semantic change and colexification with word embeddings and contemporary databases

Yu Chen
ggkevin.chen@mail.utoronto.ca
Undersupervision of Yang Xu
Department of Computer Science,
Cognitive Science Program,
University of Toronto,
yangxu@cs.toronto.edu

Abstract

As time passes, sounds, writings and meanings of languages are constantly changing. scholars have been studying scientific laws and properties of these changes for centuries. With rise of computer calculation powers and development of natural language processing, more decent computational evaluation can be done to investigate and revel the laws and patterns of language change.

1 Introduction

Semantic change is a form of language change caused by the evolution of word usage. Many efforts has been paid by scientists to understand how words change in meaning over time. Some typical laws of semantic change are law of differentiation (which proposes synonyms tend to diverge in meaning) and law of parallel change(which proposed that related words tend to have similar changes). For this project, one tries to discover laws between semantic change and colexification.

A single word form can have multiple meanings in all the natural languages one use. This phenomenon is known as Colexification. Colexification pairs in some concept categories are more frequently observed than other categories. However, the origin of these observations is not well understood. With the new technique of word embeddings from machine learning and larger corss-linguistic database of colexifications, one can investigate further to discover the relation of semantic change and variations of colexifications.

2 Data and prerpossing steps

All codes written by oneself and generated outputs are uploaded on Git Hub. The Git hub link is the following :

https://github.com/NevermoreCY/yuchen2020

However, if one wants to reproduce the results, one still have to download each database from the websites indicated in reference. The 'read me' file explains some of the reference codes and the locations they need to be put under corresponding databases.

Histword embeddings (William L. Hamilton, 2018):

HistWords is a collection of tools and datasets for analyzing language change using word vector embeddings. There are about 30,000 word vectors across 4 languages from 1800s to 1990s. For this project, the pre-trained word embeddings of "All English" are used. Specifically data of decay 1800s and 1990s are used to compare the semantic change across time.

WOLD database(Haspelmath and Tadmor, 2009):

The "WOLD" word database consists of vocabularies contributed by 41 different languages. Which is initially used to compare loanwords across languages. There are 22 concept categories, (eg: "The physical world", "Kinship", "Animals" etc.) Each categories has about 50 150 words(meanings).

IDS database:(Key and Comrie, 2015):

The Intercontinental Dictionary Series (IDS) is a database where lexical material across the languages of the world is organized in such a way that comparisons can be made. Same as "WOLD" 22 concept categories are used. And the IDS contains 1310 meanings in total. The pre-extracted colexification matrices used in the project is borrowed from this paper (Yang Xu, 2020). For each language, a 1081x1081 matrix is used to indicate the colexifications between the filtered 1081 meanings.

CLICS3 database (Rzymski and Tresoldi, 2019):

Clics3 is a database of colexifications (poly-

semies or homophonies) in 3156 language varieties of the world. It's a combination of 30 sub datasets and has 2919 numbers of meanings. It's almost 3 times larger than the size of IDS. CLICS3 uses gml structure to store the data. Which is basically a graph of nodes and edges, where nodes represent meanings, edge between 2 nodes indicates colexification. Other useful Information are stored inside the edge like number of languages, number of occurrence etc...

The set up of CLICS3 database is a little bit complicated. One need to follow the work-flow on the cited websites and one will generate a gml files at the end which contains all the colexification pairs as edges.

GloVe database(Pennington et al., 2014):

GloVe is an unsupervised learning algorithm for obtaining vector representations for words. There are 3 pre-trained word vectors in Glove with source from Wikipedia, Common Crawl and twitter. The Common Crawl data has the most number of words and samples, therefore it has the first priority in this project.

ConceptNet(Speer et al., 2017):

ConceptNet is a multilingual knowledge base, representing words and phrases that people use and the common-sense relationships between them. Basically they provide relationships between pairs of words. They have 34 relations such as "related to", "form of", "synonym", "antonym" etc... The data base supports 304 different languages with 10 core languages with huge vocabulary size, 68 common languages with acceptable vocabulary size.

3 assumptions

For Histword embeddings with IDS database, one predict that meanings that change less in meaning over time in English are also less variable across different languages.

As a further prediction, one also predicts that the colexification across languages have predictability over semantic change.

