Connecting Cultures: A Vector Approach

Reid Kay
MIDS W266
reidkay@berkeley.edu

Abstract:
Advances in computing technology and model architecture have prompted new ways of representing words for use in NLP applications. These new encoding methods allow the word representations to contain semantic and syntactic information about the words they represent. This paper uses these word embeddings as a vehicle to study how the understanding of various abstract concepts are related across cultures based on their usage in mythologies, folk tales, philosophies, and religious documents. Part I of this paper presents the motivation for the work. Part II describes existing work in the field upon which this project is built. Part III describes the methods and procedures used in this project. Part IV presents results and findings. Part V concludes the paper with recommendations for further research along the lines of this project.

Part I. Introduction & Motivation:

Technology has truly made the world more accessible. Where a century ago most people did not have the opportunity to interact with others outside of their own community due to distance constraints, we can now send messages around the world at nearly the speed of light. Technologies such as pocket translators and Google Goggles are lessening the communication constraints presented by the multitude of languages on the planet. It is very exciting to have these tools at our disposal so that we may work with and begin to better understand the people and cultures that make up the human race.

Cross-cultural understanding, however, requires more than simply mapping the words of one language to another. There often exist phrases and concepts in one language for which this is no exact (literal or figurative) corresponding representation in another. Even for individuals who share a common language, each individual will have a unique conception of a word's meaning given through his or her unique experience with language and culture – particularly for abstract concepts. Although we may have made tremendous strides in enabling ourselves be heard by others, we should take a moment to understand how words may carry different connotations across cultures based on their usage in the stories and parables that shape the minds of their people. By understanding this diversity of conceptions, we may better understand the reasons for cross-cultural roadblocks. By better recognizing when words and concepts invoke similar associations across cultures we may more easily find common grounds upon which to build a shared understanding.

This project studies a certain set of key word/concepts (both concrete and abstract) across corpora from selected religions, philosophies, and folk tales. For each corpus, a list of words similar to the key word is presented to provide a context of how each culture conceives the notion of that key word. Also put forth is a novel method for comparing the similarity of notions across corpora.

Part II. Background:

In traditional natural language processing methods, words are represented simply as discrete features produced though the one-hot encoding of the documents and corpora in which they are contained. Through analyzing co-occurrence counts, N-gram models can be used to predict subsequent words, given a context window, and provide some insight into how words are used. These models however, provide little means for comparing non co-occurring words and can suffer from sparsity complications. Even for words that do co-occur within the context, we learn little about the nature of their relationship. Comparisons of the similarity of corpora have focused on analyzing the relative frequency of words and n-grams through frequency counts and chi-squared tests [1, 2]. Although these measures have shown some successes in understanding corpora similarities, the information provided does little to show how the conceptions of specific words are related across corpora [3].

Recent advances in computing technology and model architecture have prompted new ways of representing words for use in NLP applications. Neural network language models provide a great improvement in this area. As a beneficial byproduct of training a neural network for word prediction tasks, continuous, dense word vectors are learned. The geometric relationships between these learned word vectors have been shown to exhibit interesting properties regarding the semantic meaning and syntactic usage of words to which they refer. Notably, the degree of similarity between two words can be represented by the cosine distance between two resulting word vectors. Extending this geometric representation through an application to analogy resolution are findings that vec('King') – vec('Man') + vec('Woman') can yield a result similar to vec('Queen') [4].

The vectors produced during the neural network training process are all of the same user-specified magnitude of dimensionality. As a result of the learning process, the axes associated with the dimensional space of the vectors begin to take on certain aspects of word meaning and usage from the corpus upon which the model was trained. Comparisons of equal length word vectors trained from different corpora cannot be done directly as the vectors come from different vector spaces without any guarantee of dimension alignment. To make meaningful comparisons of word vectors across corpora, the word vectors would need to be learned either all within one model, or would need to be mapped to a common vector space through a linear projection of one vector space to another [5, 6].

Part III. Methods:

Key Words

This project studies the relationships for a set of 'key words'. These key words were chosen as representative of emotional and metaphysical concepts about which various cultures may conceive differently. The list includes words like 'good', 'bad', 'faith', 'knowledge', and 'happiness'. The full list may be found in Appendix I.

