Instructions for Codenames game and data

Abstract

This document contains the instructions and definitions of the game Codenames including the basic play and rules of the game and instructions of the collection and analysis of the data. The data is actually a big set of words that are collected from the game players manually. The main concept of this project is to generate an 'Open Mind Common Sense' data from the human brain.
The game is online and enabled to play in this address:
https://code-names-project.firebaseapp.com
The game is available on PC and also in mobile (beta). You can find in the game menu also a guide for the game rules.
The project code can be found in this address: https://github.com/avi326/CodeNames, and the reference to the data can be found in the file 'README.md'. **Also you can find more explanation about the game in 'README.md' and in this guide file: https://github.com/avi326/CodeNames/blob/master/codenames-rules-he.pdf.**

1 Introduction

The following instructions described first the our research ,the main purpose is to collecting information about words relation and the type of the connection between related words probably, but need to remember there is words that might be as not related words but some players will still connect them as to them understanding. And this is the main goal of the research, to creating a big picture from the game players (humans of course) choices and learn more relationships and more meaning about words. In general our research relates to word embeddings, there are a dozens of prior arts written about this topic and also there are some systems that collecting and analyzing the data contains in the system. One of the art is 'ConceptNet 5.5', actually it is an information graph that connects words and phrases of natural language with labeled edges. Its knowledge is collected from many sources that include expert-created resources, crowd-sourcing, and games with a purpose. Our CodeNames project purpose is to generate a big data composed with a list of words and with the related words chosen by the humans brain, in addition the actual output will be analyzed to data in a graph / vectors form for another helpful uses that needs this kind of information and probably help to solve some active issues with this important and 'real' knowledge.

2 Related Work

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics.

The term word embeddings is a very wide and learned in many research and academic places. There are some arts written on this subject and some of them that used here:
- (Bojanowski et al., 2017) mainly talking about limitation but useful of word representations, trained on a large unlabeled corpora for many natural language processing task, but models that learn such representations are ignore the morphology of words, by assigning a distinct vector to each word. In this art there is a propose

of approach based on the skipgram model, where each word is represented as a bag of n-grams character. A vector representation is associated to each n-gram character and a words are representing as the sum of these representations.

-(Joulin et al., 2017) mainly concentrating in the ability of word embeddings to capture both semantic and morphological similarity, as affected by the different types of linguistic properties (surface form, lemma, morphological tag) used to compose the representation of each word.

- And another one is the 'ConceptNet' that are explained in previous section and we can add that 'ConceptNet' is actually a information graph that connects words and phrases of natural language (terms) with labeled, weighted edges (assertions). The original release of ConceptNet was intended as a parsed representation of Open Mind Common Sense, a crowd-sourced knowledge project.

3 Method and Model

The method of our research is based on 'Open Mind Common Sense' concept and probably the best way to do it and to get the accurate as possible results, is to demonstrate a game played by a 'regular' and reasonable persons that will play without knowing the working system behind the scene that learning them and collecting they choices i.e the connected words. With this method the players will play in normally way without any try to affect the research or tilt the results, this is how we can get the 'real' thoughts and the common sense of the humans brain. It should be noted that the game can be implemented in various methods and rules, it is depends in the audience target and what is the purpose of the game, so we create one combination from many optional ones. Our version of the game is a game to two players with some rules will be explained in more details below, in the mode there is still a competition and challenge but the both players are also slightly help to each other to resolve the questions.

This is an example for game board with 25 randomly words with agents and assassins colored on the board.

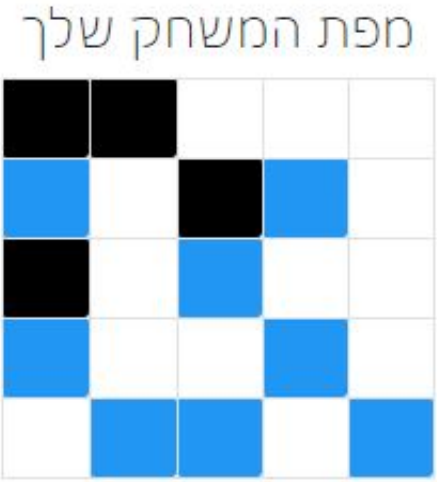
3.1 The Words

As mentioned before the game based on words, the words are very important to the game input and to the output data. The input complex from 350 words taken from the original CodeNames game, the words are basically different in their types of objects e.g egg, sea, palace, country etc. In each game or instance of game (rematch), the words randomly selected from the words database. The small difficulty but sophisticated thing in the randomity is the lack of context sometimes between the selected words, from one hand we can get connections and context between unexpected words that only humans can be find it in a different from computer algorithms, but in other hand it is can create a mistakes if the players will be 'forcibly' connect words or it will lead to zero connections in specific game. The research focus on this type of words but of course there is thousands of words in different categories and type, and there is an option to add / remove words by requirement and the system needs, depend on which learning and research the project designated for.

