EFFECTIVE NEURAL INFORMATION RETRIEVAL IN CROP SOIL REQUIREMENT CLASSIFICATION

Ayman Salama[1,2], Tomas Maul[2], Ebrahim Jahanshiri[1], Nur Marahaini Mohd Nizar[1], Neil Crout[2]

[1]*Crops For the Future Research Centre*
ayman.mohamed@cffresearch.org, ebrahim.jahanshiri@cffresearch.org, marahaini.nizar@cffresearch.org
[2]*School of Computer Science, University of Nottingham, Malaysia*
Tomas.Maul@nottingham.edu.my, hcxas1@nottingham.edu.my, neil.crout@nottingham.ac.uk

Word embeddings are numerical distributed word representations that have recently sparked significant interest in the research community. They are used by Google and Facebook corporations in various applications such as enhancing search and recommendation engines. Many word embedding techniques such as word2vec, doc2vec, glove, BERT, RoBERT and others have recently been published. Word embedding methods identify the target word by their accompanying words and convert their textual representation into numerical vector spaces where arithmetic calculations can be done. Extracting knowledge from written text into structured formats such as datasets usually takes a significant amount of time and resources. For example, collecting data about underutilised crops from the literature can be a tedious process. This research aims to partially automate the process of data collection that can be used to build domain-specific decision support systems. To demonstrate the usefulness of neural information retrieval, we propose an application of this method to extract soil requirements of underutilised crops. We trained a word2vec model on 178 different languages with 72GB of Wikipedia text with 8.1 billion words. The trained model was used to predict the soil classification requirements of 2201 crops and the verification was done using FAO Ecocrop data. We tested several arithmetic methods for calculating the crops' scientific and common English names against soil classes in the word embedding vector space. We achieved an accuracy of 76.11% for soil classification against crop scientific names while English common names against soil classes resulted in 49.97% accuracy. The results suggest that it is effective to use neural information retrieval in knowledge extraction. It also suggests the use of scientific names word embeddings for crops provides higher accuracy than common English names in this domain. Future work will involve other types of predictions relevant to agriculture including aspects pertaining to macro and micronutrients, crop types, and crop usage.

***Keywords*:** Word-Embedding, Machine-Learning, Soil-Classification, Underutilised-crops, Neural-Information-Retrieval.

INTRODUCTION

This research focuses on the study of agriculture and food datasets. Most of agriculture and food research centres focus on the main 25 crops which feed more than 90% of calories consumed by the human population. Global knowledge system for underutilised crops is ongoing research conducted in the Crops For the Future Research Centre (Gregory, et. al, 2019). The system provides a relational database and NoSQL database to introduce crops by 4000+ variables and 44 metadata for each variable. The variables are defined in three dimensions, taxonomical, geographical and data sources (Marahaini, 2018). The system can store multiple values for the same variable based on crop variety and reference the value and geographical location where the data is collected from. Several methods are used to populate the database such as the integration of international databases pertaining to climate, soil and crops. Additional extensive manual information extraction is performed from the scientific literature. Finally, approaching governmental institutions for data localization is by far the most exhaustive process for data collection. Value chain data CONNECT and ASSESSCROP are comprehensive experiments conducted through crop value chains to build a relational database to trace information about crops from germplasm, seed inventory, plot and crop management, harvesting and postharvest process, product manufacturing, nutrient sampling and testing, market and consumers (CFF Research, 2019).

Neural Information Retrieval (NIR) is extensively used in web search engines and social media data processing. Search engines and social media recommendation systems are examples of successful applications of NIR. Users obtain 60% to 86% accuracy from the recommendation systems and still achieve decent customer satisfaction (Ricci, 2011). In spite of NIR's success, sensitive scientific studies like medical and agricultural sciences limit NIR applications possibly due to stricter accuracy demands (Onal et al., 2018) and, in some cases, lack of model interpretability.

