Lexical Representation & Retrieval on Monolingual Interpretative text production

Debasish Sahoo
Kent State University, USA
dsahoo@kent.edu

Dr. Michael Carl
Kent State University, USA
mcarl6@kent.edu

Abstract

Over the past decade, researches in the domain of Language Translation have grown multi-folds. One such area of focus is how the words are encoded, stored and retrieved from memory of individuals who are involved in process of text translation and production. Several models have been developed around this research area, among which Bilingual Interaction Activation (BIA and BIA+) and Multilink are two such popular models with precise hypothesis which can be tested. In this paper, we shall primarily focus to investigate, how the above models assumptions on lexical access (how the words are activated and retrieved in human memory during text production tasks) impact the text production time. Though the above models are designed for bilingual translations, they can also be applied to monolingual tasks. We will limit our experiment to monolingual interpretative text production tasks : Copying, Paraphrasing and Summarizing, in English language only.

1 Introduction

BIA model, in its original form emphasised only on the orthographic representation of the words and its framework was based on a monolingual Interactive Activation Model (IAM). It assumes that during lexical access, words similar in orthography get activated in the mind of the of the user. In case of Bilingual Translation, both the languages are active in the user's memory. Subsequent versions of BIA model (BIA+), took into account the role of phonology (similar sounding) and semantics (similar meaning) during lexical access. Multilink, in addition, assumes that the already activated orthographic neighbors based on the input word activate their associated semantic neighbors, which in turn activate their associated phonetic neighbours and so on. The model explains the observed increase in the word production time by the co-existance of the so many similar words in user's mind. In this paper, we will assess the Multilink hypothesis on the monolingual interpretative text production task. Copying amounts to the most conceivable literal interpretation of a text and thus constitutes a baseline for interpretative text production. Translation can be considered interpretative text production (Gutt, 2010), but other types of monolingual interpretative text production include paraphrasing and summarizing. We operationalize orthographic and semantic similarity by using the measures Orthographic Neighbours (ONS) and Semantic Neighbours (SNS) respectively. Our hypothesis is that the the presence of larger set of such similar words is directly proportional to the word production time.

2 Experiment

For our experiment, we used the *Multiling* dataset from *CriTT TPR-DB* consisting of 6 different English texts. 13 students from the Computer Science department, all proficient in the English language were assigned to perform 3 different tasks - *Copying(C), Paraphrasing(H) and Summarizing(U)*. 9 of these students were native English speakers and 4 of them were Indian students. These tasks were performed on our laboratory computer configured with *EyeTracker (SMI 250mobile) and Keystroke Logger (Translog-II)*. The data was then uploaded

to *Translation Process Research (TPR)* database and aligned. We used Python based libraries (Pandas, Numpy and Matplotlib) on Jupyter Notebooks to execute our experiment.

3 Data

Below are some of the important behavioral data captured in the TPR-DB

SToken represents the source text word token

TGroup represents produced word(s) corresponding to its SToken

Dur provides the word-production time (in ms) for each SToken

HTra provides the Translational Cross Entropy for each SToken

Ins, Del provides the Num of Insertions and Deletions to produce each TGroup for its SToken

4 Orthographic Similarity

According to BIA, *Orthographic Neighbours* are defined by words that differ only by 1 letter. These words look similar to the eyes of the user. According to the BIA, while performing any word production task, the orthographically similar words corresponding to the word being processed, get activated in the user's mind which leads to delay and subsequent longer word production time. In our experiment, we use *Levenshtein Distance (LD)* to find the *Orthographic Neighbours* of our SToken and refer the term Orthographic Neighbours Set (ONS) for the list of such words. The *LD* is a string metric for measuring the difference between two sequences. Informally, the LD between two words is the minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into the other. *ONS* for a SToken is defined as the set of all the words with LD = 1 found in the word-token repository(BNC). For example, ONS for *"killing"*, is {*"willing","filling","billing","milling","tilling" ,"pilling","killings", "skilling"*}.

We used the *British National Corpus (BNC)* as the reference corpus to create word-token repository. The pre-processing steps include tokenizing the corpus to word-tokens, cleaning to remove alpha-nums, numeric, tokens with special characters, grouping the unique word-tokens by its frequency of occurrence in the corpus. Lastly, the tokens are stored as key,value pairs with each word as key and its frequency as value after removing the words with frequency < 10, since they might be typos. We have around 600,000 unique tokens in the repository.

