Cross Cultural Analysis

Satvik Jain
(sj2995@columbia.edu)

INTRODUCTION

We are aware that now days there is a vast amount of multi-media content available belonging to different cultures and nations. With the advent of video streaming platforms such as YouTube and Bilibili, such video content is easily accessible. The aim of this project is to determine the cultural differences between different countries through the analysis of their respective news video feeds. In this project we follow a unique approach of doing multi-modal analysis by combining both textual and imagery information using a deep learning pipeline. Then finally we take common frames and analyze the different keywords associated with those frames for both the cultures to analyze the difference between those cultures.

An important part of this analysis is to extract important keywords from the description of a particular video frame. These keywords are then to be used as a part of the final UI display in which we display the video frames and the corresponding video descriptions. These keywords would assist the user in summarizing the content of the video and get an understanding of the key information being conveyed by the video frame.

In our experiments, we focus on a specific event which is the famous match Alpha Go by Google AI and the human Go champion. We use this event to analyze the cultural difference between USA and China. This event was extensively covered by both US and China and hence serves as a useful database of multi-media content for conducting our research.

This report covers the approaches used for generating the keywords for the video descriptions, namely manual selection and selection using TF-IDF (Term Frequency Inverse-Document Frequency) analysis. The keywords generated using each of the words are presented and then finally combined to create a keyword list for each of the video descriptions. This keyword list would then is then integrated in the UI interface.

BACKGROUND

One of the most common approaches used for text summarization is using the TF IDF analysis. TFIDF, short for term frequency–inverse document frequency, is a numeric measure that is used to score the importance of a word in a document/text based on how often did it appear in that text and a given collection of texts. The intuition behind this measure is: If a word appears frequently in a document, then it should be important and we should give that word a high score. But if a word appears in too many other documents, it's probably not a unique identifier, therefore we should assign a lower score to that word.

TF-IDF analysis is done using the following two important formulas:

TF(w) = (Number of times term w appears in a document) / (Total number of terms in the document)

IDF(w) = log_e (Total number of documents / Number of documents with term w in it)

Hence TF-IDF for a word can be calculated as:

TFIDF(w) = TF(w) * IDF(w)

DATA

The data for the analysis were the video transcriptions that were obtained using the Google translate results that were generated from the audio corresponding to a particular video frame.

Based on the UI design we had, there were 10 data points (bubbles) corresponding to US videos, 5 data points (bubbles) corresponding to Common (both US and China) videos and 8 data points

(China). In my experiments, I focused on the US bubbles and the common bubbles since the TF-IDF tools and library were for the English language.

However, since the corpus was limited, we did not just rely on the TF IDF analysis and hence we also used manual analysis to combine human expertise along with the automated analysis. The details related to the TF-IDF analysis, manual analysis are explained in the following section.

METHODOLOGY

1. TF-IDF analysis

The following pre-processing steps were used in conducting the TF-IDF analysis for extracting the keywords from the video descriptions:

1. Tokenize the sentences – the sentences were tokenized using the 'sent_tokenize' function in NLTK (Natural Language Toolkit)

2. The sentences were further processed by converting them into lower case along with removal of tags and special characters and digits.

3. Then the 7 video transcriptions from US were taken as the corpus and the TF-IDF analysis was performed using the leave one out approach – i.e. for each iteration of the evaluation, the document for which the TF-IDF scores are to be generated (keywords to be generated) is treated as the test set and all the other video descriptions are treated as the training data.

4. Using the above leave one out approach, the training data is first transformed using CountVectorizer and TfidfTransformer.

5. Then the trained transformer is used to generate the TF-IDF predictions on the test document.

6. TF-IDF vectors are then arranged in descending order of the value of their scores.

7. The top 5 vectors for each transcription are then extracted –which are the keywords for that transcription.

The results of the TF-IDF results for the video transcriptions are shown below:

=====Doc=====
when asked about the match case that he was shocked by a couple of moves during mid game tools as those moves would have been played by human

===Keywords===
moves 0.497
shocked 0.359
mid 0.359
couple 0.359
played 0.294

=====Doc=====
the long-held belief that machines can't be professional go players due to the games complex and cheated and creative nature

