Cross-Culture Analysis Using NLP Methods

Zikun Lin (supervided by: John R. Kender)

Columbia University, New York, NY, United States

Abstract

We are living in the age of information explosion. Thanks to the rapid growth of technology, we can get more and more information from newspapers, websites, and videos. However, media in different countries and cultures sometimes will have different opinions and focuses on a particular event. In this report, we use some natural language processing and machine learning methods, such as Word2Vec, POS tagging, t-SNE and so on, to discover several typical differences and try to give explanations for these differences. Moreover, we also provide a "blacklist" and a "whitelist" in English and Chinese video descriptions, which can serve as dictionaries and datasets in future work.

1. Introduction

There is a massive amount of online videos. For example, "YouTube", the most well-known video website, is an American video-sharing website headquartered in San Bruno, California, allowing users to upload, view, rate, share, add to playlists, report, comment on videos, and subscribe to other users; it offers a wide variety of user-generated and corporate media videos[1]. In every minute, there are 300 hours of videos uploaded to YouTube. Moreover, in China, there are several leading video websites widely used by people. For example, "Tencent Video", "iQiyi" and "Youku" also have a massive quantity of videos, mostly in Chinese; and "bilibili" is a video website widely used by Chinese teenagers. These websites provide us with excellent sources of videos from different cultures.

In these years, several international events attract the eyes from different countries. There are also topics that may have differences between media in different cultures. For example, (1) "AlphaGo vs. Ke Jie" is an exciting battle between artificial intelligence and a human being. (2) Peter Wang, a fifteen-year-old Chinese American boy, held open the door so others could escape in Florida Shooting Incident. (3) Yingying Zhang, a visiting scholar in the United States from China, was kidnapped by a Champaign resident and former physics graduate student at UIUC. (4) A soccer team aged between eleven to sixteen was rescued

[1] https://en.wikipedia.org/wiki/YouTube

from Tham Luang cave in Thailand by an international team, which is known as Tham Luang cave rescue. Interestingly, different media will hold different opinions, thus having different focuses. And some of the differences can be observed and analyzed by ourselves. For example, in the first event, since Ke Jie is a Chinese, Chinese media tend to focus more on him comparing to media in other countries.

The video descriptions and transcripts collected last semester make it possible to use natural language processing methods to carry out the cross-culture analysis. Several traditional NLP methods can take essential roles in our research. Specifically, POS tagging can help us find out different types and parts of speech; sentence segmentation can assist us in dealing with Chinese sentences. Moreover, in recent years, with the help of deep learning, models like LSTM, transformer, Word2Vec become popular in the NLP field. It enables us to convert words into vectors, making measurements and operations like clustering, classification possible. We can use these well-developed models and methods in our research to make the analysis more quickly and efficiently.

The main contributions of my research can be summarized as follows:

- We use several NLP and ML methods (such as Word2Vec, POS tagging, t-SNE and so on) to find out the properties of news articles and differences between Chinese news and English news on word level. This is described in Section 3.
- We analyze typical advertising sentences and informative sentences from video descriptions and sort out "blacklists" in both English and Chinese and "whitelists" in Chinese from video descriptions. These lists help us carry out the video filter experiments and more work in the future.

To help readers better understand the models used in experiments, I describe the datasets and some related work in Section 2. And in Section 5, I will conclude my research and plan for the future.

2. Related Work

2.1. News Datasets and Social Network Datasets

In order to carry out research in natural language processing, we need to analyze the properties for news, whose language is more formal than the language we use in our daily lives. Also, there are several datasets in both English and Chinese on news and social networks. In this research, the datasets I use are from Reuters newswire, People Daily, Twitter, and Weibo.

As for news, the first dataset I use is"*Reuters-21578, Distribution 1.0*[2]". It contains more than 20,000 news sentences from the Reuters newswire in 1987. And the second dataset I use is news on "People Daily" (the most official newspaper in China) in Jan.1998, which also contains more than 20,000 news articles.

As for social networks, Cheng et al. (2010) collected more than 100,000 twitter users and their updates, UCI's lab and MOEKLINNS Lab also collected the dataset "*microblogPCU*[3]" containing about 50,000 updates on Weibo ("microblog" in Chinese, the largest social network platform in China). These datasets can be used to analyze the properties in social networks.

2.2. News Descriptions and Transcripts

In order to carry out the NLP part of our experiment, having clean data based on our task is necessary. In the Word2vec experiments, I will use the news description data collected by Andy beforehand. These data contain paragraphs from the video description on several topics (AlphaGo, Florida shooting, lunar rover, Thailand cave rescue) in both YouTube and CGTN.

Besides the descriptions, online audio to text converters make analyzing the words in the videos possible. And in fact, as we will see in further analysis, due to some cultural differences, transcripts are more reliable data comparing to video descriptions.

