Illustrating Text

Christopher Catton, Partick Murphy, Charles Ung

December 10, 2015

1 Introduction

A text has always been a complicated thing to translate into another language. Language is not precise and encoding an idea into text is not always perfect. Often, Language has a number of nuances brought in by different culture. Someone from one country may associate luck with red and someone from another country may associate luck with green. But, there still will be an intersection in the features, the visualization, of what any two people assign to a word or even a text in general. If it was not the case then we couldn't translate anything. But, we know that we can translate texts to different languages and we know that multiple texts can describe the same thing.

Often in translation of a text there is some meaning lost. Different cultures view things differently and different languages encode ideas differently. Everyone tends to visualize meaning to some extent though.

In this paper, we take databases of texts retrieved from sources with different languages and cultures. We use the databases, with no prior understanding of the language or cultures, to explore whether we can determine a subset of each text and assign it some meaning by obtaining images for further subsets of the subset generated for the text. In doing so we expect that we can generate meaning for a text and that we can determine whether two texts encode the same thing (i.e. describe the same event or idea).

An interesting, but difficult application for this problem is to generate visual translations. Providing an image for a small set of texts can aide in translation or even language learning tasks, but it is dificult to illustrate certain language elements. Take for example elements such as 'a', 'but', 'or', or even 'context'. There are a lot of elements that may not be easily transcribed into an image, but others such as 'man', 'woman', and 'building' could easily be transcribed. There are also more complex elements that can be described such as the word 'luck' or 'lottery'. There are difficulties in any approach to translate and teaching languages and we believe that generating visual illustrations as we attempt to do is a possible application.

Generating datasets for related problems. Another application is generating datasets for other related problems such as trying to describe a given image. The task has been tried before and this could potentially generate a dataset for training a neural network for doing the task. This probably could be best achieved by building a dataset for training an object recognition network. In the simplest form of the problem we are setting out to achieve could result in providing a set of collages with images keyed to a certain location in each matrix produced from a keyword/phrase set.

There are many facets to the given problem and this paper is organize to partition each step. This paper first discusses the basic method of acheiving results and then details each step. The details are then followed by our conclusions of our exploration. The last section is a discussion on what can be done to improve our results and what modifications can be done to create more complex, interesting, and generally better performing models.

2 Basic Method

The basic method or principle for the problem is to take or generate a language model for each language that is to be used. Then to take a set of texts written in languages that we have models for and filter them on some level to generate a set or sets of keywords for each text. We then take a model that converts the keywords into a set of images with a similar format. Then we perform the very important step of partitioning the images into subsections between the languages. We then arrive at the last step where we

take the set of images and combine them in a meaningful way to illustrate the texts.

In generating the language model for the languages used, we used basic apriori methods to give us relevant data for words and word pairs. Usage of non-aprori methods in way works against what we are attempting to do since non-apriori methods may be encoding meaning of words or word combinations into account, but we are attempting to solve this through the use of images annotated with keywords and phrases. Non-apriori methods may provide a better measure of salience then apriori method, but take away from what we are trying to explore in a finer grained sense. Ideally we should be able to assign meaning to a text without already knowing details of a language as mentioned priori.

The partitioning of images is the task we would say is the most important step. It is where we believe we can take a keyword or phrase, a word part, from one language and assign it a meaning that we can intrinsically understand. Within one language there will be a number of sets that are more descriptive of the nuances of the meanings of the keywords within the language and others less so. We determine how much a set represents the nuances of a language by seeing if it also exists as a set for another language. The intersection between the image sets of two texts acts a way of being able, or a measure, to determine whether or not two texts are two different encodings of the same thing. If the two texts are encoding the same thing, i.e. the measure is high, then logically image sets that measure less similar are still encoding the same material. But, they are also better encodings for their language since they are also incorporating nuances of the language and culture encoding them.

There are more advanced methods for dealing with image sets that should provide much better results and more flexibility in their use. Those methods are discussed at the end of the paper. We felt that usage of simple methods would act as a much better illustration of the problem and its difficulty. Our decision to use simpler methods does come with a number of difficulties in itself that may not be difficulties of the advanced methods discussed.