Corpora Construction

Important and influential works of writing (for which appropriate English translations could be found) from 6 cultures were identified. These cultures are Buddhism, Chinese (Taoism & Confusism), Christianity, Hinduism, Judaism, and Islam. The full list of these documents can be found in Appendix II. These documents were organized into distinct corpora based on the culture from which they originate as well as into a 'Cosmopolitan Corpus' containing all of the documents.

Text Pre-processing

Basic pre-processing is performed to attempt to remove text matter that is not intrinsic to the text itself. Many of the documents had extraneous heading and footing material. Attempts to remove these both programmatically and manually were undertaken. The remaining text documents were tokenized such that punctuation is removed and all words made lowercase. Tokens then had basic lemmatization (NLTK's WordNetLemmatizer) applied. Tokens that do not fall within the top 40% of tokens in terms of frequency were replaced with 'UNK'.

Word Embeddings

For purposes of learning word vectors, the skip-gram method proposed by Mikolov et al. [7] is used. This method has the advantages of both producing word vectors that perform well on word similarity tests and being (relatively) fast to train [8]. Given the limited size of the corpora used for training, pre-trained word embedding vectors were used for vector initializations whenever possible. These pre-trained word vectors were trained on a portion of the Google News dataset. The resulting vector archive contains '300-dimensional vectors for 3 million words and phrases' [9]. For words in the corpora vocabulary for which there does not exist a pre-trained embedding, random initialization is used. Similarly, key words are initialized with random initializations so that the ultimate similarities of key word embeddings are the results of the contexts and the model rather than any pre-trained embedding that may exist for that word.

Inter and Intra Corpus Relationships

Within each culture's corpus, word2vec was run for 100 steps with a window size of 6 and 2 skips. From the resultant word embeddings, the 5 nearest neighbors to each key word are calculated as determined by cosine similarity. Although all word vectors are of the same dimensionality, because these corpora do not share a common vector space, direct comparisons of vectors across corpora are not possible.

Next, for the Cosmopolitan Corpus, all instances of key words are all appended with a tag corresponding to the culture from which they came. This tag allows each culture's keywords to be represented as distinct words within a unified vector space. Upon the appended Cosmopolitan Corpus, word2vec is again run for 100 steps with a window size of 6 and 2 skips. As the word embeddings now share a common vector space, we can examine how the different cultures' usages of the key words compare within the context of a shared language vector space. To accomplish this, for each key word, a distance matrix is produced which shows the cosine distance between word vectors of the key words for each culture.

Code for parts of this project may be found at the following location:

https://github.com/ReidKay/w266_project/blob/master/nearest_neighbors_across_cultures.ipynb

Part IV. Results:

