Bilkent University

Department of Computer Engineering

Senior Design Project

Etymon: A Deep-Learning Application for Etymological Clustering of Words

Project Specifications

Nashiha Ahmed, Mert İnan, Cholpon Mambetova, Utku Uçkun

Supervisor: Dr. Mehmet Koyutürk
Jury Members: Dr. Uğur Doğrusöz and Dr. Varol Akman

Project Specifications
Oct 9, 2017

This report is submitted to the Department of Computer Engineering of Bilkent University in partial fulfillment of the requirements of the Senior Design Project course CS491/2.

Contents

Project Specifications

Etymon: A Deep-Learning Application for Etymological Clustering of Words

1. Introduction

In this report, a general introduction to Etymon is presented.

Etymon is an analysis and tracing tool for word origins in all languages. It will be used to review current etymological language families and if possible find new connections that were not already present in current taxonomy. It will accomplish this using a deep learning approach. Current etymological analyses rely on pattern matching or tracings between different languages by experts in linguistics [1], yet it may be cumbersome or even improbable to detect word origins in situations where direct links cannot be observed between two different words. In this case, Etymon will pose an advantage as it will be using a large corpus of data in order to match words in any given language.

Various studies carried out by linguistic experts and historians improved the understanding of language and its origins [2]. However, there is still "room for improvement" in the field. Currently, most of the studies target the Proto-Indo-European language family [2]. There is sparse research done for other languages, and there is not a single, unified resource for this information. Most of the information is scattered online or among other forms of literature.

Since there is no similar project in the market yet, our software will be designed from scratch, which would make it a greenfield project. However, we will use other existing algorithms to build our software, such as deep learning algorithms among others that will be specified further in the report.

In the following sections, detailed description of the system, its constraints and requirements are discussed. In addition, issues of professional and ethical importance are also evaluated in order to consider the effect of the system on the society and the scientific community.

1.1. Description

Etymon will be a system that contains three main components: deep-learning, augmented-reality with object recognition, and generative dreaming. Etymon bases most of its functionalities on the etymological map that will be generated by the deep learning algorithm. Etymon system will include a palpable product in the end as an online application based on this etymological map. The system and its implementation will be broken down into three stages in respect to its components.

In the initial stage of the project, a deep-learning algorithm will be used to cluster words from different languages to derive a general map of language families and their relation to each other. This will be done based on word and pronunciation data acquired from Wikipedia. Furthermore, for English, pronunciation data from online sound libraries will also be used, and for other languages, International Phonetic Alphabet (IPA) readings of the words will replace sound files for clustering. At this stage, a crucial consideration will be finding a way to convert words into vectors —this is called "word embedding" in literature [3]— in order to classify them according to their features using machine learning algorithms. At the end of this stage, output from the deep learning algorithm will be a network of words. As the nodes in this graph will be immense in number, instead of showing each word, large clusters of similar words will be sufficient to show language families (Figure 1). This graph will be the etymological map to be used in the consecutive stages of the project and it will be the default graph to be shown when the user wants to search a word.