Word embeddings are distributed word representations in a relatively low dimensional vector space. Processing text by neural networks requires a form of numerical transformation of text which in this case corresponds to numerical vectors. Word embeddings as a fundamental concept were introduced earlier in the 1950s (Harris, 1957) and (Firth, 1957), under distinct terminology. With the evolution of neural networks, word embeddings have become a core concept in their own right, with a broad spectrum of applications. Tomas Mikolov and colleagues published several papers discussing and presenting promising solutions to enable machines to learn high quality distributed vector representations to capture syntactic and semantic relations among words in a corpus. There are several techniques and libraries that have been published like word2vec (Mikolov, 2013), doc2vec (Le., 2014), glove (Pennington, 2014), BERT (Devlin, 2018) and others.

One of the advantages of word embeddings is that users don't have to manually annotate any datasets with explicit labels. Self-supervision is a relatively recent machine learning method to autonomously label the training dataset without human interaction. It is learning from existing words and phrases without the need for human interpretation (Hinton, 1999). How word embeddings are used defines their unique abilities. Here are some examples of how word embeddings can be used efficiently in different problems.

Word embedding usage

1. Automated text tagging. Nikfarjam et al. (2015) managed to extract features of drug reaction from social media corpora using word embeddings and claimed to achieve 82% accuracy which is an improvement relative to the baseline studied.
2. Recommendation Engines. Ozsoy (2016) studied the design of a recommendation system with word2vec embeddings from social media Foursquare's check-in dataset and provided promising results.
3. Synonyms and search query expansion. There is a claim that Google is using word embeddings like word2vec to recommend to the user alternative search keywords (Gao, 2017).
4. Crop suitability. Word embeddings were used in an agricultural tool - SELECTCROP - developed by the Crops For the Future Research Centre (Gregory et. al, 2019). The user of the tool searches for crop suitability in a specific location. The crop name can vary from one language to another and even from a location to another in the same language. The user might use a local name to search for a crop. The search engine will take the user query and use word2vec to get the synonym of that local name and identify the scientific and English common name and extract the crop features and calculate the crop suitability (Salama et al, 2017).
5. Machine translation. Training word embeddings on two different translations for the same content. (Mikolov, 2013) demonstrated the idea of using monolingual data and mapping it with bilingual data. The distributed representation has the potential to extract the similarity between vector spaces and effectively do the translation. The result of Mikolov's experiment achieved 90% precision for translating between English and Spanish. Chelba et al. (2013) published a new data set of one billion words to be used as a measurement tool for statistical language modelling. This corpus has the potential to be used for the evaluation of the translation along with word embeddings.
6. Question answering. Although there are several types of research towards automating an AI agent to answer human questions, Weston et al. (2015) believed that there is no comprehensive system till date that can achieve it. They proposed that by using word embeddings powered with an enhanced memory network model, having an automated AI question answering system is possible.
7. Classifying a movie by genre through studying the plots of movies (Schnabel, 2015).
8. Classifying restaurants by themes (e.g. "Scantic, Jazzy") by studying user reviews (Liu, 2015).
9. Sentiment analysis. Classifying user reviews tends to be a very time-consuming task. There are several supervised learning methods for sentiment analysis, however using word2vec can be an easier way to

approach sentiment analysis (Mesnil, 2014). In this work, the Facebook AI research group studied an IMDB movie review dataset using several combined machine learning approaches, including NB-SVM, RNN-LM and sentence vectors. They have published their code to make it easy to reproduce their results for transparency.

10. The usage of word embeddings to detect textual anomalies will enable the extraction of new features from text like social media Facebook status or tweets.

Word2vec explained

Word2vec is a neural network that acts as a word embedding algorithm. It turns a corpus of text into a mathematical vector space. Word2vec was developed by Google and its code was released online with the original model trained on 1.6 billion words. At the time of its release, its performance was described as reaching state-of-the-art level for measuring syntactic and semantic word similarity (Mikolov, 2013).

How does word2vec work?