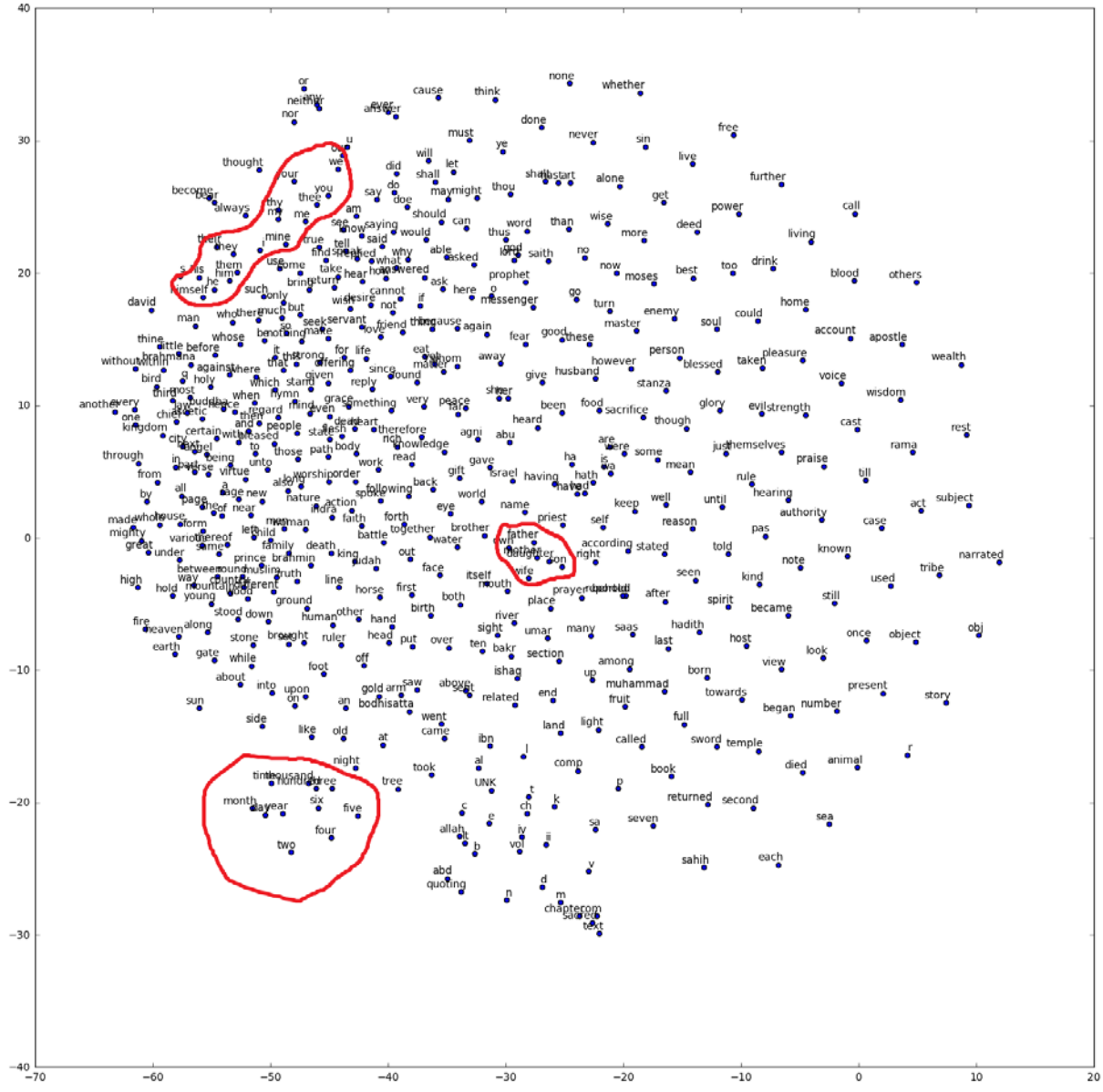
Nearest neighbors for the key word "love" for each culture's corpus are displayed below. Full results for the intra-corpus nearest neighbors analysis can be found in Appendix III.

Buddhist: husband, talk, sinful, fair, bear
Chinese: revere, benevolence, the, senseless, pickle
Christian: faith, prevail, demonstrative, show, conscience
Hindu: stolen, destined, turvayana, happen, install

Judaic: shined, deem, guilty, inclination, count
Muslim: realisation, calumny, accompany, khaythama, lix

Although each culture's list contains elements which can be understood as exhibiting a relationship to the key word, "love", it can be seen that the various cultures focus on different connotations of the word. This provides evidence for the motivation of this project.

The learned word embeddings have been projected onto a 2-dimensional space (determined via Principle Component Analysis) for visualization purposes. The projection for the (non-appended) Cosmopolitan Corpus can be found below. Additional visualizations for the other corpora can be found in Appendix IV.



Even on such a small sized training corpus, evidence of word2vec's embedding of concepts can be seen. Three clusters are identified above. One contains numbers and a description of time (two, four, thousand,

month, year, etc.). Another contains familial relationships (wife, son, mother, daughter, father). A third contains pronouns (you, me, my, them, him, etc.). Other such clusters can be found.

The next step is to quantify the degree to which the diverse cultures of the world have similar understandings of the key word concepts as expressed through syntax and semantics. This is represented by the cosine similarity between the unique tokens representing key words across the cultures in the Cosmopolitan Corpus. Presented below is the similarity matrix for the key word 'friendship'. The full similarity matrices for each key word may be found in Appendix V.

