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 next steps.

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 people 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.

In this project I shall study a certain set of key word/concepts (both concrete and abstract) across corpora from selected religions, philosophies, and folk tales. For each corpus, I will provide a list of words similar to the key word to provide a context of how each culture conceives the notion of that key word. I also present a method for comparing similarity of notions across corpora.

Part II. Background:

In traditional natural language processing, words were represented simply as discrete features produced though the one-hot encoding of the documents and corpora in which they were contained. Through analyzing co-occurance 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. Even for words that do co-occur, 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 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 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 it was 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:

Corpora – English translations of the following:

- Christian
    - King James Bible
- Islam
    - Koran
- Hindu
    - Vedas
    - Upanishads
    - Bhagavad Gita

- Ancient Egypt
    - Book of the Dead
- Chinese
    - Dao De Jing
- Buddhist
    - Lotus Sutra

These will be augmented with other works and possibly other categories. A full list will be presented in the appendix of the final draft. Basic pre-processing will be performed to attempt to remove text matter that is not intrinsic to the text itself. Words will be tokenized such that punctuation will be removed and all words made lowercase. Words will then have basic lemmatization applied before being used for training the model.

Word Embeddings

For purposes of learning word vectors, I use the skip-gram method proposed by Mikolov et al. [7]. This method has the advantages of both producing word vectors that perform well on word similarity tests and being (relatively) fast to train [8].

Inter and Intra Corpus Relationships

Initially, all corpora are treated as distinct and unique vector spaces are learned for each corpus. Within each corpus, I calculate the 8 nearest neighbors to each key word as determined by cosine similarity.

Next, I (will) prepend key words with a character denoting the corpus from which they originate. The corpora are than merged into a 'Cosmopolitan Corpus' and word embeddings in a unified vector space is learned from the combined corpus. Due to the prepended words now being distinct words within a shared vector space, we can see how the different culture's usages of the key words compare within the context of a shared language vector space. To accomplish this, for each key word, I produce a distance matrix showing the cosine distance between word vectors of the key words for each culture.

Part IV. Results:

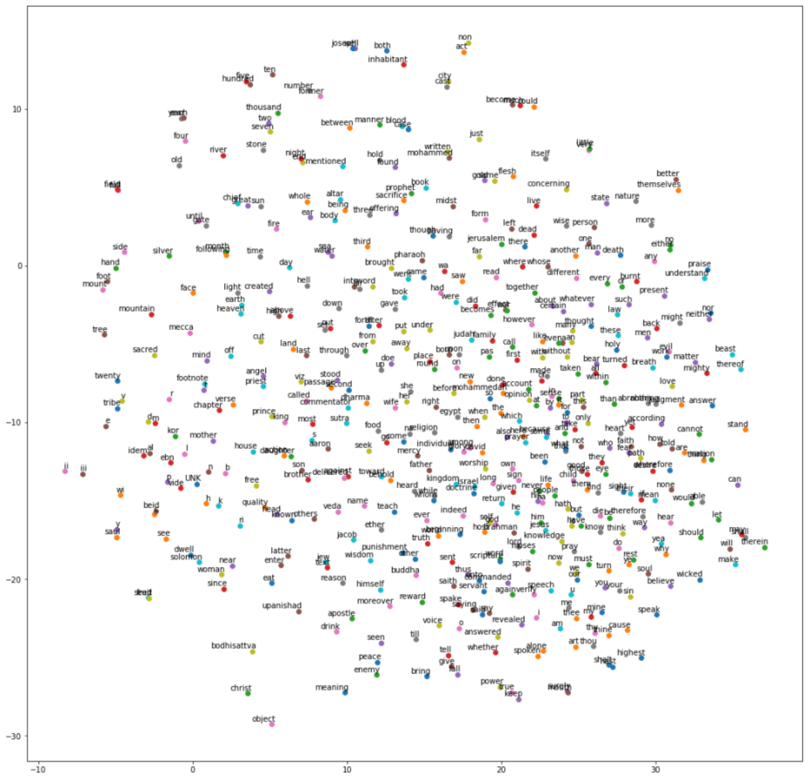
For this milestone, I have acquired a set of documents and performed word embeddings on The Bible (individually), the documents enumerated above as a single corpus, and a larger collection of religious, mythologic, philosophic, and folk-lore documents using the skip-gram algorithm. The embeddings are projected onto a 2-dimensional space for visualization purposes. The list of nearest neighbors and visualizations from the set enumerated in Part III are below. Additional neighbor lists and visualizations (along with supporting code) can be found at:

https://github.com/ReidKay/w266_project/blob/master/word2vec_milestone.ipynb

Nearest Neighbors:

Nearest to good: them, not, zeeb, but, which, of, for, stuhe,
Nearest to bad: taketh, simply, gave, vicinity, defilement, win, day, herdmen,
Nearest to heaven: tahpenes, earth, lovingkindness, god, soundless, land, chap, reared,
Nearest to hell: fight, buddhanta, prohibition, into, universe, num, invoked, intends,
Nearest to man: he, one, purushah, is, or, a, there, for,
Nearest to king: shedd, s, him, son, that, alarm, this, unto,
Nearest to evil: work, not, them, shewest, men, triple, catcheth, overcomes,
Nearest to holy: these, ruhe, all, katyayaniputra, almug, determine, daubed, power,
Nearest to peace: rushing, grasped, invisible, perception, vyapade, gentleness, and, them,
Nearest to salvation: groweth, for, alpha, physical, whale, will, strictly, hotama,
Nearest to eternal: men, sittest, barest, chapel, sarah, him, du, wanton,
Nearest to time: day, speculative, double, it, that, reprehended, partiality, at,
Nearest to destroy: cummin, spend, sole, not, ayasya, eight, varkaruniputra, madhukkhandas,
Nearest to pray: is, kalpe, dogma, saith, thee, killing, adorn, nuvidh,
Nearest to god: lord, and, he, for, therefore, the, me, unto,
Nearest to faith: tamer, primal, strengthen, manifest, tegas, sahm, dainty, fiercely,
Nearest to death: excess, thenceforth, he, abase, oreb, every, vyapade, averse,
Nearest to birth: indirectly, pursuit, ill, achievement, medes, ida, yibh, evil,
Nearest to life: vamadeva, yibh, his, convenience, yielded, not, livest, zamakh,
Nearest to child: and, varkaruniputra, lightness, them, bush, not, hire, son,
Nearest to sin: abishai, covet, you, nuvidh, your, he, devaputras, continuously,
Nearest to body: value, being, they, reared, jaladharagarjitagho, pradhvamsana, betrayed, museum,
Nearest to mercy: come, he, father, all, error, ritra, and, also,
Nearest to love: agnishtoma, prasnam, and, waning, november, all, god, pursueth,
Nearest to hate: skillful, deceived, cognition, proclamation, off, know, chiefest, is,
Nearest to soul: ye, such, not, brahman, intendeth, faculty, and, eli,
Nearest to justice: trembleth, gall, continueth, ftp, stuhe, whosoever, parosh, mundane

2-D Visualization of word embeddings:



One interesting finding is that nearest neighbor to ‘Life’ is found to be ‘Vamadeva’ which in Vedic/Hindu culture is an aspect of Shiva associated with preservation and vital life energy [9].

Part V. Conclusion and Next Steps:

Done:

- ✓ Acquire Documents
- ✓ Create Word Embeddings via Skip-Gram
- ✓ Extract Nearest Neighbors of key words

Remaining:

- o Organize documents into distinct corpora based on the culture from which they arise
- o Perform K-NN on each corpus
- o Prepend key words with corpora tag and create ‘Cosmopolitan Corpus’
- o Create distance matrices from target words in ‘Cosmopolitan Corpus’
- o Comment on interesting findings

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] Wisdomlib.org; Definition of Vamadeva