For GloVe word embeddings and ConeptNet relations, one believe it's possible to build a neural network model using word embeddings and relations as input, predict the relations of conlexification pairs that are not listed by ConceptNet.

Assumption may not always be correct, and some of the prediction are proved to be wrong during the progress of this project. However, combinational data between different database are generated and could be useful for further investigation.

4 methodology and steps

4.1 Histogram with WOLD database:

First, one extracts all the meanings from the WOLD database. Then the meanings are processed into single words so that can be fit with Histoword embeddings. The historical word2vec embeddings for English from 1800s-1990s are used here. Time interval is set to 10 years, so the word embeddings are calculated each decade. Then Degree of semantic change for meanings in cross language database can be calculated using the historical word embeddings.

The method to quantify the degree of semantic change is 100-nearest-neighbors method.(Also used in (Yang Xu, 2015)) For example, for a specific word, one first finds its 100-nearest-neighbors in 1800s and 1990s regardless of position tags.(neighbors can be all the positions like noun, verb and adj...etc). The degree of semantic change equals to 1 minus the overlap ratio between two set of neighbors.

Once the degree of semantic change for each meaning are calculated. One can group them into 22 concept categories pre-defined in "WOLD" database. The histogram for each category is generated. Most of them are gathered and having a Bell curve shape. Most of the words are in the area of 0.5 to 0.9 degree of semantic change. And there are almost no sample locate at area below 0.4.

Figure 1 to Figure 4 are example histograms for some of the categories:

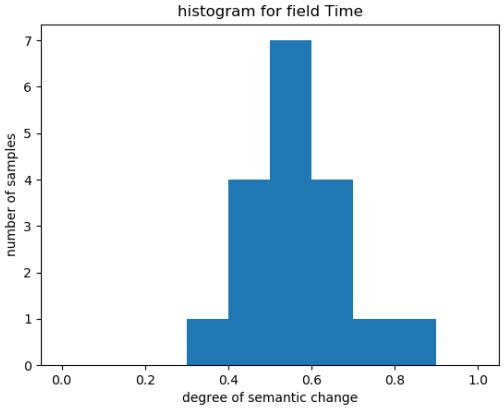

Figure 1: Time

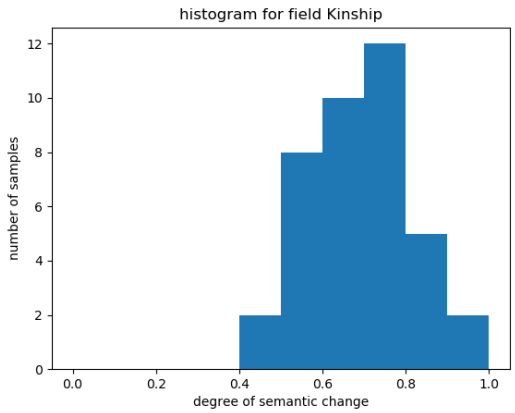

Figure 2: Kinship

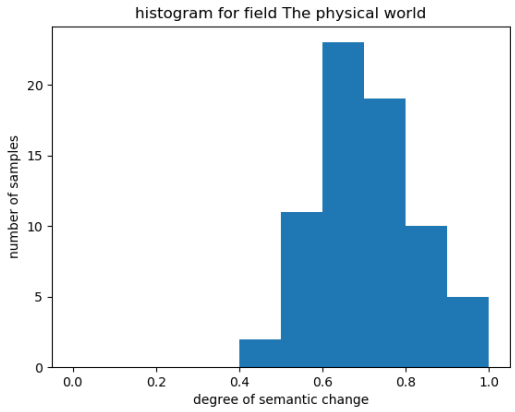

Figure 3: The physical world

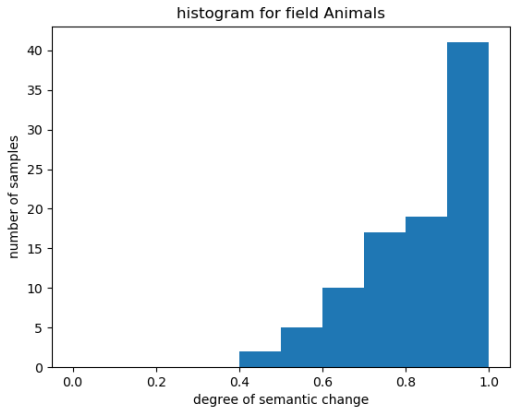

Figure 4: Animals

Then all the domains are juxtaposed into a single plot. Where y-axis is the probability density and x-axis is the degree of semantic change. Kernel density estimation with bandwidth equal to 0.05 is used to smooth the histogram.

