IDENTIFYING GEOGRAPHICAL LOCATIONS IN MARIANNE MOORE'S POEMS USING NAMED ENTITY RECOGNITION

Spencer Anderson
MehtA+

Ivy Guo
MehtA+

David Aidan Dugan-Lazo
MehtA+

July 28, 2022

ABSTRACT

Marianne Moore, an American poet from the 20th century, refers to many geographical entities in her poetry. However, she doesn't always refer to locations using their proper names. We aimed to identify all the lines of a poem that refer to geographical entities that Moore mentions throughout her poems. Although we have no way to measure accuracy, we found that using named entity recognition (NER) on scraped Wikipedia pages to identify possible locations a word refers to and then filtering these locations with word2vec gives us surprisingly reasonable results.

1 Introduction

1.1 Motivation

Marianne Moore doesn't always refer to locations using their proper names in her poems. For example, when referring to Egypt, instead of directly saying Egypt she might say "scarab" instead as an allusion to Egypt. Given this, how do readers know what Moore is referring to? This question served as our motivation for picking this as our final project. We wanted to create a model that would help people who are interested in gaining a further understanding of Moore's poetry to discern any mention of geographical entities in her many works.

1.2 Goals of Project

Our goal for this project is to identify, using machine learning techniques, all the lines of one of Marianne Moore's poems that refer to a geographical entity and all geographical entities that Moore mentions throughout her poems using named entity recognition and word2vec models.

2 Related Work

Last year, students from MehtA+ worked on identifying all direct locations mentioned in Marianne Moore's poetry[1, 2, 3, 4, 5, 6, 7]. For example, if a poem mentions the word "Egypt," their models would recognize Egypt as a location. However, their models would not be able to recognize other words referring to locations. For example, they would not recognize that "scarab" could refer to Egypt. Our project aims to include this feature and present a more complete representation of all of Moore's references.

3 Methodology

3.1 Data-set

The data-set that we used is a book of 170 of Marianne Moore's poems. This is a challenging data-set due to it being varied in content and short as a whole, so we didn't have many texts to train a model on. To continue, it also contains words that word2vec doesn't recognize, for they were used long ago and are no longer in use by the general population

or were spelled differently when Moore wrote her poems than they are now. In addition, poems unlike other forms of writing use many metaphors and writing techniques that make it difficult to discern the meaning of a word.

Preprocessing We removed all stop words such as "the" and "are" as they would be associated to the geographical entities despite the fact that they aren't alluding to any specific places. We then lemmatized the remaining words and split up each poem by line. Lastly, we removed all lines containing "editor's note" or "Moore's note" because these lines aren't part of the poems.

3.2 Model

3.2.1 Word2vec

We initially tried downloading pre-trained word2vec models. Word2vec represents words as vectors and uses cosine similarity to find the similarity between two words. The similarity ranges from -1 (not similar at all) to 1 (same word). Therefore, for each word in a line, we could find the similarity between the word and each location in a list of locations. The two word2vec models we tried were GloVe[8] and another model trained on the English Wikipedia[9].

3.2.2 Named Entity Recognition

We also tried scraping Wikipedia pages and using named entity recognition (NER). A NER model is used to identify all the named entities (eg. geopolitical entity) in a piece of text. We first found all the nouns and proper nouns in Marianne Moore's poems. We then scraped the Wikipedia pages corresponding to these words and used the spaCy NER model to identify all the locations associated with each word. Since some words are ambiguous and could be associated with multiple Wikipedia articles, we tried both taking the first suggestion and taking a random suggestion. Since this model picks up on all locations, even the ones that are not closely associated with the word at all, we used a word2vec model to filter the locations based on how similar the location is to the line of the poem. We tried multiple thresholds for the similarity to find one that filters out locations that are clearly incorrect but doesn't filter out locations that might be correct. We also tried using various NER models, since the small model from spaCy isn't the best at detecting only locations.