Similarity Matrix for friendship

	Christian	Hindu	Judaic	Muslim	Buddhist	Chinese
Christian	1.000000	0.616650	0.564612	0.500147	0.168903	0.642962
Hindu	0.616650	1.000000	0.597061	0.612654	0.238321	0.642327
Judaic	0.564612	0.597061	1.000000	0.458355	0.223050	0.622187
Muslim	0.500147	0.612654	0.458355	1.000000	0.210364	0.540165
Buddhist	0.168903	0.238321	0.223050	0.210364	1.000000	0.244712
Chinese	0.642962	0.642327	0.622187	0.540165	0.244712	1.000000

The ultimate goal of the project is to identify topics with high levels of cross cultural similarity that may be used to highlight areas of pairwise shared understanding. Presented below are the key words with the highest cosine similarity for each culture pair.

Most Similar Word Matrix:

	Christian	Hindu	Judaic	Muslim	Buddhist	Chinese
Christian	N/A	destroy	destroy	pray	soul	soul
Hindu	destroy	N/A	bad	worship	worship	friendship
Judaic	destroy	bad	N/A	mercy	rejoice	wisdom
Muslim	pray	worship	mercy	N/A	happy	knowledge
Buddhist	soul	worship	rejoice	happy	N/A	soul
Chinese	soul	friendship	wisdom	knowledge	soul	N/A

Finally, for each key word, a vector representing the centroid of various cultures' embeddings of that word is produced. The Euclidean distance from each culture's embedding vector to the centroid vector is taken. Summing these distances provides a measure of the degree of universality in understanding (as described by syntax and semantics within the texts) for the given keyword – with lower numbers denoting a smaller divergence in conception. Ordering these sums in ascending order provides insight into words with the most overall homogeneity across cultures. The top 5 most universal keywords are presented below. The full list can be found in Appendix V.

	score
soul	3.32634
knowledge	3.70695
destroy	3.75873
worship	3.87242
friendship	3.96047

Part V. Conclusion and Next Steps:

It is hoped that this novel approach and application of word embeddings provides impact in the real world. All too often conversations are undertaken, but due to different understandings of the concepts being expressed, a true understanding is never reached. Hopefully this project has provided a means to transcend this babel.

It should be noted that the methods outlined above do have their limitations that could hopefully be addressed (in time) with more researchers and resources. These limitations include the following:

- The selection of key words was chosen by Your Humble Researcher. To share a common vector space, all words other than the keywords were collapsed across cultures into a common token. This creates the implicit assumption that the usage and understanding of these non-keywords, which create the context for the key words' usages, is shared across cultures.
- The corpora size used in this project are quite small. It would be desirable to have corpora that are orders of magnitude larger than those used here. The reasons for this are twofold:
    1. Such small corpora do a poor job training word embeddings. The use of pre-trained word vectors somewhat helps to ameliorate this problem, but the use of such small corpora limits the amount of information learnable by the word2vec algorithm. Efforts we made to produce corpora of similar size, but differences in the frequency of the key words across corpora may present complications given the small size of the corpora.
    2. The works chosen to include in the corpora were chosen by Your Humble Researcher. He is neither an anthropologist, nor an expert in comparative religion. The richness of cultures and religions cannot be fully captured in small numbers of documents, but rather are shaped over time and countless expressions. Significant time was devoted to identifying a small set of appropriate documents for use. Although this was interesting work and could have continued for ages, this was not intended to be the focus of this project so the determination to proceed with the identified set of documents was made. Similarly, with more time, more cultures could be included in the analysis.
- Basic lemmatization was applied to the tokens. In particular, NLTK's WordNetLemmatizer was used with default settings. Alternate pre-processing methods yield greatly different results. For instance, the words 'died and 'dead' return different tokens after lemmatization. If both of these tokens were collapsed into a single token, the learned embeddings would be quite different than they are. This would result in different cosine similarities, nearest neighbors, etc. A compelling case could be made for each pre-processing method depending on the level of word specificity desired.
- The documents used in this study had all been translated into English. As such, the works are not simply a reflection and inspiration of the culture that produced them, but also of the culture and creativity (and potentially foreign language teacher) of the translator.
- When calculating the 'most universal' key words, the determination to use the sum of the Euclidean distances to the centroid was used. Using the centroid was not without its shortcomings. Alternate avenues for exploration include finding a vector that minimizes the

sum of the angles between itself and vectors representing the various cultures usages of keywords. This is an interesting subject for future study.

