what is provided in basic libraries and tutorials. Generally, XML parsers function by generating a parse tree in memory of the entire XML file, which is not possible with a 30 GB file. While it would have been possible to develop tools that first fragmented the large XML file into individual files representing each citation; this would have resulted in millions of files. Fortunately, we were able to find a highly optimized parser capable of sequentially extracting the pertinent abstract text by parsing the file character by character without constructing a full in memory tree representation.[3] We then combined this tool with custom processing code to clean the extracted abstract text by removing undesirable characters and converting the words to lower case.

[1] http://www.ncbi.nlm.nih.gov/pubmed
[2] https://clinicaltrials.gov
[3] http://www.ibm.com/developerworks/xml/library/x-hiperfparse/ ; Author: Liza Daly. See also http://effbot.org/zone/element-iterparse.htm

2.3. **Known binary semantic concepts.** Obtaining the known binary semantic relationships to be used for accuracy evaluation proved to be one of the more challenging and time consuming components to our milestone progress. While we were aware that the Unified Medical Language System (UMLS) was a large and complicated database consisting of a conglomeration of several disparate medical term thesauri, the ample documentation that we found lead us to believe that extracting the examples of semantically related terms that we needed would be straight forward. Indeed, now that we have the tools in place it is relatively straight forward. However, getting to this stage was not. It seems that the typical use case for the UMLS is to start with a term of interest and then see what information you can gather about that term and related terms. So, much of the documentation and tools provided by the UTS were not directly applicable.[4].

For a time, we were concerned that the database may not even yield a suitable testing set. Many of the medical terms are actually phrases and in cases have multiple distinct but equivalent textual representations. There are several indexing systems for the terms spread out over approximately two dozen tables. That said one of the indices does correspond to unique or 'atomic' representations. Restricting the examples to ones involving this atomic index proved to still yield 381 possible classes of semantic concepts. Though many of these classes of concepts did not have enough specific examples for our uses after further restricting to examples where both terms in a semantic relationship were single words and not phrases. Unfortunately, at the time of writing this progress report we do not have exact details on which concept classes will ultimately prove viable. Due to the size and complexity of the database, the SQL queries to extract these concepts are time consuming. Even more so than training the word2vec algorithm. The script we wrote to perform these queries is still running and has been doing so for several days.

Fortunately, earlier in the process we identified a particular relationship class that proved both viable and interesting. This semantic relationship involves diseases and symptoms that often co-occur (for example HIV and AIDS). This concept is interesting as it raises the possibility of being able to identify new and unknown connections between diseases, if say our training data were clinical notes. Thankfully the UMLS contained 13,768 examples of the 'co-occurs_with' concept where both terms in the relationship utilize the unique atomic index. Even after further restricting to examples where the terms were single words, we were left with 1,232 examples of pairs of linked words. Because our accuracy evaluation is based on using the vector offset of one pair of words to evaluate another, this would mean more than 1.5 million test cases.

As it turns out there are properties of this semantic class, not present in the semantic relationships evaluated in the original word2vec paper, that have a negative impact on our very stringent accuracy measure. Namely that the semantic relationship is both symmetric and not one-to-one. By symmetric we mean that if HIV co-occurs_with AIDS than AIDS co-occurs_with HIV; and by not one-to-one we mean that there exists examples such as PALPITATION co-occurs_with ANXIETY and PALPITATION co-occurs_with TACHYCARDIA. Given that our accuracy measure looks for the single word in our vocabulary that minimizes the

[4]https://uts.nlm.nih.gov/home.html

error function in Equation (1), these two properties mean that we will necessarily, at best, be achieving an accuracy of $1/n$ for test cases where w_{b2} appears in n distinct relationships of this class. As will be outlined in the results section, it proved necessary to restrict the test cases to those examples of word pairs where both words were unique to that specific relationship. This left us with only 86 word pairs and 7396 test cases.

$$f(a,b) = \begin{cases} 1 & \text{if } w_{b1} = \operatorname{argmin}_{w \in \text{vocabulary}} ||(w_{a1} - w_{a2}) - (w - w_{b2})|| \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Amongst the 380 remaining semantic concept classes some of the relationships are not symmetric. For these concept classes it will only be necessary to remove examples where the first word in the relationship appears in more than one example. Indeed in one of the semantic relationships tested by Mikolov et al.[2] there exists this one-to-many property. This is the relationship of City-in-state. Clearly a state can have multiple cities.