===Keywords===
professional 0.343
nature 0.343
machines 0.343
held 0.343
complex 0.343

=====Doc=====

we used the previous versions of alphago and yet it was able to pull a much higher level to see using much more principled algorithms

===Keywords===
alphago 1.0

=====Doc=====
Go is an ancient Chinese board game where the opposing players try to capture each other stones on the board

===Keywords===
board 0.72
ancient 0.439
chinese 0.36
players 0.304
game 0.26

=====Doc=====
Go is an ancient, aristocratic Chinese board game that’s reputed to have as many possible moves as there are atoms in the universe. And Google recently trained an artificial intelligence computer to play against one of the best human players in the world. The computer won.At Google’s Future of Go Summit, 19-year-old Chinese Go prodigy Ke Jie was defeated by the AI AlphaGo in a three-match series.AI evangelists are happy with the win, but AI doomsayers are worried it’s coming for our jobs next. And China is just mad that an American company beat the world at a Chinese game.VICE News reports on what the competition really means for AI development.

===Keywords===
ai 0.552
chinese 0.34
google 0.276
computer 0.276
world 0.226

=====Doc=====
Artificial Intelligence program AlphaGo defeated the world's top-ranked Go-player Ke Jie in the first of three games on Tuesday in Wuzhen of east China's Zhejiang Province.After four-and-a-half hours of play, Ke, playing black, lost by 0.5 points, which is the narrowest margin possible in the game. The game follows Chinese Go rules with black having the advantage of first move, and a set point of 7.5 was later given to white to compensate for this.When asked about the match, Ke said he was shocked by a couple of moves during mid-game talks as those moves wouldn't be played by a human. 'When I first saw it, I thought it was almost an impossible move for human players to come up with, since it is obviously a later step. But afterward, I realized it was really an astonishing move,' said Ke. With a newly upgraded version of AlphaGo bolstered

by reinforcement learning, the founder of DeepMind - the company behind AlphaGo, hopes Ke can help discover potential weaknesses of the program. '(It's) Especially interesting for us to see in use some of the moves like the three-three move from the master series of games against AlphaGo, and we were very keen to see how AlphaGo will deal with its own strategies,' said Demis Hassabis, CEO of DeepMind. Ke Jie said AlphaGo has advanced much faster than he thought. 'Compared to last year, AlphaGo's understanding of Go has progressed so much. Last year it played in a human-like way, but this time, it's almost like the God of Go,' said Ke. There will be two more matches between Ke Jie and AlphaGo on Thursday and Saturday. AlphaGo gained worldwide fame when it scored a landmark 4-to-1 victory over South Korean Go master Le Se-dol in a five-round showdown last year, overturning the long held belief that machines can't beat professional Go players due to the game's complex, intuitive and creative nature. Ke, 19, became the youngest champion in Go history after winning three world titles within the space of one year between January 2015 and January 2016. The winner will be awarded 1.5 million U.S. dollars while the losing side takes home 300,000 dollars.

===Keywords===
ke 0.623
alphago 0.431
year 0.227
like 0.208
jie 0.208

=====Doc=====
Artificial intelligence is coming – so how's it going to change our reality?In March of this year, Google's artificial intelligence, AlphaGo, beat one of the top human intelligences, Lee Sedol, at the strategically mind-boggling board game Go. Experts had thought we were years away, but the computer played elegant, creative moves to outfox a Go master. So are we on the brink of an AI revolution? I asked Dr Peter Bentley, a computer scientist from University College London, for some expert insight:Peter Bentley, a computer scientist at University College London, says "since the beginning of artificial intelligence research, one of the main ways that we have tested the intelligence of our computers is to ask them to play games with us, and the progression towards the recent victory has been a long one. But in all of these cases playing games is a hugely simple task."In a game there's a clear 'winning' outcome and it's a closed environment, so the spectrum of possibilities can be accurately predicted. A Go stone will not suddenly turn into a chess piece, for example, or a sausage. Google wants to transfer AlphaGo to real world situations, like medicine. So how does an AI brought up on boardgames hold up in the real world?"It's a very pure clean simple problem, playing a game. The rules are precise, there is no fuzziness, you either are allowed to do that or you are not allowed to do it, and actually real intelligence is completely nothing to do with precision. Real intelligence is about surviving in a horrible, complicated, messy world that's trying to kill you, that's what intelligence is for! That's why organisms have intelligence – to survive! So playing a computer game is a neat trick," says Bentley.Bentley also states, "one of the things that's coming through now is an increasing use of computers to do creative things, that's computers composing music, creating artwork, doing exotic special effects in movies – all sorts of really unusual things that we might not think of but a computer does think of it – for a long time there's been a long debate what is creativity? Could