Figure 5 is the plot for all the categories:

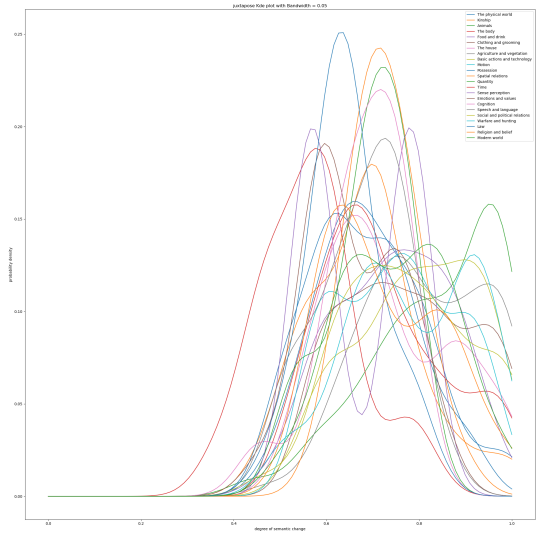

Figure 5: Kde plot for all categories

4.2 Histword embeddings with IDS database:

After processing the WOLD data with histogram embeddings. One move to another database which is the intercontinental dictionary series(IDS). The focus is moved to relationships between histogram embeddings and cross-linguistic colexifications. As the paper (Yang Xu, 2020) indicates, colexification is the phenomenon in which multiple meanings are expressed by single word form. Colexification encompasses cases of homophony,hyperonmy and multiple forms of polysemy. Using the IDS data, one can detect the crosslinguistic colexifiations in forms of polysemy. To save time, the pre-extracted ids colexification data from (Yang Xu, 2020) are used. The link of the code for replications are indicated in the data section.

One uses the colexification data and histogram embeddings to check the hypothesis whether cross-linguistically more variable fields also tend to change more rapidly over time in English. In order to check the hypothesis, the first task is to quantify cross-linguistic variation in colexification, of a domain. In (Yang Xu, 2020) , colexification is defined by a NxN matrix of 0 and 1. N is the number of meanings which is 1310 in IDS. And if an entry (i,j) equals to 1, there is colexification between word i and word j. There are about 300 matrix since each language need one such matrix. Thus, there is a simple way to quantify cross-linguistic variation in colexification. For each semantic domain, for each meaning entry, one calculates the number of other meanings that it is colexified. Sum and average all these numbers within the domain across all the

languages. Eventually, one will have 22 values respectively for 22 categories. The higher the value, the higher degree of colexification domain is.

After calculating the average colexification count, one matches it with the degree of semantic change in Histogram word embeddings.

Figure 6 the plot where x-axis is the chapter id, y-axis is degree of semantic change and colexification count:

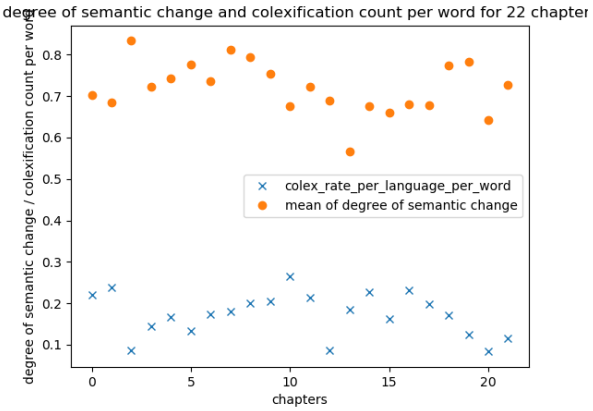

Figure 6: degree of semantic change and colexification counts over 22 categories

It seems there is a positive correlation between the two values. However, since the unit of two values are different. A more precise evaluation need to be done. One combined two set of values together and calculates the Pearson correlation and p-values. As the values mentioned in the result section, the output is not promising. One initially believes that this is caused by the position tags. However, after dividing the data into different position tag, the correlation is still not detected.