The neural architecture of word2vec is very simple and consists of a single hidden layer. The embeddings themselves are extracted from the network's weights. The weights of the word2vec network are randomly initialized and then the network keeps trying to adjust its weights by going through sentence by sentence until the weights are optimised in a manner that minimizes the prediction error. In one variant, the network aims to predict the centre word within a contextual window of words. Conveniently, the resultant word embeddings can be mathematically calculated against each other to answer different linguistics questions involving similarity, relationship, and semantic meaning. The most common example that is used to describe this is "king, man, queen, woman" vectors. If the corpus is large enough, the vector of "king" and "man" should be close together. Likewise, the vector for "woman" and "queen" should be close together. The elegance of word2vec reveals itself in the ability to demonstrate this equation "Vector[king] - Vector[man] + Vector[woman] ≈ Vector[queen]". The simple conclusion is that word2vec can successfully represent the semantics of words in any given corpus though distributed representations extracted from the weights of the network's hidden layers. This vector space can be used mathematically to extract meaning and quantify the relationships between different entities in any given corpus.

METHODS

Word2vec experiment with agriculture data:

Earlier we conducted two experiments to understand the geographical utilization patterns of underutilised crops. The first experiment used Google search engine data extracted from the Google AdWords tool for a period of two years for all worldwide users searching for a specific crop and using all known crop scientific and local names. Text mining and Google maps were used to interpret 300K search events and visualize them to understand the global cropping pattern with time series analysis. The second experiment expanded the targeted data from search engine events to corpus of data of English Wikipedia. The expansion of input data revealed a different approach to understand the vernacular knowledge of crops. The first experiment integrated international databases for vernacular knowledge to construct a comprehensive list of 30 names of a designated crop. Google Adwords data processing required significant manual work which imposes human error and possible distorted accuracy that required several corrections. NIR in the second experiment created a vector space of English Wikipedia corpus that transferred the focus of the experiment from not only the vernacular knowledge but also to all aspects of crop entities. Depending only on the English Wikipedia corpus created a gap for the vernacular knowledge extraction giving the fact the underutilised crops are possibly domesticated and utilised by native communities distanced from the agriculture industry which means the corresponding knowledge is mostly stored in a non-English language. The first objective in this research was to expand the base knowledge to accommodate non-English languages, whereby we studied available Wikipedia corpus in 178 languages and massive gigantic vector space for all the languages.

Training word2vec on Wikipedia all languages

The training was done in two steps. First, the text space of all XML files for all languages was extracted. Then we combined all generated processed text (from all languages) and further processed it into a vector space model to be used for later analysis. The idea is very simple and yet effective, encapsulated in the steps below. System specifications consisted of: Dell PowerEdge R730, 16 Core CPU, 128G RAM; Centos 6.5 Final.

1. Perform part one of the process which is to obtain the text of Wiki corpora for all languages.
2. Concatenate all output text files together using Linux shell scripting; the new size is 72G.
3. The next challenge is to setup the word2vec model for subsequent training on the 72G of data.
4. Train the word2vec model; training was done successfully in 29 hours 13 mins.

Creating a test set from text mining on Ecocrop

The training of word2vec on all Wikipedia languages was done in a standard self-supervised manner, using a two-layer neural network. Self-supervision in this context means that the output labels are automatically generated from the raw text (e.g. words at the centre of a contextual window), rather than requiring manual curation, which would be prohibitively expensive in this domain. The vector space that we have created from training word2vec on all Wikipedia languages can be used to quantify any relation between any entities. In this experiment, we aim to use this vector space for identifying the relationship between crops and other entities like countries, local names, soil types, and diseases. The pseudocode diagram below indicates the steps to process the data and create a test set, similarity index and ranking accuracy for evaluation purposes.

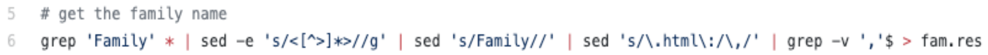
The three soil properties are *clay*, *loam*, and *sand*. We used these keywords because they are the common words being used to describe the soil texture requirement of the crops. The testing dataset loop iterates through 2201 crop scientific and common English names with the three soil properties. The trained word2vec model is then deployed to serve the prediction of similarity. The resultant collections of testing dataset are passed to the word2vec model that was trained on all 178 languages. The output is an array of prediction values from 0 to 1. We compared the values obtained with the data from FAO ECOCROP (http://ecocrop.fao.org) by using optimal soil texture requirement which uses heavy, medium and light levels as an indication. We deduced the heavy level to be those soils which comprise a high percentage of clay, medium level to those with loamy (silty) texture, and light level to those with a high percentage of sand (Food and Agriculture Organization, n.d.).