3. Initial results

Our initial test of word2vec was done using a somewhat restricted training corpus of 172,638,903 cleaned words taken from the more recent journal articles. We used the skip-gram variant of word2vec which has been shown to do better for semantic relationships, and a vocabulary of 154,023 words. We set the dimension of the abstract feature vectors that word2vec would compute to be 500 (This is essentially the 'range' dimension of the projection matrix). The test set was made up of the full 1.5 million examples of diseases and symptoms that co-occur where the terms are not multiword phrases. This yielded a rather discouraging accuracy of 2.548018%. Curious as to what was causing this undesirable result, we tested the accuracy where we called it a 'win' if the true word was amongst the 10 nearest neighbors rather than just the single nearest neighbor. This did boost the average accuracy to 12.24381%. Ultimately, manual inspection of some of the test examples illuminated the problem that there were numerous examples of diseases co-occurring with more than one other disease. This was confirmed by the histogram we produced of the number of terms that appear with a given frequency as the first term in the relationship, shown in Figure-1.

To correct for this problem we trimmed the test set to only the one-to-one cases of related words. By this point we had also trained word2vec on the larger 600 million word corpus. Because we were down to 86 pairs of words in this semantic class we were worried that many of the terms might not be in the vocabulary known to word2vec. Fortunately, 6475 of the 7396 test examples were viable, and our accuracy score improved to 10.080272% using the single closest vector measure and 28.461538% using the 10 nearest neighbors.

4. Future work and final project goals

There is much additional work the we can do for the final project completion. Chief amongst this is looking into the viability of the remaining 380 semantic relationship classes. Particularly, the ones that are not symmetric. Given the tools that we have in place it should be straight forward to analyze all 380 to determine

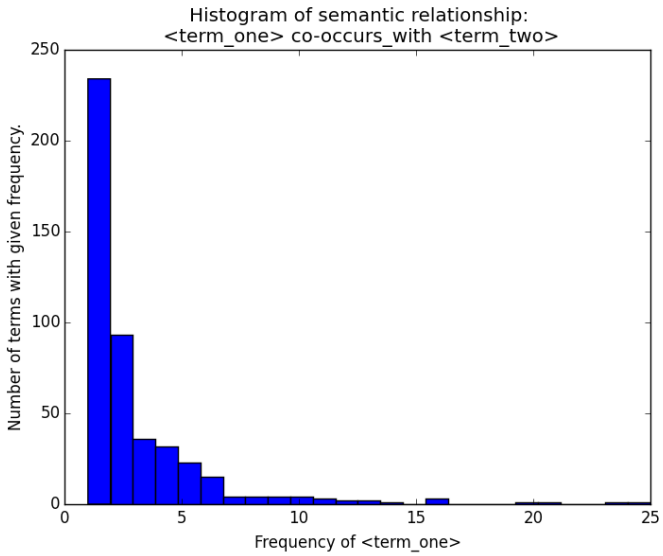

FIGURE 1. Histogram representing the number of terms that occur as the first term with a given frequency. Because the relationship co-occurs_with is symmetric this is equivalent to the histogram representing the second term. While many terms appear only once, the ones that appear multiple times (up to 25) contribute a large number of test examples that will necessarily be wrong given an accuracy metric concerned only with the closest matching word vector.

if they are viable and to determine the performance we can achieve predicting their analogies.

Additionally, there is much work we can do to improve our initial accuracy results. There are several parameters to word2vec that can be adjusted. Perhaps the most important of these is the dimension of the abstract feature vectors it produces. The original word2vec paper achieved 60% accuracy partly by increasing the vector dimension to 2k. Ideally we would also increase the size of our training corpus, though this may not be possible. It may not be possible to achieve 60% accuracy on semantic relationship classes that undoubtably appear less frequently than those evaluated by Mikelov et al.[2]

As suggested at our milestone presentation, we will also attempt to use the precomputed word vectors being distributed by the word2vec authors. It will be interesting to see how well they do, given that they are trained on documents not specific to the medical field. At the same time, they are trained on a larger text corpus, which perhaps can compensate for the broader subject matter. Additionally, we will look at less stringent accuracy measures, based on the k nearest neighbors, for a variety of k other than 1. As well, we will determine baseline accuracy by creating test cases by selecting 4 random (and presumably unrelated) words. Given the vocabulary size and vector dimension this should be a relatively small baseline, but certainly worth verifying.

We are also still considering Latent Dirichlet Allocation (LDA) as a unsupervised learning algorithm. Time permitting, it would be interesting to see how it compares to word2vec on the smaller text corpora that it can handle.

References

[1] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[2] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*, 2013.

[3] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*, 2013.

[4] Tomas Mikolov, Wen tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of NAACL HLT*, 2013.

[5] Christos Papadimitriou, Prabhakar Raghavan, Hisao Tamaki, and Santosh Vempala. Latent semantic indexing: A probabilistic analysis (postscript). 1998.

[6] Josef Sivic. Efcient visual search of videos cast as text retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31:591–605, 2009.