Then one move to another direction. One believes that there is still some information from the nearest k neighbors can predict the colexification. The naïve approach is to calculate the average similarity between neighbors and every meaning in each category. One calculate the average similarity between the every words in a given category and the target word's neighbors.

For example for word 'fire' and category "Animals". We can calculate the average similarity between every words in "Animals" and all the neighbors of 'fire'. If there are N words in "Animals" category and M neighbors of "fire". One will calculate N*M numbers of similarities and then pick the average. The hypothesis is that the highest similarity brings the highest possibility of colexification. However, the prediction rate is lower than what one expected. This method predict the highest colexification category for each word with 39% correct rate and second highest category with 12% correct rate. Although it's higher than rolling a 22-faced dice(5%). It's not statistically useful. Especially when the probability to guess the highest category as the word's category has correct rate of 60%.

All these clues are suggesting that there is no huge correlation between the neighbors from English word embeddings and the cross-language colexifications.

4.3 GloVe word embeddings with Clics3:

After working with the previous 2 colexification database, it seems that sparsity of the data might be a potential problem. One moves to a more comprehensive dataset of colexfications which is CLICS3. The number of languages in CLICS3 is 10 times larger than IDS, while the number of meanings is 3 times larger. After transferring the meanings into single English words, there are 2207 meanings left.

One bonus function clics3 provide is a threshold filter for colexifications. So one can filter out the colexiciations onccurs only in n languages.(or language families). One set this threshold t equal to 2 so that colexifications that only occurs in exactly one languages are removed. These are considered to be noisy candidates. The colexification pairs are extracted to csv files and stored in the Github repositories. It's interesting that the top pairs are the concepts(or meanings) that takes less congnitive effort to relate.

Since the CLICS3 data is updated in 2019, a contemporary word embeddings should be used instead of the histoword embeddings. Therefore, pretrained word vectors from Global Vectors for Word Representation(GloVe) is used. There are 3 set of pre-trained embeddings with different sources which are "wikipedia", "common crawl" and "twitter". The similarity between two words vectors is calculated as their Euclidean distance. The closer the distance, the more similar they are. The coleixfication pairs is then divide into two groups. The similar group and the non-similar group. The threshold dividing these group is just the median or mean of all the distances. The outputs are stored in github repository as well. Each source has 3 csv file which are similar, non-similar and all data. Each row of

the file contains two target words, glove euclidean distance and # of colexified languages.

4.4 ConceptNet:

The last database one investigates is the ConceptNet. ConceptNet is a semantic network, designed to help computers understand the meanings of words that people use. There are 34 different type of relations in ConceptNet and common ones are "related to", "antonmy", "synonym" etc... For the colexification pairs found from CLICS3 and Glove, one can find the relations of pairs and how they are explained. The ConceptNet has 300 languages and one is currently using the English data. There are 14368 number of colexification pairs from CLICS3 and Glove(common crawl source). However, not every pair has relations, only 4000+ pairs appears to have relations indicated by the ConceptNet. All the relations are attached for each row of the csv files. The final csv data contains the following: two target words, distance from Glove, # of colexification languages from CLICS3 and relations from ConceptNet.

4.5 Neural network model with ConceptNet and GloVe:

Since the ConceptNet only has relations for 4000 colexification pairs. There are still 10k pairs haven't been classified. One believe it's possible to train a neural network model that learns the mapping between word embeddings of a pair, and the pair's relation category label. Then generalize that model to classify the remaining pairs that were not classified.

The model architecture used is basically a multilayer perceptron. One believes it's a effective and modifiable model for supervised learning. The input data is the concatonation [x1, x2], where x1 x2 are word embeddings from GloVe. The input label is the relations from ConceptNet for the pair x1 and x2. one predict an output that is a n-hot encoding of 34 dimensional vector. The 34 dimensions represent each kind of relation provided by ConceptNet. If the entry equal to 1, it means this relation exist for this pair of words. One will introduce the implementation bellow.

First, K-fold validation is used. The input data is randomly divide into k = 10 groups. And the training will run 10 times while each time one group is selected as validation set and other groups are selected as training set.