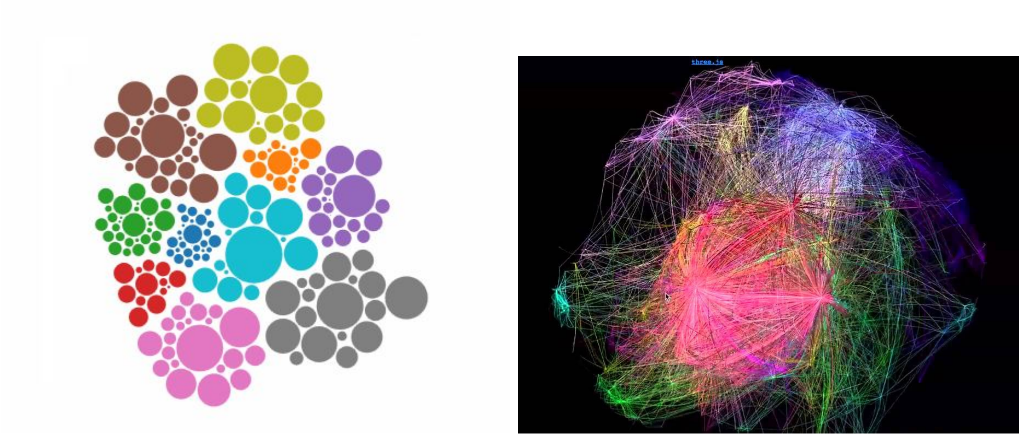

Figure 1: This figure shows two graph visualizations by different softwares that may guide the appearance of the etymological map [4][5].

In the next stage of the project, the etymological map will be used to display specific graphs for words that the user would search in the system. For each word entered by the user, the system will create a smaller and localized graph (Figure 2). This map will show connections between the root of that word and words in different languages. Augmented-reality will be used in this stage as an additional feature. Users will be able to point their cameras to objects and trace their etymological origins. After object recognition, the text will be searched in the etymological map and that map will be overlaid on the object itself using augmented reality.

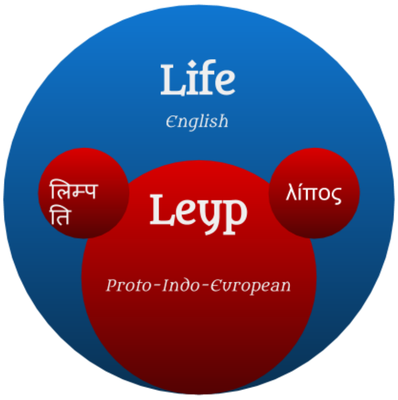

Figure 2: This figure shows a local graph for the English word, "life". Its origin is identified to be "leyp" in the Proto-Indo-European language family, and two descendent words —one in Sanskrit and one in Greek— are given next to the origin word.

As the last stage of the project, the etymological map will be fed into a deep dreaming algorithm in order to generate new words from roots that are present in the language families. This stage will present a creative perspective on the evolution of languages and show alternative formations that languages could have devised. This stage will combine data visualization and deep learning in a visually-appealing manner. When the system is in idle state, it will "hallucinate" these new words, much like that done by a deep learning archiving installation called "Archive Dreaming"[6].

Similar mobile applications such as, "Etymology Explorer" [7], and "Oxford English Etymology" [8] or web services such as "Online Etymology Dictionary" [9] exist in the market yet are largely focused on etymological roots of English words. They also do not analyse words that share the same etymological origin or display languages that exist in the same language families.

The preeminent potential users of Etymon are simply those who are interested in tracing the roots of words. This may range from linguists to anthropologists and historians to the average language learner. Our aim is to provide a tool for the interested to quickly and conveniently visualize word roots and language families. The user can use this program to look up meanings of the words through their etymological origins. For example, one can easily look up the etymological root of his or her name. Also, it can be used as a translator in the sense that one can find words in the other language sharing a common root and similar meanings.

We will use personal testing as well as expert knowledge to validate the accuracy of our program. Different languages will be tested separately to get more representative results. We are planning to use an open-source, pre-trained object recognition algorithm which will require less testing for the augmented reality stage of the project.

All in all, the general description of the Etymon system can be categorized into its three main stages and in the following sections, constraints and requirements of these stages will be supplemented.

1.2. Constraints

In this section, crucial considerations of the Etymon system are listed in detail. Discussing these constraints are valuable to define the details of software implementation.

1.2.1. Implementation Constraints

- The main bottleneck of the system will be the number of words available online in each language.
- The program will require continuous internet connection for word search functionality.
- The search functionality may be slowed down by the increasing size of the etymological map.
- Machine learning algorithms will not run on user devices but on servers to manipulate the large dataset.
- In the augmented reality component, object classification will be restricted by the capabilities of the pre-trained model.
- The amount of words present in each language creates hardware limitations. The system will require additional storage and processing resources.
- Graph visualizations will be done using WebGL and Three.js.