[Abelmoschus moschatus, “Clay:” 0.10494453, “Loam” 0.13737687, “Sand:” 0.00425547]

The prediction value above represents the relationship between the scientific or common English name and the soil classes. The relationship value is meaningful when used relatively rather than independently. The above example of (Abelmoschus moschatus) indicates that “Loam” soil achieved the highest cosine similarity to the crop relative to the cosine similarity values of “Sand” and “Clay”. Accordingly, we can conclude that the *Loam* soil class is a probable correct classification for Abelmoschus moschatus’ soil requirements. To reiterate, the classification is based on a relative comparison of similarity values; it is impractical to use an individual cosine similarity to directly extract meaning. Finally, we developed an algorithm to verify the prediction automatically against Ecocrop data.

Pseudocode for training and testing process.

```
Data Preparations
    1. By using text mining and regular expressions, Scientific and English common names and Soil
       classes are extracted and stored in a relational database. The full code is available on Github
       aymansalama/text-mining-ecocrop (Salama, 2019)
```



```
# get the family name
grep 'Family' * | sed -e 's/<[^>]*>//g' | sed 's/Family//' | sed 's/\.html\:/\,/' | grep -v ','$ > fam.res
```

Figure 1. Example of a regular expression used in text mining.

```
    2. Creating two test sets:
        ○ For each Soil Class name loop:
            ■ Create a line like (English common name, Soil class)
            ■ Create a line like (Scientific name, Soil class)
Calculating the similarity measure:
    1. Load the vector space of word2vec for all Wikipedia languages to the memory.
    2. For each line in the two sets (Crop name: [English common name, Scientific names], Soil class)
       loop:
        ○ Measure the similarity between the two entities (Crop name, Soil Class)
        ○ Sort all similarities by the highest score
    3. Convert the distance into ranks, for example:
            [Abelmoschus moschatus, “Clay:” 2nd, “Loam” 1st, “Sand:” 3rd]
    4. Validate the predicted class against the recorded soil class from Ecocrop data.
```

RESULTS AND DISCUSSION

This research aims to verify the ability of NIR to capture quantitative data that can be used in different applications, including classification. This will reduce the human resources needed for data collection. We tested the trained model predictions against ECOCROP United Nation Food and Agriculture Organization “FAO”, and the results show an accuracy of 76.11%. The results suggest that NIR can be used effectively in the data collection process. It is important to note that it is not designed to replace human resources in data collection. Rather, it is designed to support human resources in speeding up data collection. The model can read millions of documents and present quantitative data analysis and present it visually to the data collection team to support the direction of the process and enhance it.

The experiment measured two main concepts, i.e., crop name and soil class. A crop name can be presented in the form of a scientific or common name. The results suggest that the scientific name possesses stronger representations than the common English name. The results also confirm the general knowledge that scientific names are more accurately used in the literature. When we measured the accuracy of crop soil classification using common English names, we obtained 49.97% which is far worse than the accuracy gained from using scientific names, i.e., 76.11%. Fig. 2 indicates a sample of ten scientific names and the comparison to soil classifications based on the cosine similarity. The highest value of the cosine similarity represents the correct prediction for the soil classifications.

