Battery Vision

Using Word2Vec to Extract Details from Battery Literature.

Sterling Baird[1], Kennedy Lincoln[2], David Morley[3]

[1]BYU, Dept. Mechanical Engineering, [2]BYU Dept. Physics and Astronomy, [3]BYU Dept. Mathematics

Theory of Predictive Modeling

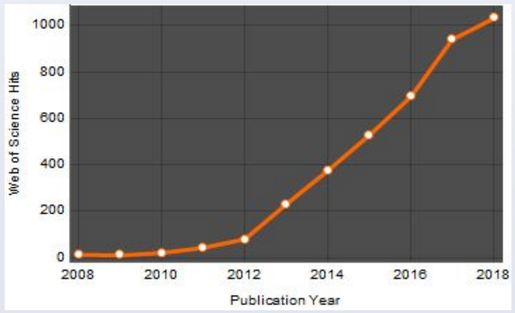
Background

- Types of quantities in scientific literature: experimental, theoretical, or reference to other work
- Researchers resort to manual methods to extract relevant data and find trends (labor intensive and limited to small data volumes)

Problem Statement

- Decrease the time of **data collection** via an automatic process by modeling the **relationship of words** in lithium-sulfur battery literature (due to the large volume of data) to gather context of the parsed text.

Figure 1. The rapidly increasing number of publications in the field of lithium-sulfur batteries. Results obtained via Web of Science analytics using keyword "lithium-sulfur".



Methods

Step 1: Import the data

- Three methods of importing "raw" raw data from digital object identifier (DOI) links: manual saving of .html files, ContentMine's quickscrape tool, and Mathematica's URLDownload function (Figure 1).
- Manual extraction: 100% success in retrieving the full-texts, but most time-consuming method (Figure 1.)
- Quickscrape: more full text articles than Mathematica, but also more unusable downloads (i.e. empty file or no relevant info).
- Mathematica: best choice because of speed (~5 min for 200 links), convenience (2-lines of code) and consistency of output.

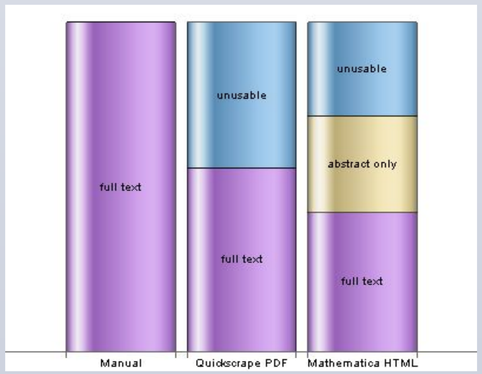


Figure 2. A comparison of various data extraction methods from DOI links. It shows the trinary results of full text, abstracts only, and unusable files.

Methods (cont.)

Step 2: Clean the data

- We used a basic parser to remove unwanted (common) words in the text, punctuation, citation numbers, and everything before the abstract and after the conclusion, unified words to their common roots and make everything lower case.

Original Sentence:
'Lithium-Sulfur batteries are an interesting type of battery to study.'
Removing stop-words:
['Lithium-Sulfur', 'battery', 'interesting', 'type', 'battery', 'study.']
Remove duplicates, punctuation, and set to lowercase:
['lithium-sulfur', 'battery', 'type', 'battery', 'studied']

Figure 3. A simplified example of cleaning text for word2vec.

Step 3: Word2Vec

- For input to the word2Vec model, we create "one-hot" (OH) vector encoding of each word (Figure 4a).
- The model then maps each OH to its nearest neighbors using a simple neural network framework (Figure 4b).
- The weights of the hidden layer become the new word encodings that carries information of the context of the word based the words immediately around them
- For example, the encodings of "king" and "queen" would be similar because of their similar use based on words around them.

	"lithium-sulfur"	"battery"	"interesting"	"type"	"study"
"lithium-sulfur"	0	1	0	0	0
"battery"	1	0	1	1	1
"interesting"	0	1	0	1	0
"type"	0	1	1	0	0
"study"	0	1	0	0	0

Figure 4a. One-hot encoding of a test sentence.

- The problem is summarized mathematically via the Softmax model:

$$argmax_\theta \frac{1}{T}\sum_{t=1}^{T}\sum_{j\in c,j!=0} logp(x_{t+j}|x_t;\theta)$$

with

$$p(x_{t+j}|x_t;\theta) = \frac{exp(\theta x_i)}{\sum_t exp(\theta x_t)}$$

where x_i, N, θ, and K represent a one-hot encoded vector, number of words in vocabulary, the embedding matrix, and dimension of learned embeddings, respectively.

- Further, our 'notion of best' arises from how similar words are mapped in this space. Words with similar contexts are close together by the cosine similarity metric.

Additional References and Resources Used For Word2Vc:
https://towardsdatascience.com/word-embedding-with-word2vec-and-fasttext-a209c1d3e12c
https://en.wikipedia.org/wiki/Word2vec
https://skymind.ai/wiki/word2vec
https://medium.com/explore-artificial-intelligence/word2vec-a-baby-step-in-deep-learning-but-a-giant-leap-towards-natural-language-processing-40fe4e8602ba
https://radimrehurek.com/gensim/models/word2vec.html

Methods (cont.)

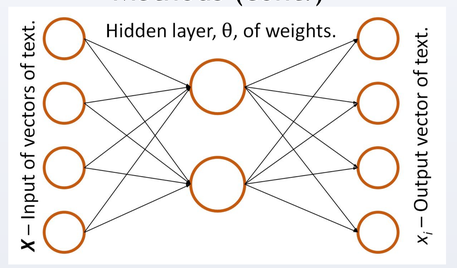


Figure 4b. Schematic representation of a Word2Vec algorithm.

Results

- The words from the vectors are mapped to a 50 dimensional vector space, then projected onto a 2 dimensional plane (Figure 5a) using principal component analysis (PCA) 1, 2, and 16. (Notice the data appears to live on a hypersphere).

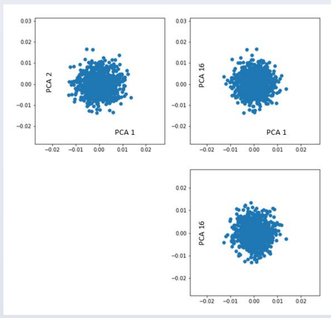


Figure 5a. The plots of Word2Vec results.

- We then sampled random data points from within our training set to measure correlation of words to experimental and theoretical concept epicenters. We found that data found in certain areas of the text had higher correlations than others (Figure 5b).

Data Point	Text Excerpt	Similarity to Experimental	Similarity to Theoretical	Data Point Actual
30 wt	"To reduce the interfacial resistance, 30wt.% solid electrolyte …"	0.058	0.0089	Experimental
1,041 mA h g^{-1}	"It shows an initial specific discharge capacity of 1,041 mA h g^{-1}…"	0.24	0.069	Experimental
0.2C	"When the current rate returns to 0.2C, a reversible capacity of 789 mA h^{-1} g^{-1} remains."	0.055	0.07	Experimental
Average Similarity of All Words: 0.01				

Figure 5b. Table of similarities of data points.

Future Work

- Pull the data into tables, eliminating non-original work and theoretical information based on the above models
- Automatically filter scientific articles into categories (e.g. experimental, computational, review)
- Automate DOI collection through use of web crawler or similar