4 Problems and Solutions

Working on this project was not easy throughout and we encountered several problems that halted our progress. However we didn't let these issues stop us from completing our goal and we worked to solve the problems we came across to create the best possible model. One issue that we came across early in our journey was word2vec not functioning as we wanted it to and not writing down the correct similarity between words in the context they are written in. Another problem we came across near the beginning of the project was that to present the accuracy of the model, we would need to know exactly what Moore is referring to in her poems. However, due to the fact that an official list of allusions to geographical entities in Marianne Moore's poems has never been written it would be impossible to know whether or not our model makes accurate predictions. Eventually, we decided to simply subjectively describe the accuracy by looking through the passages ourselves and seeing which similarities seemed reasonable and which ones seemed blatantly incorrect (eg: people and countries). Another problem is that most of the words in texts aren't actually referring to places, but the model might find that it is referring to somewhere. Because of this, we decided to set a threshold for the similarity between a word in the poem and the location to filter out these cases.

5 Results

Our model using just word2vec did not work at all. Our model using NER gave use many reasonable results, as well as a few weird ones.

5.1 Word2vec

As mentioned earlier, word2vec did not adequately recognize similarities between words. For example, the similarity between "scarab" and "Egypt" was less that 0.3, even though Moore often uses "scarab" in reference to Egypt. Because of this, we weren't able to accurately determine what locations each word could refer to. Moreover, just using word2vec requires having a list of all locations Moore might refer to. Since this wasn't possible, word2vec by itself doesn't work.

5.2 Named Entity Recognition

Named entity recognition paired with word2vec worked quite well. There is no way for use to measure the accuracy of this model, but most of the results look reasonable to us. As per 1, the model recognized that the line "A scarab of the sea" is likely referring to Egypt. Since our similarity threshold was 0.25, the model only outputs Egypt as a possibility. Notice that the model gives an error when finding the similarity between "scarab" and "Middle East." This is because "Middle East" is not in the model's vocabulary. For some lines of poetry, the model isn't able to pinpoint the location Moore refers to, but can suggest a few possibilities. For example, the model predicts that the line "With an elephant to ride upon—"with rings on her fingers and bells" refers to Africa, Thailand, or Zimbabwe, all of which are reasonable. However, our model isn't completely accurate because it doesn't pick up on everything. For example, in the line "And this Dante," Dante likely refers to the Italian philosopher, which would imply Italy. However, the model didn't pick up on this, likely because the word2vec model doesn't have "Dante" in its vocabulary. Our model also predicted that some lines of poetry referred to entities that are not actually locations. For instance, our model predicted that the line "longer that; nor did the blue red yellow band" refers to the color white.