a computer ever be creative? And the news is yes it can be. Not only can it be creative, it can do things that really amaze us and make us think, holy crap I wish I'd thought of that."Artificial intelligence will change our lives. Already AlphaGo's first victim says he's learned to play better by playing against the machine. Imagine what we will learn as AI is unleashed onto our world.

===Keywords===
intelligence 0.527
computer 0.428
playing 0.286
world 0.234
creative 0.234

2. Manual Analysis:

Since the corpus that was used for the TF-IDF analysis was generated from a very limited sample, manual inspection technique was also used. A peer review method was used in which each of the transcriptions was shown to a group of 5 people each of whom came up with a set of keywords on reading the descriptions. Then all of the keywords generated were compared and for each description those keywords were selected which were chosen the most frequently in the 5 sets of keywords for that transcription.

By using the above methodology, following were the sets of keywords for each of the video transcriptions:

Transcription:

When asked about the match case that he was shocked by a couple of moves during mid game tools as those moves would have been played by human",

Keywords:

Shocked, mid game, human

Transcription:

The long-held belief that machines can't be professional go players due to the games complex and cheated and creative nature

Keywords: complex, creative, machines, professional go players

Transcription:

We used the previous versions of alphago and yet it was able to pull a much higher level to see using much more principled algorithms

Keywords: previous version, alphago, higher level, principled algorithms

Transcription:

Go is an ancient Chinese board game where the opposing players try to capture each other stones on the board

Keywords: board game, Chinese, stones, opposing players.

Transcription:

Go is an ancient, aristocratic Chinese board game that’s reputed to have as many possible moves as there are atoms in the universe. And Google recently trained an artificial intelligence computer to play against one of the best human players in the world. The computer won. At Google’s Future of Go Summit, 19-year-old Chinese Go prodigy Ke Jie was defeated by the AI AlphaGo in a three-match series. AI evangelists are happy with the win, but AI doomsayers are worried it’s coming for our jobs next. And China is just mad that an American company beat the world at a Chinese game.VICE News reports on what the competition really means for AI development.

Keywords: Go**,** board game, AI, Google, Ke Jei, defeated, job loss, Chinese

Transcription:

Artificial Intelligence program AlphaGo defeated the world's top-ranked Go-player Ke Jie in the first of three games on Tuesday in Wuzhen of east China's Zhejiang Province.After four-and-a-half hours of play, Ke, playing black, lost by 0.5 points, which is the narrowest margin possible

in the game. The game follows Chinese Go rules with black having the advantage of first move, and a set point of 7.5 was later given to white to compensate for this.When asked about the match, Ke said he was shocked by a couple of moves during mid-game talks as those moves wouldn't be played by a human. 'When I first saw it, I thought it was almost an impossible move for human players to come up with, since it is obviously a later step. But afterward, I realized it was really an astonishing move,' said Ke. With a newly upgraded version of AlphaGo bolstered by reinforcement learning, the founder of DeepMind - the company behind AlphaGo, hopes Ke can help discover potential weaknesses of the program. '(It's) Especially interesting for us to see in use some of the moves like the three-three move from the master series of games against AlphaGo, and we were very keen to see how AlphaGo will deal with its own strategies,' said Demis Hassabis, CEO of DeepMind. Ke Jie said AlphaGo has advanced much faster than he thought. 'Compared to last year, AlphaGo's understanding of Go has progressed so much. Last year it played in a human-like way, but this time, it's almost like the God of Go,' said Ke. There will be two more matches between Ke Jie and AlphaGo on Thursday and Saturday. AlphaGo gained worldwide fame when it scored a landmark 4-to-1 victory over South Korean Go master Le Se-dol in a five-round showdown last year, overturning the long held belief that machines can't beat professional Go players due to the game's complex, intuitive and creative nature. Ke, 19, became the youngest champion in Go history after winning three world titles within the space of one year between January 2015 and January 2016. The winner will be awarded 1.5 million U.S. dollars while the losing side takes home 300,000 dollars

