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).

Providing translations

Language learning

Generating datasetss for related problems

Organization of Paper
–Related Work
–Basic Method
–Detailed Discussion of Each Step
–Conclusions
–Future Work

–Modification by each step

2 Related Work

Related Work – Computer Vision

3 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.

4 Generating 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.

4.1 Generate Language Models

We include a number of steps for generating language models that each build on each other. This allows us to create illustrations at different steps of computation and view the effects of different algorithms.

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 include a settings parameter to allow for the inclusion of splits on choice punctuation marks. In this 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.

-Frequency Overall is the sum of the frequency of a given word set in each article for a given language -Frequency Over Articles is the number of articles a word appears in for a given language

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.

4.1.1 L1, Possible Extensions

A possible extension of L1 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, etc..). 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).

4.1.2 L2, Sub-words and Word Order

In many words 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 the L2 stage we 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. But, 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.

4.2 Generating Salient 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 L1 statistics for a given article and take the language model for the article. We use the article to remove non-salient words from the article and then use a combination of statistics of the word for the article and in the language.

4.3 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.

4.4 Generating Relevant Image Set

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 implemnetation, 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.

4.4.1 I1

In I1 we obtain a set of images based on a set of keywords. Among the image sets we would assume that we would be able to generate an intersection of image sets that would be representative of two articles discussing the same thing. Some articles, not in the intersection would be better representative of a given article and even their language. Possibly, we would be able to even find images or sets of images that would represent a difference in how speakers of two different languages view the topic or even the words and phrases themselves.

4.5 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 occurs too.

5 Conclusion and Problems

– better datasets needed

5.1 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.

6 Future Work

– better datasets (particularly for images) – more complex analysis of images – break down of images into features and feature groups

6.1 Modelling a Language

– use non-apriori methods – user generated keywords and machine learning – more complex apriori methods

6.2 Image Dataset

– Create custom dataset by crowd-sourcing or move intensive data-mining for reuse of images

6.3 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. 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.

6.4 Illustration

– Composite Images based on keywords – Exploration of culturally ambiguous collages (Left-to-Right Right-to-Left) – Image Modification based Salience and Computer Vision