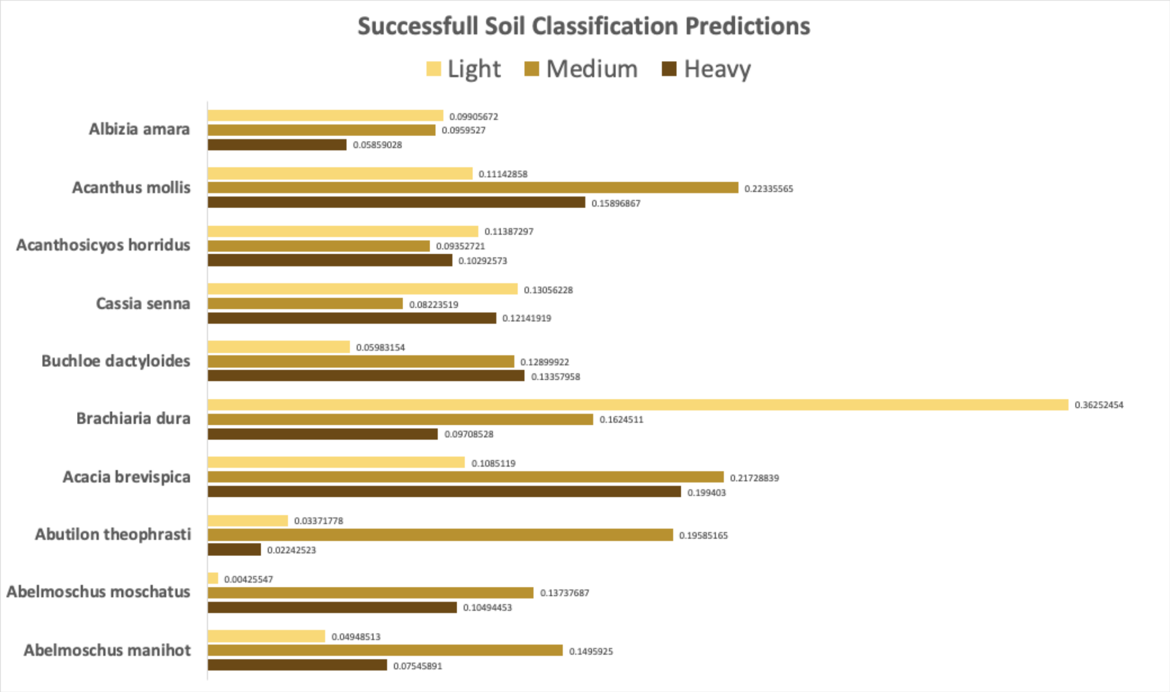


Figure 2. Cosine similarity comparison for soil classification against a sample of ten crops.

The entire data set of crop names consists of 2021 crops. Our word2vec model successfully obtained 1997 crop scientific names out of 2021 names. The crop names that didn't appear in the model vocabulary list are most likely due to fewer occurrences in the corpus. The prediction was successful in obtaining the correct soil class for 1520 crops and failed in 477 crops which corresponds to a 76.11% success rate. Table 1 shows soil requirement predictions on the left and data from FAO ECOCROP on the right. The table shows only nine sample crops and indicates the correct prediction of the model against ECOCROP data.

Table 1. Predication values versus FAO ECOCROP data. The bold values indicate the highest value of the prediction which matches the actual data in FAO ECOCROP.

	Prediction			**FAO ECOCROP Soil Texture Optimal**		
Scientific name	**Clay**	**Loam**	**Sand**	**Heavy**	**Medium**	**Light**
Abelmoschus manihot	0.075	**0.149**	0.049	0	**1**	0
Abelmoschus moschatus	0.104	**0.137**	0.004	0	**1**	0
Abutilon theophrasti	0.022	**0.195**	0.033	0	**1**	0
Acacia brevispica	0.199	**0.217**	0.108	0	**1**	0
Brachiaria dura	0.097	0.162	**0.362**	0	0	**1**
Buchloe dactyloides	**0.133**	0.128	0.059	**1**	0	0
Cassia senna	0.121	0.082	**0.130**	0	0	**1**
Acanthosicyos horridus	0.102	0.093	**0.113**	0	0	**1**
Acanthus mollis	0.158	**0.223**	0.111	0	**1**	0

CONCLUSIONS

In this paper, we studied the effectiveness of the application of NIR in the data collection process. We trained a word2vec model on the entire Wikipedia corpus with 178 different languages. We used the trained model to perform crop soil requirement predictions. We tested the predictions against ECOCROP data for soil classification for 2021 crops. The prediction achieved 76.11% accuracy. However, the selection of the right terminology is crucial in identifying and obtaining the highest accuracy path for prediction. For instance, using common English names instead of scientific crop names leads to a significant decrease in prediction accuracy. The results suggest that we can effectively use NIR in semi-autonomous data collection processes to build structural scientific databases. In the future, we are planning to use the model to predict different use cases for crop classification criteria. Model expansion and enhancement are under investigation to find the best hyperparameter values for the training process. Enlarging the training dataset is an ongoing process in which we are planning to construct a data pipeline for retraining the model and testing its predictions.