Then the input is passed through the multilayer perceptron. The layer size for this multilayer perceptron is a tunable hyperparameter. The layer size used is [input dimension, 500, output dimension]. Which means there are input layer, one hidden layer with 500 hidden unit and one output layer. The activation function for each layer is tanh function. The loss function one used is a L2 loss function. The final loss $L_{reg} = \frac{1}{2}(y-t)^2 + \lambda(\frac{1}{2}w^2)$.

Before the training starts, one randomly initialize the parameters with normal distributions. For each iteration of training, gradient of loss function for parameters will be calculated and current parameters will be updated. For each iteration, one use a small batch set of training data instead of all the data. The stochasticity will reduce the training time and probability to be trapped in local minimum. After all the iterations, one get the optimal parameters. Then it can be used to predict the labels for test data. The output of the test data is not in integer so one need to round them into integer. Since there are K set of result, the majority result(result that are agreed by at least t=5) folds are used. The threshold t can also be tuned.

All the hyperparamters can be tuned in the model are: Learning rate, k(in k-fold), parameter scale, l2 lambda, t, batch size, layer size and training iterations.

The set of hyperparamters one used in the project are: Learning rate = 0.1, k = 10, parameter scale = 0.1, l2 lambda = 0.1, t=4, training iterations = 4000.

5 result and discussion

5.1 Histogram with IDS:

As mentioned in methodology and process, one predicts that concepts that change less in meaning over time in English are also less variable across different languages. By going through the method one mentioned, one get the following result.

5 chapters with lowest semantic change: 'Time', 'Law', 'Emotions and values', 'Possession' and 'Sense perception'.

5 chapters with highest semantic changing: 'Clothing and grooming', 'Warfare and hunting', 'Basic actions and technology', 'Agriculture and vegetation', 'Animals'.

5 chapters with lowest colexification: 'Law', 'Animals', 'Quantity', 'Religion and belief', 'Warfare and hunting'. (While the 6th lowest one is 'Clothing and grooming'.)

5 chapters with highest colexification: 'The physical world', 'Sense perception', 'Cognition', 'Kinship', 'Possession'.

The categories 'Animals','Warfare and hunting' 'Clothing and grooming' are both belong to the highest semantic change group and lowest colexification group. It seems there are some correlations between the two values by the observations. However, when two set of values are grouped, the Pearson coefficient is only 0.1486 and p value is 0.5094. The data is then divided by several different position tag but the values aren't significant. (For noun, coefficient = 0.204 and p value = 0.3616. For adj, coefficient = 0.3679 and p value = 0.0921. For verb, coefficient = -0.1505 and p value = 0.5038.)

5.2 GloVe word embeddings with Clics3 and ConceptNet:

The information gathered from the combination of GloVE. CLICS3 and ConceptNet are stored in the git hub repository as csv files. As mentioned in methodology and process, there are 3 data sources from GloVe, which are 'wiki','twitter' and 'common crawl'. For each data source, 3 output files are generated, which are 'all pairs', 'similar pairs' and 'non-similar pairs'. Every row contains two colexification words(determined by CLICS3), distance from GloVe word embeddings, # of colexified languages from CLICS3 and relations from ConceptNet. And all rows are sorted by distance form short to long.

The mean of distance is used to divide all pairs to two groups. It's simple but seems effective. However, some pairs might be put in the wrong group. One further step in future might be find a better entity that can distinguish the two groups.

5.3 Neural network model with ConceptNet and GloVe:

For the training accuracy and validation accuracy. One used the average number with 10-fold cross validation.
The average training accuracy is 0.9647.
For 10-fold, the list of training accuracy is [0.9648, 0.9644, 0.9646, 0.9646, 0.9649, 0.9637, 0.9650, 0.9654, 0.9650, 0.9650].
The average validation accuracy = 0.9647.
For 10-fold, the list of validation accuracy is [0.9335, 0.9342, 0.9383, 0.9343, 0.9345, 0.9341, 0.9353, 0.9363, 0.9372, 0.9355].

Figure 7 the piechart that summarizes the # colexified pairs against relation types for all the colexified pairs with #languages $>= 2$.