- darkNet [10] and YOLO [11] will be used for object recognition in the augmented reality stage.
- Python, Java, HTML, Swift will be used for implementation.
- ARKit for iPhones, and ARCore for Android will be used for augmented reality components and the online component will use WebAR.
- Word2vec algorithm cannot be used. A new algorithm must be created to cluster words after turning them into vectors [12].

1.2.2. Economic Constraints

- Computing power and database storage may be rented from services like Amazon AWS and Google Cloud Service to run the deep learning algorithm on the dataset.
- The web application will be free to use by all users.
- Open source and free frameworks or machine learning algorithms, such as YOLO, will be used instead of paid products.
- Sound files may need to be bought to include pronunciations in the algorithm in order to avoid copyright infringement.

1.2.3. Ethical Constraints

- The software may cause conflict of interests with linguists and may pose a threat to their expertise.
- Search history of the user will not be saved in order to protect privacy.
- Data entered by the user to train the machine learning model further will be used anonymously and for a short amount of time.

1.2.4. Social Constraints

- Profanity and other inappropriate words may need to be filtered or removed from the dataset.
- Some words that are falsely attributed to different origins as a result of folklore will be shown with their original roots causing inaccuracies.
- Our capacity to evaluate the program’s accuracy in some languages will be limited by the knowledge of the experts in those languages.

1.2.5. Political Constraints

- Countries united within similar language families that are currently under political conflict may give rise to political issues.
- Languages that do not have a highly developed dataset or a dataset that is difficult to find may not be sufficiently represented by the software, which may be viewed as a bias.
- Some words in languages having different dialects might be shown under the same language; therefore, creating political conflicts.

1.3. Professional & Ethical Considerations

Ethical and professional responsibilities of all team members during the onset of this project will include:

1. Giving credit to intellectual property: Each team member and the system itself will be responsible with the data or code that they will use in the project and will give credit in order to protect intellectual property rights of the users or experts in linguistics.
2. Striving to achieve high quality work in both the processes and products of development: All team members are responsible to paying attention to product development. During the implementation phase, commenting, testing and readable code-writing are essential aspects in addition to concise and effective report writing.
3. Evaluating critically the outcomes of machine learning algorithms: The outcomes of the machine learning algorithms will be evaluated according to quantitative error measuring strategies in the literature and the efficiency and correctness of the algorithms will be evaluated critically based on these measures.
4. Understanding and reminding the uncertain foundations of certain machine learning approaches: As machine learning algorithms that this project will include are black-box models, their internal working mechanism is not understood-well. Hence while using these algorithms, precautions on their output should be taken.
5. Evaluating quantitatively the effectiveness of products for human consumption: As the end products of the system include AR components and visualizations, their effectiveness will be evaluated by surveys or tests that will be done in cooperation with the users.

In addition to the code of conduct that every group member affiliated with the project would abide by, we would like to further discuss the ethical implications of the system.

Ethical considerations can be categorized into two main sections: user privacy and anonymity, and use of intellectual property by the machine learning algorithm. As the user data will be used extensively, it will be a crucial consideration to take care of the anonymity of the data input by the user. In addition, as most of the data used by the machine learning algorithm will contain work from experts in linguistics and the users that input the algorithm, it is vital to discuss the credibility of the output.

As aforementioned user privacy will be regarded. Therefore, when a user searches a word or takes or uploads a picture for the augmented reality feature, user information, for instance IP addresses, will not be stored. In addition to this, encryption on user material can also be implemented. This also implies that no attribution will be given to the user for the input data that improves the “learning” algorithm, which may give rise to intellectual property disputes.

Moreover, resources that have already been classified by linguists will be used, which may be considered as a threat to human intellect. This may also give rise to intellectual property disputes, as we will be using material without any direct attribution to the experts in the field, even as we take careful consideration copyright laws when using datasets. Although the purpose of Etymon is to aid users such as linguists, the advancement of the program could lead to a decrease in the need for the expertise of linguists and translators in research on etymology. This could give rise to the classic issue in the field of artificial intelligence of artificial intelligence “taking over” jobs.