3.2 The Game - Open Mind Common Sense

The research has been learned by two players game, there is more combination as mentioned before to the game but the purpose is to collect as much as possible data that helping to generate a bigger and accurate output data, that is why it is better for now to this setup instead multi-players game. The board of the game is complex from

25 randomly words (codenames) taken from the words repository in a form of 5-by-5 grid, the board is common and the same for both players. There is 2 types of cells: agents that are marked in blue and assassin marked in black color. Each player have two roles, to challenge the opponent and to be challenged by the opponent, the purpose is to reveal the opponent agents (blue cells). Each player selects one to up eight words from his agents and give the best definition that well described the word, the second player should to solved this by guess the exact word from the board that actually is the opponent agent, this is the success of the challenge, of course this process is vise versa between the the players. Every success will credit the player with some points. Another option in the game is to avoid from the assassin cells, means the players should not select a word that found on the opponent black cells, it is caused to lose the game, but the words connection still saved and increase the data repository.



This is the game map of the board game example above.

4 Data

The data in the research is a different from the ordinary type by that we are collecting the data and not using in specific data, actually the product from the system is a database that contains a big amount of words and the connections between them, those are the words that are saved from all the playing players. The structure of the data is in kind of a vector form, for each word there is a list of words the aggregating from the players guesses. One important point is to be aware to some related words to some word, that can contains a different meaning and context between them, for example the word chicken the related words, the words can be egg and crybaby because the dual meaning or slang of some words. In addition to that, the collected data is in Hebrew so there is another example of different meaning and context and ambiguity, we thing this is can be helpful to deepen the research and understanding of words but need to know how to handle it and use it by the requirements.

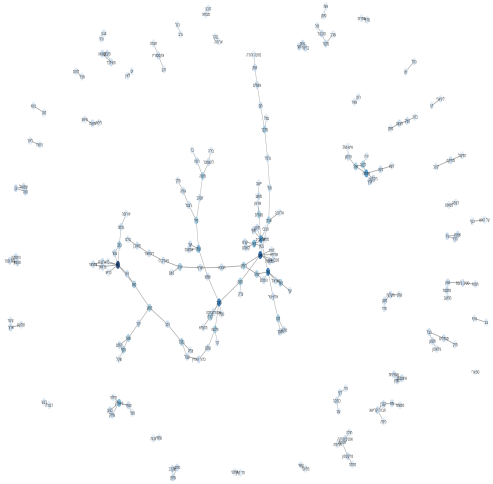
Word	Related Words
Capital	City, Jerusalem, Alcohol, Zion
Gold	Money, Reach, Jerusalem

Table 1: Vectors example.

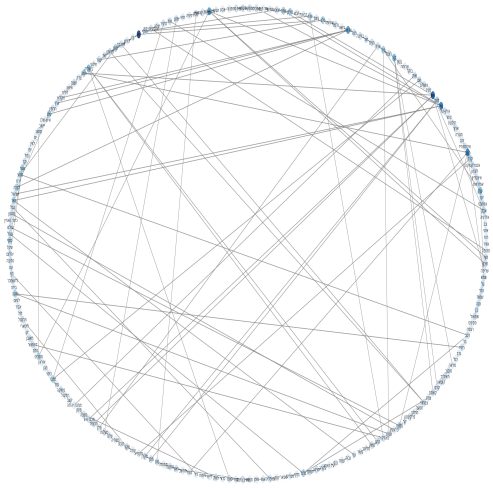
The vectors in the table are small example to the data collected from the players, and the dual meaning that sometimes happens in Hebrew, e.g the vector of the selected word gold contains the words money and reach but also the word Jerusalem because there is a song ”Jerusalem of gold” so it is some connotation to the word.

The data collected from the games can be represented and to be used in various ways, here is an example to the data in graphical way:

In the first example, we can see some of the words from the repository divided and line-connected to words batches. This is a top view to the one of the ways that can be treated with the data. We can figure out chains of related words while moving to other words related, so we can get a long chain of smart connection between words that started from one pair of words and getting to a big number of connections .