3 Generating Language Model Dataset(s)

In our study we created our own dataset(s) in order to generate a dataset containing the data we need. We generated the dataset by crawling and scraping separate news sites. These new sites include the British Broadcasting Corporation, the New York Times, De Telegraaf, and De Volkskrant. Our data is tuple of the source, date, and text of an article. Instead of gathering data related to images, we use google image search to generate a set of images related to a set and subsets of the words we extract from a text. In other related studies this task can be done by adding images that appear with a text, but we chose to use google image search to reduce the amount of scraping we need to do to generate our dataset.

4 Generate Language Models

In our study we create language models for languages of articles that we illustrate. We do this using apriori methods and first find frequency for each word and for each set of words with each set being of length N or less. We automatically remove punctuation marks and split on each space and parse each word as part of sets between the size of $1 < N$, with N being the max set size. Each set is constructed with words found by each other. This is the stage we generate frequency overall, and frequency over articles for each word and word pair. This stage should be achieved while parsing each article only once.

In more detail, we select a number of sites that are written in a specific language and parse the articles we have already obtained in generating our dataset. Each Article is then processed to obtain the frequency for each word and phrase of an appropriate size. After each Article is parsed its words and statistics are added to the language model.

We expect the least interesting articles would appear the most frequently across articles in any given language. In-frequent word sets over all articles that are frequent in a small set of articles are likely to be salient for a given article. It is difficult to establish word and word-set salience even with the frequencies. Some words may be infrequent and frequent in articles, but not

really descriptive of the article in general. In using this stage it is necessary to establish cut-off points for salient and non-salient word-sets.

5 Generating Salient Keyword/Phrase Sets

Generating salient sets is fairly easy compared with establishing a language model. To get a set of salient-words for a given article we provide the statistics for a given article and take the language model for the article. We use the language model to remove non-salient words from the article and then use some combination of statistics of the word for the article and in the language. It is hard to determine a good combination of the statistics that represent the salience of a given word or word-set. Determining a statistic that represents good salience for a word or word-set is a problem on its own.

In this project we choose a the combination:

$$1 - frequencyInArticle(Word)/numberOfWordsInArticle$$

to represent the salience of a word or word-set.

We also use two thresholds to remove non-salient words. One is the percent of articles that a word or word-set appears in and the other is how many occurences does the word appear overall.

6 Generating Image Sets

In our project we require a large number of images with similar annotations of keywords for each language used. The keywords for certain images may differ for each image, but there should be an intersection among the images for a select set of keywords. Since this is difficult to obtain on our own we, in our project, turned to google to obtain sets of images for keywords.

In generating a number of seperate sets for each set of salient keyword/phrase sets, we hope to be able to determine whether there are any sets, generated by different texts, that are similar enough to be considered the same. If they are similar enough it follows that the texts are encoding the same material and we can possibly say that if the articles are in different languages then they are approximate translations of each other.

After downloading images from the Google search image API of salient word sets from the article, the common subset of images within these sets must be found. This can be done by finding the intersection of all the sets. In our implementation, we first created subset that was the intersection of two sets and then iteratively performed the intersection of this subset with each other set of images. Since the number of images is relatively large, all the images in two sets could not be loaded into memory at once. To deal with this issue, the intersection algorithm used was a block-based intersection which loads two blocks of 50 images into memory, then compares each image in 1 block to each image in the other block and output images which are in both blocks. To compare the images two different comparison tests were used. For two images to be considered the same, they must pass either or both of the comparison tests. The first test was a simple pixel by pixel comparison, if for each pixel in each image the values of the pixels were equal then the two images are the same. The second comparison test was to take the average colour of all the pixels in a different square regions of the images and compare them. If the average colours in each of the regions were the same, then the images are considered the same.

7 Generating Illustrations

The last and final step in our project is how we visualize the image-sets and the keyword salience. Ideally we would be able to combine the images into a single images, but that is a difficult task even if we know what features and feature-groups a keyword or keyword-set relates to. Instead we opted for creating collage of the images within a set. This means that we would generate something similar to what one might expect an output of a model for object segmentation. There are no masks of objects, but we might expect that certain positions in a matrix to be occupied by a certain object for a salient set.

In arranging images in a collage we get around the problem of trying to combine them and possibly provide something in way of a story arc. It should show more meaningful visualization of the article at the origin of the matrix then in other places, provided that the salience measure is correct. It is also fairly culturally ambiguous since a lot of cultures have come to expect items to start in the top-left in recent times. But, there are more culturally

ambiguous ways to create illustrations as well (which are more difficult).