2. Requirements

- The system should provide etymological information for every word input by the user.
- Etymon should cluster words according to root features during the deep learning phase.
- The system should show a default language map that contains all the origin languages in the initial screen.
- The system should show word hallucinations while it is idle and no user input are entered.
- The system should "reorder" the map when user searches the etymology of a specific word.
- If a word does not already exist in the map, the system should provide the user with relevant error and help information.
- Etymon should convert words into vectors in order to be used by the machine learning algorithm (word embedding).
- The system should identify closely related languages as language families. It should cluster languages in the same families closer together.
- Etymon should have a simple and straightforward user interface.
- The system should use a deep learning algorithm that is optimal for word clustering.
- Etymon should be easily navigated by the user.
- The system should provide definitions of the words in addition to their origin information.
- Etymon should give reliable output after proper analysis. A user should not get wrong matches.
- The system should provide augmented reality overlays on objects recognized in still pictures and moving camera objects.
- The system should provide fast augmented reality overlays on the objects.
- Etymon should provide etymology of the word in the desired language in the augmented reality feature.
- Etymon should improve etymology maps and clusters when related user and expert data is given.
- Response time is crucial for the software, especially when database enlarges, it should be able to give a user a desired map without taking an extensive time.

3. References

[1] "Introduction", *Etymonline.com*, 2017. [Online]. Available: http://www.etymonline.com/columns/abbr?utm_source=etymonline_footer&utm_medium=link_exchange. [Accessed: 09- Oct- 2017].

[2] "Etymology", *en.wikipedia.org*, 2017. [Online]. Available: https://en.wikipedia.org/wiki/Etymology. [Accessed: 09- Oct- 2017].

[3] "Word Embeddings", *en.wikipedia.org*, 2017. [Online]. Available: https://en.wikipedia.org/wiki/Word_embedding. [Accessed: 09- Oct- 2017].

[4] Nodes, "D3 force layout - How to achieve 3D look of nodes?", *Stackoverflow.com*, 2017. [Online]. Available: https://stackoverflow.com/questions/24673627/d3-force-layout-how-to-achieve-3d-look-of-nodes. [Accessed: 08- Oct- 2017].

[5] "Graph", *I.ytimg.com*, 2017. [Online]. Available: https://i.ytimg.com/vi/qHkjSxbnzAU/maxresdefault.jpg. [Accessed: 08- Oct- 2017].

[6] F. Visnjic, "Archive Dreaming – Building relations and drawing alt-history with machine learning," CreativeApplications. [Online]. Available: http://www.creativeapplications.net/vvvv/archive-dreaming/. [Accessed: 09-Oct-2017]

[7] "Etymology Explorer -", *etymologyexplorer.com*, 2017. [Online]. Available: http://www.etymologyexplorer.com/ety_explore/. [Accessed: 09- Oct- 2017].

[8] "Oxford English Etymology", *Google Play*, 2017. [Online]. Available: https://play.google.com/store/apps/details?id=com.mobisystems.msdict.embedded.wireless.oxford.oxfordenglishetymology&hl=en. [Accessed: 09- Oct- 2017].

[9] "Online Etymology Dictionary", *etymonline.com*, 2001. [Online]. Available: http://www.etymonline.com/. [Accessed: 09- Oct- 2017].

[10] J. Redmon, *YOLO: Real-Time Object Detection*. [Online]. Available: https://pjreddie.com/darknet/yolo/. [Accessed: 09-Oct-2017].

[11] Object-Oriented Software Engineering, Using UML, Patterns, and Java, 2nd Edition, by Bernd Bruegge and Allen H. Dutoit, Prentice-Hall, 2004, ISBN: 0-13-047110-0.

[12] "Word2vec," *Wikipedia*, 26-Sep-2017. [Online]. Available: https://en.wikipedia.org/wiki/Word2vec. [Accessed: 09-Oct-2017].