The limitations described above are far from insurmountable. Corpora can be enlarged, additional cultures considered, varying key words selected, and alternate lemmatization procedures implemented by those interested in continuing this endeavor. It is hoped that this work has provided insight and inspiration, as well as a nearly plug-and-play programmatic framework, which can be used for future research.

References:

[1] Kilgarriff & Rose; Measures for Corpus Similarity and Homogeneity

[2] Eckart & Quasthoff; Statistical Corpus and Language Comparison using Comparable Corpora

[3] Babych & Hartley; Meta-Evaluation of Comparability Metrics Using Parallel Corpora

[4] Mikolov, Yih, and Zweig; Linguistic Regularities in Continuous Space Word Representations

[5] Tsvetkov et al.; Evaluation of Word Vector Representations by Subspace Alignment

[6] Tan et al.; Lexical Comparison Between Wikipedia and Twitter Corpora by Using Word Embeddings

[7] Mikolov et al.; Distributed Representations of Words and Phrases and their Compositionality

[8] Mikolov et al.; Efficient Estimation of Word Representations in Vector Space

[9] Google Word2vec Project https://code.google.com/archive/p/word2vec/

[10] Wisdomlib.org; Definition of Vamadeva

Appendix I – List of Keywords

good
bad
heaven
hell
man
king
evil
holy
peace
salvation
eternal
time
destroy
pray
god
faith
death
birth
life
war
child
sin
body

mercy
love
hate
soul
justice
joy
suffer
wisdom
worship
truth
knowledge
friend
friendship
desire
suffering
happy
happiness
rejoice

Appendix II – Corpora Contents

- Buddhist Texts
  - Amitabha
  - Dhammapada
  - Digha Nikaya
  - Jataka
  - Lotus Sutra
- Chinese Texts
  - The Art of War
  - Book of Poetry
  - Chinese Buddhism
  - Confucius – Analects
  - Confucius – The Doctrine of the Mean
  - Confucius – The Great Learning
  - Dao De Jing
  - I Ching
  - Li Jing
  - Mencius
  - Myths and Legends of China
  - Sacred Books of the East Vol. 3 (The Shu King, Shih King, and Hsiao King)
- Christian Texts
  - 95 Theses
  - The King James Bible
  - Confessions of St. Augustine
  - The Large Catechism
  - Proslogion
  - Small Catechism
  - Summa Theologica
  - The Writings of St. Francis of Assisi
- Hindu Texts
  - The Bhagavad Gita
  - Mahabharata
  - Ramayana
  - Rig Veda
  - Upanishads
  - Vishnupurana
- Judaic Texts
  - Ancient Jewish Proverbs
  - The Guide for the Perplexed
  - Legends of the Jews
  - Mishna
  - The Old Testament
  - The Talmud

- Islamic Texts
  - Al Sira al Nabawiyya
  - The Quran
  - Rubayyat
  - Sahih Muslimen
  - Sirat
- Small Cosmopolitan Corpus
  - Amitabha
  - Dao De Jing
  - Digha Nikaya
  - The Guide for the Perplexed
  - Li Jing
  - Lotus Sutra
  - The Quran
  - Rig Veda
  - Summa Theologica

Appendix III – Intra-Corpus Nearest Neighbors

Link to Buddhist Nearest Neighbors

Link to Chinese Nearest Neighbors

Link to Christian Nearest Neighbors

Link to Hindu Nearest Neighbors

Link to Judaic Nearest Neighbors

Link to Muslim Nearest Neighbors

Appendix IV – Visualizations of Word Embeddings

Link to Buddhist Word Embeddings Visualizations

Link to Chinese Word Embeddings Visualizations

Link to Christian Word Embeddings Visualizations

Link to Hindu Word Embeddings Visualizations

Link to Judaic Word Embeddings Visualizations

Link to Muslim Word Embeddings Visualizations

Appendix V – Cross Cultural Similarities

Link to Pairwise Cross Cultural Similarities

Link to Cosmopolitan Similarities