8 Conclusion and Problems

One problem we encountered was that the Google image seach API only allows for 100 searches per day for free, so gathering enough images for each of the salient sets was impossible. This in turn has made it impossible to get the results we would consider significant. To obtain the results we want we will need access to a database that we can obtain a much larger set of images for each salient set. This may be achieved by data-mining for images based on the salient sets themselves. Though there are possible ways of reducing the requirement of a large amount of images which are discussed later.

Another problem is validation of results. Since we were unable to obtain the results needed, it is not relevant that we cannot evaluate them. But, it needs to be discussed how we would go about evaluating the results had we obtained them. For this task we probably we need to crowd-source in order to have people evaluate the accuracy of the results. It isn't out of ordinary to need actual people to be involved the actual evaluation of the results. In computer vision a lot of data for evaluating models for generating salience or object segmentation have manually generated data from people. For us, we would not need any special tools such as eye-tracking tools to determine the accuracy though.

9 Future Work

9.1 Datasets

Future work needs access to a larger set of images annotated with keywords and phrases. It may be necessary to mine for the images and annotations specifically or to crowd source the annotations of mined images. But, it is achievable and may provide the results we wished to achieve. One issue we found with google is that the search, provided with two words that are nearly identical in meaning, does not provide nearly identical results. Ideally images would have similar ranking in a search for words with a similar meaning. We went in expecting there to be a large reuse of images between people of

different language groups, but our observations now point us in a different direction.

9.2 Language Model

A possible extension of how we generate a language model is to take into account the time each article is written and published. We expect that certain words are more likely to be published during certain time periods of the year such as during holidays (e.g. Christmas, Thanksgiving). Also, articles related to a crisis would likely be published during a certain span of time and possibly articles would be published in rememberance of them (e.g. 911).

In many languages their is a tendency for certain word-parts to take a certain order which may vary for a number of reasons such as whether a phrase is a question or statement. In later iterations of our language model, we would take this into account and try to determine whether a language has such an order and when or whether it changes during certain contexts. We can determine whether there is a word order by determining if two or more words appear along-side each other relatively frequently.

Another concern about word-parts is that certain parts, such as nouns, may be considered more salient for illustration. It may be possible to determine whether a particular word is a noun, verb, etc. by considering their frequency along-side closely occuring words and their order. However, this is a difficult task on its own.

Another consideration is that some languages compound words together or have prefixes and suffixes attached to words in certain contexts. In order for us to determine whether a word occurs within a compound word would be to check whether it appears alone, it's frequency overall (the word 'a' is likely to appear in a lot of words, but it is not a subword), and its frequency as a subword.

The task of getting relevant salient sets of words isn't necessarily resrticed to apriori methods or even traditional data-mining techniques. In a way using non-apriori methods may detract from ascertaining the meaning from

the image sets we find. But, their results could still be considered interesting and relevant. It would be an interesting task to design a neural network to determine word salience or even to get it to create the illustrations. Some machine learning and a training set to determine thresholds for what is and is not salient could go a long ways. The thresholds would probably maintain relevance across different methods of solving the problem, but right now we are only guessing and manually adjusting them.

9.3 Image Analysis and Partitioning of Image Sets

A possible extension of this stage would be to incorporate machine learning and computer vision to be able to find features and feature-groups within an image. By using features and feature-groups, we may be able to find which features or feature-groups relate to each keyword or keyword-set and pick images based on this. This step is already improved if we are able to match multiple modifications of a single image and throwing away the modified versions. Looking at an image on a feature basis allows us to generate a better set of images that represent the keywords better and means that we possibly won't need to have as large of a database of images with similar keyword/phrase annotations.

Learning to relate features and feature-groups to keywords and keyphrases can be done independently for each language. It may be possible to completely get around a need for a single database of images with keyword/phrase annotations for all languages. This would be the best modification to our project and product the best results. One of the biggest problems is the database for the images.

9.4 Illustration

The step of creating the illustration can be modified in a number of ways. The best method likely varies based on what you want to use the results for. For language learning and depending on the text size, it is probably best to either find a single image that is representitive of the salient set it represents or to be able to combine the images or features related to the keywords into an image. In some cases it probably would be best to have a series of images to represent a meaning. Illustrating certain actions would probably require multiple images to be accuratley illustrated.