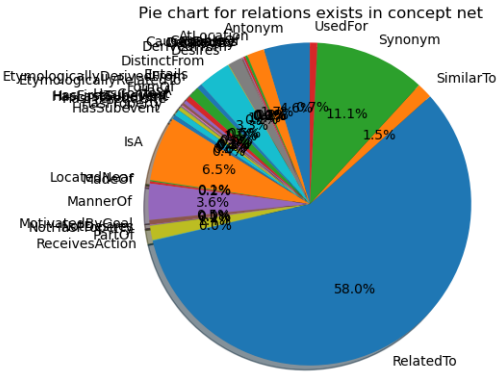

Figure 7: Piechart for all relations exists in conceptnet

Since there are too many relations, it's hard to look at all realtions in 1 plot. And the relation "related to" and "synonym" are taking too much space in the piechart for the first pie chart. One decide to make a new pie chart for top 10 relations without the first 2 ("related to" and "synonym"). Figure 8 shows the top 3 to 10 relations for existing ConceptNet relations while Figure 9 shows the top 3 to 10 relations for model prediction.

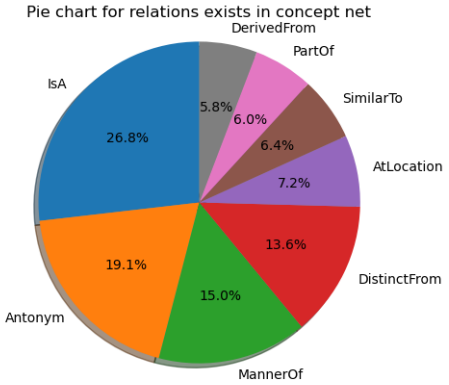

Figure 8: Piechart for top 3 to 10 relations existing in conceptnet

There isn't any huge variation for accuracy among different folds and the validation accuracy seems promising. The top 10 relations for both prediction and existed data are very similar. However, the actual result for prediction are way less precise

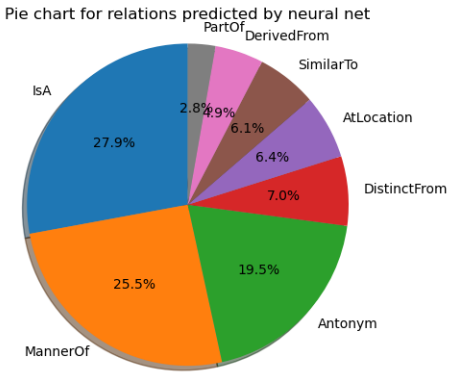

Figure 9: Piechart for top 3 to 10 predicted by model

than the validation accuracy.
Here are random samples for the category 'RelatedTo' and 'Synonym':

For existing ConceptNet data:
'RelatedTo' : ['throw', 'drop'], ['put', 'build'], ['attack', 'ambush'], ['roll', 'twist'], ['family', 'friend'], ['keep', 'defend'], ['person', 'he'], ['bite', 'chew'], ['secret', 'hide'], ['move', 'shake'].
'Synonym': ['fringe', 'border'], ['fever', 'heat'], ['green', 'unripe'], ['cease', 'anchor'], ['age', 'season'], ['wind', 'wrap'], ['separate', 'divorce'], ['breast', 'heart'], ['row', 'paddle'], ['bowl', 'trough']

For the model's prediction:
'RelatedTo': ['grain', 'back'], ['steal', 'banana'], ['stupid', 'donkey'], ['valley', 'hand'], ['seize', 'borrow'], ['imitate', 'teach'], ['peel', 'undress'], ['bay', 'swell'], ['fall', 'drip'], ['bowl', 'gourd']

'Synonym' : ['pain', 'gnaw'], ['young', 'unripe'], ['head', 'hill'], ['throw', 'divide'], ['knife', 'axe'], ['seize', 'split'], ['play', 'shoot'], ['light', 'electricity'], ['imitate', 'weave'], ['vomit', 'pour']

One can observe that the pairs in prediction are way less accurate then the existing ConceptNet data. But by the validation accuracy, there should be about 9 out of 10 pairs have same qualities as the existing data. There must be something wrong with the model. One potential problem might be the lack of training data. There are only about 2000 words and 5000 pairs involved in the training set. All these pairs are colexification pairs and this limit the size of training set. One tired to add more pairs that are not colexified but has relations into the training set. 20,000 words and 50,000 pairs are added into the training set and the csv results are stored in git hub repo. However, the output's accuracy(by looking at random samples) are not dramatically increasing.