Keywords: AlphaGo, Ke Jie, shocked, astonishing move, reinforcement learning, DeepMind, God of Go

Transcription:

Artificial intelligence is coming – so how's it going to change our reality?In March of this year, Google's artificial intelligence, AlphaGo, beat one of the top human intelligences, Lee Sedol, at the strategically mind-boggling board game Go. Experts had thought we were years away, but the computer played elegant, creative moves to outfox a Go master. So are we on the brink of an AI revolution? I asked Dr Peter Bentley, a computer scientist from University College London, for some expert insight:Peter Bentley, a computer scientist at University College London, says "since the beginning of artificial intelligence research, one of the main ways that we have tested the intelligence of our computers is to ask them to play games with us, and the progression towards the recent victory has been a long one. But in all of these cases playing games is a hugely simple task."In a game there's a clear 'winning' outcome and it's a closed environment, so the spectrum of possibilities can be accurately predicted. A Go stone will not suddenly turn into a chess piece, for example, or a sausage. Google wants to transfer AlphaGo to real world situations, like medicine. So how does an AI brought up on boardgames hold up in the real world?"It's a very pure clean simple problem, playing a game. The rules are precise, there is no fuzziness, you either are allowed to do that or you are not allowed to do it, and actually real intelligence is completely nothing to do with precision. Real intelligence is about surviving in a horrible, complicated, messy world that's trying to kill you, that's what intelligence is for! That's why organisms have intelligence – to survive! So playing a computer game is a neat trick," says Bentley. Bentley also states, "one of the things that's coming through now is an increasing use of computers to do creative things, that's computers composing music, creating artwork, doing

exotic special effects in movies – all sorts of really unusual things that we might not think of but a computer does think of it – for a long time there's been a long debate what is creativity? Could a computer ever be creative? And the news is yes it can be. Not only can it be creative, it can do things that really amaze us and make us think, holy crap I wish I'd thought of that."Artificial intelligence will change our lives. Already AlphaGo's first victim says he's learned to play better by playing against the machine. Imagine what we will learn as AI is unleashed onto our world.

Keywords: Artificial Intelligence, AlphaGo, AI Revolution, games, Real intelligence, creativity

3. TF-IDF + Manual

Finally, to come up with an optimal set of keywords, the keywords generated by the analysis and TF-IDF analysis were combined to create an optimal set of keywords for each of the transcriptions:

=====Doc=====
when asked about the match case that he was shocked by a couple of moves during mid game tools as those moves would have been played by human

===Keywords===
moves
shocked
Mid game
couple
played

=====Doc=====
the long-held belief that machines can't be professional go players due to the games complex and cheated and creative nature

===Keywords===
professional
creative nature
machines
AlphaGo
complex

=====Doc=====
we used the previous versions of alphago and yet it was able to pull a much higher level tsee using much more principled algorithms

===Keywords===
Alphago
algorithms

=====Doc=====
Go is an ancient Chinese board game where the opposing players try to capture each other stones on the board

===Keywords===
board game
ancient
chinese
players
stones

=====Doc=====
Go is an ancient, aristocratic Chinese board game that’s reputed to have as many possible moves as there are atoms in the universe. And Google recently trained an artificial intelligence computer to play against one of the best human players in the world. The computer won.At Google’s Future of Go Summit, 19-year-old Chinese Go prodigy Ke Jie was defeated by the AI AlphaGo in a three-match series.AI evangelists are happy with the win, but AI doomsayers are worried it’s coming for our jobs next. And China is just mad that an American company beat the world at a Chinese game.VICE News reports on what the competition really means for AI development.