REFERENCES

Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., & Robinson, T. (2013). One billion word benchmark for measuring progress in statistical language modeling. arXiv preprint arXiv:1312.3005.

CFF Research. (2019). Retrieved November 5, 2019, from Crops For the Future Research Centre website: http://www.cffresearch.org/Our_Research-@-CFF_Research.aspx#sthash.lFRuRsSU.dpbs

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Food and Agriculture Organization. (n.d.). 6. SOIL TEXTURE. Retrieved November 5, 2019, from http://www.fao.org/tempref/FI/CDrom/FAO_Training/FAO_Training/General/x6706e/x6706e06.htm

Firth, J. R. (1957). Ethnographic analysis and language with reference to Malinowski's views. Man and Culture: an evaluation of the work of Bronislaw Malinowski, 93-118.

Gao, B., & Liu, T. Y. (2017). U.S. Patent Application No. 14/932,652.

Gregory, P. J., Mayes, S., Hui, C. H., Jahanshiri, E., Julkifle, A., Kuppusamy, G., & Azam-Ali, S. N. (2019). Crops For the Future (CFF): an overview of research efforts in the adoption of underutilised species. *Planta*, 1-10.

Harris, Z. S. (1957). Co-occurrence and transformation in linguistic structure. Language, 33(3), 283-340.

Hinton, G. E., Sejnowski, T. J., & Poggio, T. A. (Eds.). (1999). Unsupervised learning: foundations of neural computation. MIT press.

Le, Q., & Mikolov, T. (2014, January). Distributed representations of sentences and documents. In International conference on machine learning (pp. 1188-1196).

Liu, P., Joty, S., & Meng, H. (2015). Fine-grained opinion mining with recurrent neural networks and word embeddings. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (pp. 1433-1443).

Mesnil, G., Mikolov, T., Ranzato, M. A., & Bengio, Y. (2014). Ensemble of generative and discriminative techniques for sentiment analysis of movie reviews. arXiv preprint arXiv:1412.5335.

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

Mikolov, T., Le, Q. V., & Sutskever, I. (2013). Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).

Marahaini, N., Jahanshiri, E., Salama, A., & Tengku Adhwa Syaherah, T. M. S. (2019). Linking data across the value chain of underutilised crops—a multidisciplinary approach. *Food Res*, *3*, 108-116.

Nikfarjam, A., Sarker, A., O'connor, K., Ginn, R., & Gonzalez, G. (2015). Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. Journal of the American Medical Informatics Association, 22(3), 671-681.

Onal, K. D., Zhang, Y., Altingovde, I. S., Rahman, M. M., Karagoz, P., Braylan, A., ... & Angert, A. (2018). Neural information retrieval: At the end of the early years. Information Retrieval Journal, 21(2-3), 111-182.

Ozsoy, M. G. (2016). From word embeddings to item recommendation. arXiv preprint arXiv:1601.01356.

Pennington, J., Socher, R., & Manning, C. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

Ricci F., Rokach L., Shapira B. (2011) Introduction to Recommender Systems Handbook. In: Ricci F., Rokach L., Shapira B., Kantor P. (eds) Recommender Systems Handbook. Springer, Boston, MA

Salama, A. (2019) Text mining for Ecocrop with Regular expressions. https://github.com/aymansalama/text-mining-ecocrop.git

Salama, A., Maul, T., Jahanshiri, E., & Crout, N. (2017). ICBAA2017-4 JOURNEY FROM TEXT MINING TO NEURAL INFORMATION RETRIEVAL IN AGRICULTURAL DATA SCIENCE.

Schnabel, T., Labutov, I., Mimno, D., & Joachims, T. (2015). Evaluation methods for unsupervised word embeddings. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (pp. 298-307).

Weston, J., Bordes, A., Chopra, S., Rush, A. M., van Merriënboer, B., Joulin, A., & Mikolov, T. (2015). Towards ai-complete question answering: A set of prerequisite toy tasks. arXiv preprint arXiv:1502.05698.