After computing the ONS for all STokens, we used the below measure (SimS1) to calculate the orthographic similarity score for our hypothesis

$$SimS1 = \sum_{s \in ONS} 1 - (lev/len(s))$$

where s = size of ONS, lev = LD and len(s) = length of the word in ONS.

We observe a negative effect (p-value of 0.46) of the ONS on the Word Production Duration. The higher p-value suggests that our results are not significant. These results are not in accordance to the hypothesis laid down in BIA. Hence, we can accept our Null Hypothesis

5 Semantic Similarity

The semantic neighbours of a word is defined as the list of words with similar meaning. We generate the Semantic Neighbours Set (SNS) consisting of semantically similar words using the Word2Vec.

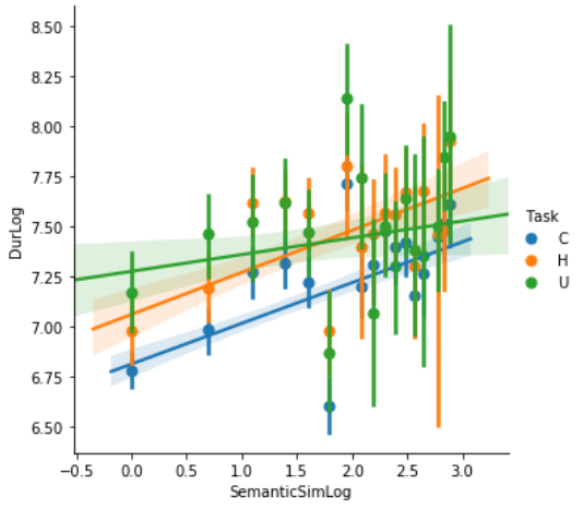
Word2Vec (Word2Vec, 2008a) is a popular word-embedding model, which once trained, can be used to find semantically similar words given an input word. We used a python based framework Gensim and a pre-trained word-embedding model provided by *(Global Vectors for Word Representation)* (glove.6B.100d.zip) to load our Word2Vec model. The Glove model are trained on a corpus containing 6 Billion word tokens, with each word vector represented in 100 dimensions. This model uses cosine similarity to find list of similar words along with its similarity score (SS) - between 0 and 1. We only include words with SS > 0.7 for our test. SNS for a SToken is defined as the set of all the words with SS > 0.7. For e.g. ONS for *"killing"* is {*"murders","slaying","shooting","kidnappings", "executions", "deaths","arrests",*} .

We used the measure *SemanticSim* (the size of the SNS) for our experiment.

We observe a significant positive effect (p value < 0.05) of the SNS as plotted in the figure below. We also observed that Copying and Paraphrasing tasks are more positively correlated than the Summarizing task.

6 Conclusion

From the experiments performed on our data, we observed that while there is no significant impact



of Orthographic Neighbours, we can see a significant correlation of Semantic Neighbours on the Word Production time. With the high p-value (0.46) for orthographic similarity, we can reject the negative correlation as insignificant and thus conclude that we do not see any impact of Orthographic neighbours on the word production time. For Semantic Similarity, we found positive correlation for all the tasks, with Copying task having least average Dur and Summarizing task having a lesser correlation than the other two tasks. We can therefore conclude, our experiment supports the hypothesis that the activation of larger number of semantically similar words may possibly create more ambiguity in the mind of the user to make a suitable choice which eventually may lead to longer word-production time.

7 Future Enhancements

We would like to expand the scope of our experiment to translation data from multiple languages in the future. We would like to test the theory of 'language non-selective lexical access' that is the co-activation of many word candidates from different languages that are similar to the input word. We would also like to train our own model in multiple languages with Word2Vec for our Bilingual experiments.

References

Carl, Bangalore, Schaeffer. 2016. New Directions in Empirical Translation Process Research

Dana M. Basnight-Brown, Ph.D.: Models of Lexical Access and Bilingualism

Ton Dijkstra, Alexander Wahl. (2018) Multilink: A computational model for Bilingual word recognition and translation

Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean. (2013) : Efficient Estimation of Word Representations in Vector Space

Ernst-August Gutt (2010), Ph.D, University of London: Translation and Relevance