===Keywords===
ai
chinese
google
Ke jie
jobs

=====Doc=====
Artificial Intelligence program AlphaGo defeated the world's top-ranked Go-player Ke Jie in the first of three games on Tuesday in Wuzhen of east China's Zhejiang Province.After four-and-a-half hours of play, Ke, playing black, lost by 0.5 points, which is the narrowest margin possible in the game. The game follows Chinese Go rules with black having the advantage of first move, and a set point of 7.5 was later given to white to compensate for this.When asked about the match, Ke said he was shocked by a couple of moves during mid-game talks as those moves wouldn't be played by a human. 'When I first saw it, I thought it was almost an impossible move for human players to come up with, since it is obviously a later step. But afterward, I realized it was really an astonishing move,' said Ke. With a newly upgraded version of AlphaGo bolstered by reinforcement learning, the founder of DeepMind - the company behind AlphaGo, hopes Ke can help discover potential weaknesses of the program. '(It’s) Especially interesting

for us to see in use some of the moves like the three-three move from the master series of games against AlphaGo, and we were very keen to see how AlphaGo will deal with its own strategies,' said Demis Hassabis, CEO of DeepMind. Ke Jie said AlphaGo has advanced much faster than he thought. 'Compared to last year, AlphaGo's understanding of Go has progressed so much. Last year it played in a human-like way, but this time, it's almost like the God of Go,' said Ke. There will be two more matches between Ke Jie and AlphaGo on Thursday and Saturday. AlphaGo gained worldwide fame when it scored a landmark 4-to-1 victory over South Korean Go master Le Se-dol in a five-round showdown last year, overturning the long held belief that machines can't beat professional Go players due to the game's complex, intuitive and creative nature. Ke, 19, became the youngest champion in Go history after winning three world titles within the space of one year between January 2015 and January 2016. The winner will be awarded 1.5 million U.S. dollars while the losing side takes home 300,000 dollars.

===Keywords===
ke jie
alphago
Reinforcement Learning
God of Go
Deep Mind

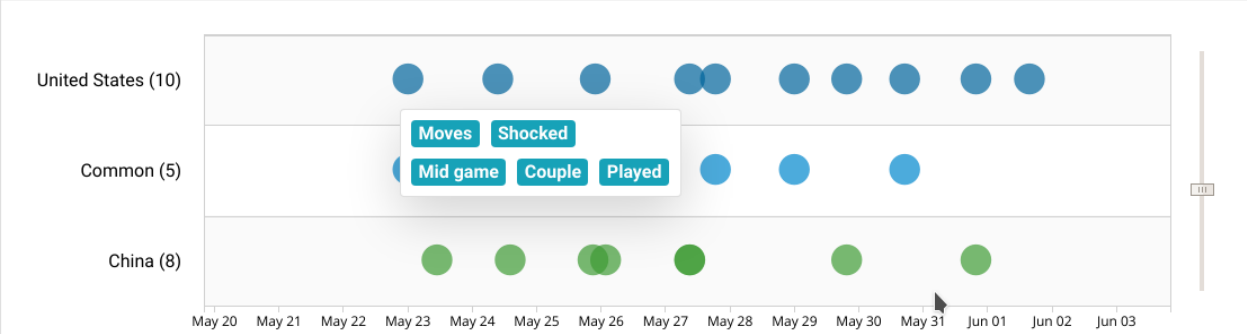
=====Doc=====
Artificial intelligence is coming – so how’s it going to change our reality?In March of this year, Google’s artificial intelligence, AlphaGo, beat one of the top human intelligences, Lee Sedol, at the strategically mind-boggling board game Go. Experts had thought we were years away, but the computer played elegant, creative moves to outfox a Go master. So are we on the brink of an AI revolution? I asked Dr Peter Bentley, a computer scientist from University College London, for some expert insight:Peter Bentley, a computer scientist at University College London, says “since the beginning of artificial intelligence research, one of the main ways that we have tested the intelligence of our computers is to ask them to play games with us, and the progression towards the recent victory has been a long one. But in all of these cases playing games is a hugely simple task.”In a game there’s a clear ‘winning’ outcome and it’s a closed environment, so the spectrum of possibilities can be accurately predicted. A Go stone will not suddenly turn into a chess piece, for example, or a sausage. Google wants to transfer AlphaGo to real world situations, like medicine. So how does an AI brought up on boardgames hold up in the real world?“It’s a very pure clean simple problem, playing a game. The rules are precise, there is no fuzziness, you either are allowed to do that or you are not allowed to do it, and actually real intelligence has completely nothing to do with precision. Real intelligence is about surviving in a horrible, complicated, messy world that’s trying to kill you, that’s what intelligence is for! That’s why organisms have intelligence – to survive! So playing a computer game is a neat trick,” says Bentley.Bentley also states, “one of the things that’s coming through now is an increasing use of computers to do creative things, that’s computers composing music, creating artwork, doing exotic special effects in movies – all sorts of really unusual things that we might not think of but a computer does think of it – for a long time there’s been a long debate what is creativity? Could a