The other potential problem is that the model assume that every pair is classifiable according to conceptnet, but in reality there might be many pairs that just cannot be classified. For the rest 10k unclassified pairs, some of them are intrinsically unable to be classified. However, the model has to give a prediction which means there are lots of false negative pairs in the final result. Due to the limit of time, one's progress ends there. A good next step is to somehow find and remove these intrinsically non-classifiable pairs.

References

Martin Haspelmath and Uri Tadmor, editors. 2009. *WOLD*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Mary Ritchie Key and Bernard Comrie, editors. 2015. *IDS*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Christoph Rzymski and Tiago Tresoldi. 2019. The database of cross-linguistic colexifications, reproducible analysis of cross- linguistic polysemies.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge.

Dan Jurafsky William L. Hamilton, Jure Leskovec. 2018. Diachronic word embeddings reveal statistical laws of semantic change.

Barbara C. Maltc Serena Jiangd Mahesh Srinivasane Yang Xu, Khang Duongb. 2020. Conceptual relations predict colexification across languages.

Charles Kemp Yang Xu. 2015. A computational evaluation of two laws of semantic change.

A Appendices

RelatedTo	**Synonym**	**IsA**	**Antonym**	**MannerOf**
'throw', 'drop'	'fringe', 'border'	'hook', 'fishhook'	'road', 'door'	'change', 'cook'
'put', 'build'	'fever', 'heat'	'cloud', 'smoke'	'fire', 'wood'	'cut', 'shear'
'attack', 'ambush'	'green', 'unripe'	'hole', 'nostril'	'forest', 'field'	'separate', 'divorce'
'roll', 'twist'	'cease', 'anchor'	'bunch', 'knot'	'skin', 'bone'	'plant', 'root'
'family', 'friend'	'age', 'season'	'person', 'fish'	'cup', 'plate'	'cover', 'cloak'
'keep', 'defend'	'wind', 'wrap'	'roof', 'thatch'	'money', 'coin'	'arrive', 'land'
'person', 'he'	'separate', 'divorce'	'cut', 'roast'	'throw', 'eat'	'run', 'stream'
'bite', 'chew'	'breast', 'heart'	'woman', 'queen'	'sew', 'cut'	'change', 'die'
'secret', 'hide'	'row', 'paddle'	'seed', 'grain'	'bee', 'fly'	'grow', 'shoot'
'move', 'shake'	'bowl', 'trough'	'insect', 'fly	'hen', 'chicken'	'bury', 'set'

DistinctFrom	**AtLocation**	**SimilarTo**	**PartOf**	**DerivedFrom**
'person', 'animal'	'bridge', 'beam'	'slow', 'lazy'	'foot', 'sole'	'sharp', 'sharpen'
'saucer', 'plate'	'town', 'tree'	'smooth', 'easy'	'body', 'nose'	'cloth', 'clothes'
'smoke', 'steam'	'fish', 'bone'	'shovel', 'spade'	'back', 'spine'	'paper', 'newspaper'
'church', 'mosque'	'field', 'wheat'	'bad', 'severe'	'hand', 'fingernail'	'bark', 'barking'
'leaf', 'branch'	'soup', 'hair'	'stone', 'rock'	'cattle', 'calf'	'ear', 'earring'
'bowl', 'plate'	'fire', 'wood'	'cut', 'split'	'wash', 'rinse'	'daughter', 'stepdaughter'
'tool', 'machine'	'water', 'turtle'	'door', 'window'	'leg', 'thigh'	'long', 'length'
'walk', 'drive'	'field', 'wind'	'hard', 'solid'	'song', 'music'	'slow', 'slowly'
'blood', 'bone'	'table', 'dish'	'hen', 'rooster'	'cattle', 'cow'	'laugh', 'laughter'
'plant', 'flower'	'spit', 'water'	'thick', 'deep'	'face', 'eye'	'cause', 'because'

Table 1: some example colexified pairs for the top ten relations in ConceptNet.