Another view is more wide and represents a big picture of the relation between a lot of words, this is kind of meshed network giving the information about a word and the related words:



In the both examples above the arcs connected the words and actually indicates the relation between them, but more important is there is an importance to the arcs that can be present something like a grade or weight of the relation and it can be useful to deeply understanding which words are more related or less, and by this to continue to understand more characteristics of whole text or some NLP general issue, e.g to analyze and understand sentence idea by looking in the related words and the connection between them through the chain. For example the sentence "we going to the library" can produce us words like student, studies, degree etc. and all the topics can be to concluded from this. Also it can be apply to more wide context like to understand the idea of whole text by examine dozen of words from the text in the data and to conclude the related text topics by this process. In the opposite way, using the data, we can to discover also far words, means the words that are not related to some word, it is also be important to tell how some topic is a completely different from another one and to draw all the conclusions from it to any assessment needs.

5 Experiments

In the research the term experiments is a little different from the 'common' term because the research based on the data collected from the players. For the reliability and intelligence, the data collected and saved as is, without any interference or changes so we don't execute any experiments before the data collection but only afterward we executed some experiments and analyzing with the output data to find out word and the related words and the connection between them.

6 Evaluation

In addition to previous section our research does not do the 'common' evaluation of 'grading' the executed experiments, but execute evaluation in compare to another system resource similar to this kind of system. First, I will mention we still need to perform it but the idea of evaluation here is to check the correctness of the data using another system database. For example we can compare some data against 'WordNet Browser (version 2.1)" found at https://wordnet.princeton.edu/ and can be downloaded free. If we take again the word 'gold', 'WordNet' show some values like: coins (made of gold), atomic number, amber and some commentaries. The evaluation we can perform is to check which related words are found and connected also in the research collected data and more sophisticated is to calculate the 'distance' between the unmatched words using another related words in the chain even there are does not connect directly. So we can get a large and wide view on the list of words.

7 Discussion

There is some points it is worth to notice about the data:

- We can find in the data a various different related words for some word, the interesting thing is the large different in the category that characterizes the word, means there is a related words to the same word but in different context.

- The other different point is between the current research and others, the system output and data are varied. The main reason is the most of the data calculated and collected from tagged texts, of course the other system used in very sophisticated method and engine but as long is not the human brain it is probably (for now) will be different until the machine will be more 'smarter' and getting more cleaver and understanding of our brain.

8 Conclusion

In the research we saw a lot of new things, kind of a new world of the big topic of word embeddings, the learning was widely and comprehensive and yet we discovered our research is a small part of big feature. The main conclusion we understood is there is a quite substantial different between working with automatic machines using tagged texts and human bean. We did not focus in parsing tagged texts but in other researches we observed the method and the result of generate connections and context between words it might be less accurate from getting data from real players with a knowledge and common sense.To be more specific the research is examining the Open Mind Common Sense term, this is a challenging and complicated topic, this is why getting a data from tagged texts could be wrong sometimes because machine is making the connection and the context by rules and previous learning, of course it is very clever and success data but still the common sense of human brain is very difficult to replace.
Another conclusion refers to the collected data and to analyze it, it seems there is a couple of ways to represent and using the output data. For every uses and for every problem the data should be collected different as to the system needs and requirements, for example we can generate a scenario to get the directly related words, we can get the opposite words, and the unrelated words etc. This project generating an output collected users data that can be very helpful for resolving NLP problem and to be a basis quite accurate database of a Open Mind Common Sense.
The last conclusion is our understanding there is a option to develop and continue the work on this research with a lot of work.

9 Suggestions for Future Work

During the research we saw there are couple of interesting subjects to research in word embeddings, it is a wide and sophisticated topic that can be useful for resolving problems and to enhancement processes in AI section. In the future we going to focus in the method to apply the project engine and the output data on some NLP problem using the graphical representation, and try to figure our how this research can helping to solve those kind of issues, or what is still missing to make a progress for this goals.
Another future research uses is to wise the machines information due to the collected data, for example there is some text, article or an art, we can determine some key words, analyze and understanding the subject and the main details of the text by using the data, also afterwards it will be more easy to retrieve related subjects and texts, e.g if the system found some key word like 'diet', using the data, there are probably related words like body, self-confidence, weight etc. and so on with those words and more. The conclusion of future work is to enhance the accuracy of the collated data and the analyzing, to get more effective and useful data, to do that there must be very large and reliable data from the users so the goal is to make the game popular and develop more options and features.

10 Bibliography

The 3 main arts need be written (other knowledge was retrieve from others and various resources):

- (Joulin et al., 2017)

- (Bojanowski et al., 2017)

- (speer2017conceptnet)

A Appendices

The analyzing process of the output collected data was executed via Python program. The

file also found in GitHub repository project under 'Analyze Python' folder in main tree address: https://github.com/avi326/CodeNames/ There is 2 same files in py 'regular' Python format and in ipynb to Jupiter notebook