computer ever be creative? And the news is yes it can be. Not only can it be creative, it can do things that really amaze us and make us think, holy crap I wish I'd thought of that."Artificial intelligence will change our lives. Already AlphaGo's first victim says he's learned to play better by playing against the machine. Imagine what we will learn as AI is unleashed onto our world.

===Keywords===
intelligence
Computer game
AI Revolution
world
creative

Results and Display

The above final set of keywords (generated by combining Manual + TF-IDF) is used in the User Interface as pop-ups corresponding to each frame bubble. This is shown in the image below:



The above is a screenshot from the UI display. The keywords can be seen as a pop up from the bubble corresponding to a US video frame. Such a UI helps in summarizing the transcription and also provides ready and relevant information to the user.

References:

1. https://towardsdatascience.com/text-summarization-using-tf-idf-e64a0644ace3

2. https://www.ijarcce.com/upload/2016/march-16/IJARCCE%2040.pdf

3. http://www.cs.columbia.edu/~jrk/NSFgrants/videoaffinity/Interim/19y_kathleen.pdf
4. http://www.cs.columbia.edu/~jrk/NSFgrants/videoaffinity/Interim/19x_stanley.pdf
5. http://www.cs.columbia.edu/~jrk/NSFgrants/videoaffinity/Interim/19y_zikun.pdf

Code Snippets for TF-IDF Analysis:

In [2]:
```
import re
def pre_process(text):

    # lowercase
    text=text.lower()

    #remove tags
    text=re.sub("<!--?.*?-->","",text)

    # remove special characters and digits
    text=re.sub("(\\d|\\W)+"," ",text)

    return text
```

In [4]:
```
doc_list_new = []

for text in doc_list:
    text_new = pre_process(text)
    doc_list_new.append(text_new)
```

In [5]:
```
def sort_coo(coo_matrix):
    tuples = zip(coo_matrix.col, coo_matrix.data)
    return sorted(tuples, key=lambda x: (x[1], x[0]), reverse=True)

def extract_topn_from_vector(feature_names, sorted_items, topn=10):
    """get the feature names and tf-idf score of top n items"""

    #use only topn items from vector
    sorted_items = sorted_items[:topn]

    score_vals = []
    feature_vals = []

    # word index and corresponding tf-idf score
    for idx, score in sorted_items:

        #keep track of feature name and its corresponding score
        score_vals.append(round(score, 3))
        feature_vals.append(feature_names[idx])

    #create a tuples of feature,score
    #results = zip(feature_vals,score_vals)
    results= {}
    for idx in range(len(feature_vals)):
        results[feature_vals[idx]]=score_vals[idx]

    return results
```

In [7]:

```
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfTransformer
for i in range(0,len(doc_list)):

    train_data = doc_list[:i] + doc_list[i+1:]
    test_data = doc_list[i]

    cv=CountVectorizer(max_df=0.85,stop_words= 'english')
    word_count_vector=cv.fit_transform(train_data)

    tfidf_transformer=TfidfTransformer(smooth_idf=True,use_idf=True)
    tfidf_transformer.fit(word_count_vector)

    feature_names=cv.get_feature_names()

    # get the document that we want to extract keywords from
    #doc=docs_test[0]

    #generate tf-idf for the given document
    tf_idf_vector=tfidf_transformer.transform(cv.transform([test_data]))

    #sort the tf-idf vectors by descending order of scores
    sorted_items=sort_coo(tf_idf_vector.tocoo())

    #extract only the top n; n here is 5
    keywords=extract_topn_from_vector(feature_names,sorted_items,5)

    # now print the results
    print("\n=====Doc=====")
    print(test_data)
    print("\n===Keywords===")
    for k in keywords:
        print(k,keywords[k])
```