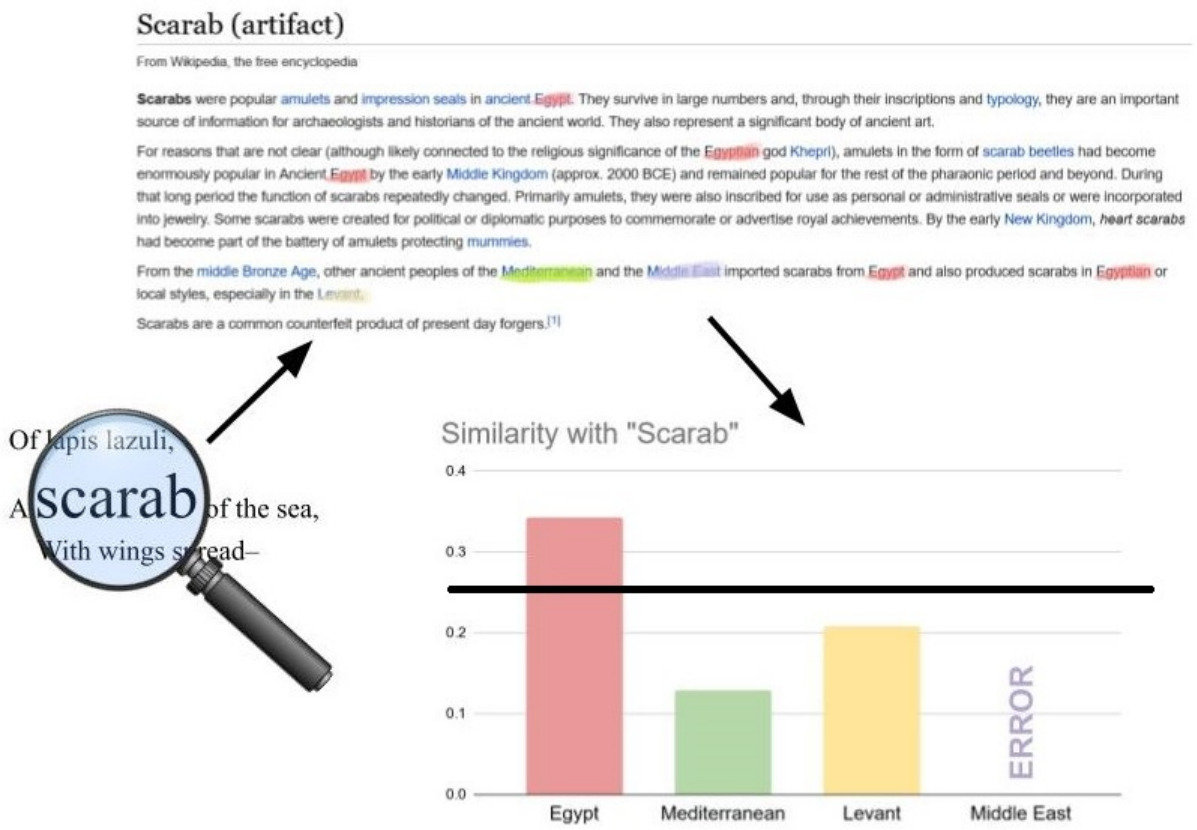

Figure 1: How our model works for one specific line

Table 1: Sample Results

Line	Word	Location
longer that; nor did the blue red yellow band	red	white
And this Dante	N/A	N/A
as the wind changes;	wind	Earth
A scarab of the sea	scarab	Egypt
With an elephant to ride	elephant	Africa
Of Moorish gorgeousness	Moorish	Arabia

6 Conclusion and Future Work

Although our project wasn't the most successful due to many of the hurdles we had to face, it recognized locations in poetry using a variety of methods that all worked together. If we had more time, we would try using bigger NER and word2vec models. This would allow our program to more accurately recognize only the locations in Marianne Moore's poems. This would also allow our program notice all words that Moore uses that might refer to a location since the word2vec model would know more words. For example, a model with a larger vocabulary could recognized that "Dante" is associated with "Italy."

An interesting possible future project would be to train a word2vec model just on the scraped Wikipedia pages. This way, the similarities between words would likely be more accurate in the context of Moore's poetry. For example, the Wikipedia page on Egypt doesn't mention scarab but the Wikipedia page on scarab mentions Egypt. In general, the Wikipedia pages on locations don't mention specific words used by Moore that refer to locations, but the pages on these words reference the locations. Therefore, if we train a word2vec model just on these pages, the similarity between Moore's words and location names would be higher.

7 Our Code

`https://gist.github.com/ivyg15/aeaa2b1fff640f85cb77ed3c7a6775ef`

8 Division of Labor

We divided the work as follows:

- Ivy and Spencer split up working on the coding, where Spencer worked on some of the pre-processing and Ivy worked on scraping the Wikipedia sites.
- The paper was worked on by all members, but while Spencer and Ivy were working on the coding, Aidan made headway on the paper.
- The poster was also a team effort, though due to Ivy's artistic abilities, she helped greatly with making the images.

9 Acknowledgements

We would like to acknowledge our incredible teacher, Ms. Haripriya Mehta for teaching us all of the code that we needed to know for this project and for helping us and guiding us throughout any issues we encountered while working on it, our primary advisor and secondary advisors, Ms. Anna and Ms. Andrea respectively for helping us throughout our final project by assisting us with any problems we had and explaining any topics that we had any particular trouble with.

References

[1] Amanda Lin and Anna Muyan Li. Finding the locations or countries in the poems of marianne moore using machine learning and natural language processing. 2021.

[2] Abdulaziz Khader Isaac Ang, Daniel Suh. Predict poem setting using machine learning. 2021.

[3] Danny Nimish, Hana. Finding locations in marianne moore's poems. 2021.

[4] Justin Wickelgren Alec Situ and Surya Kolluriy. Model for identifying locations. 2021.

[5] Dylan Sheehan Kabir Goel, Siddarth Kappa. Locating geographical references in poems written by marianne moore. 2021.

[6] Pradyun and Aadhavan. Finding geographical locations in marianne moore's poems. 2021.

[7] Nakul Solai Livy Bottomley. Recognizing words relating to places in marianne moore's poems. 2021.

[8] 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.

[9] Murhaf Fares, Andrey Kutuzov, Stephan Oepen, and Erik Velldal. Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In *Proceedings of the 21st Nordic Conference of Computational Linguistic*, page 271–276, 2017.