In Section 4, I will also use Chinese video description data and transcripts on these topics from several Chinese video websites ("bilibili", "Tencent Video", "iQiyi" and "Youku").

2.3. Word2Vec

Representing words in a vector space is an efficient way to group similar words and analyze the distribution of a set of words. Rumelhart et al. (1988), Mikolov et al. (2013a) and Mikolov et al. (2013b)'s papers described methods and improvements to represent word and phrases and their compositionality on a vector space. Particularly, Mikolov et al. (2013a) introduced the Skip-gram model, which is an efficient method for learning high-quality vector representations of words from large amounts of unstructured text data, and it is one of the most popular ways to train word vectors.

In order to carry out our experiments quickly, I use Google's pre-trained word and phrase vectors[4], so that we do not need to take much time training from massive datasets. Instead, with the help of Řehůřek and Sojka (2010)'s Gensim library, we only need to call

[2] http://www.research.att.com/~lewis
[3] https://archive.ics.uci.edu/ml/datasets/microblogPCU
[4] https://code.google.com/archive/p/word2vec/

```
model = gensim.models.KeyedVectors.load_word2vec_format()
```

function to load the model and get the vector representation we need.

2.4. POS Tagging

Part-of-speech (POS) tagging is perhaps the earliest and most famous example of "'sequence to sequence' problem. The input to the problem is a sentence. The output is a tagged sentence, where each word in the sentence is annotated with its part of speech. POS tagging is one of the most basic problems in NLP, and is useful in many natural language applications. With the help of NLTK (Natural Language Toolkit) by Loper and Bird (2002), we only need to call the "nltk.word_tokenize()" function to get the POS for each word in the given sentence.

3. Experiments

There are two main parts of experiments. To begin with, in order to find the properties of news articles, I use simple counting method and modern Word2vec method to analyze the differences between news and social network; on the other hand, I also focus on the differences between two major English news sources (YouTube and CGTN) to find out the cultural differences based on the informality and categories of the words they use. Some manual and intuitive analysis is involved.

3.1. Differences between News and Social Network

Section 2.1 described a subtask for our research: finding the difference between words in news and words in social networks, then analyze the difference between them. This section shows the result of this subtask.

3.1.1. Chinese Sentence Segmentation

Although the "People Daily" dataset already provides segmented Chinese words, the "Weibo" dataset only provides raw sentences. Different from English, Chinese sentences do not use spaces as separators of words. Therefore, before carrying out the word-level analysis, we need to separate the words first. A nice tool widely used in Chinese word segmentation is called "jieba"(Sun, 2012).

After using this tool on the "Weibo" dataset, I get more than 70,000 segmented words from about 50,000 Weibo updates.

3.1.2. Word Count Analysis

After segmentation, the total numbers of different words in these four datasets are shown in Table 1.

It is obvious that Twitter has many more different words than other corpora. It is because

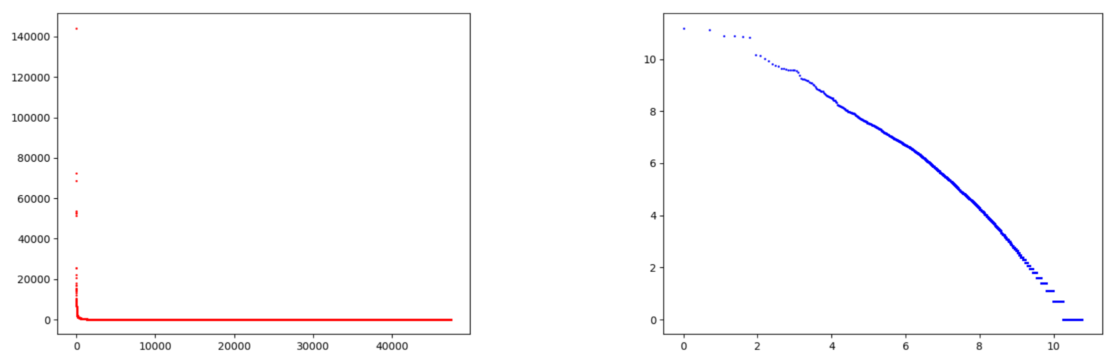
Corpus	Total Number of Different Words
Reuters	47462
People Daily (Chinese)	56482
Twitter	4240058
Weibo (Chinese)	70892

Table 1: Word Count

there are so many URL links and words with typos on Twitter.

From the word list, we can get the word frequency curves. However, the curves are often too deep to be readable because of the existence of several really-high-frequency-words and a large portion of nearly-zero-frequency words. To solve this problem, I apply a Log-log plot to transform the deep curve into a nearly linear line. We change the count in the y-axis to log(count) and the index in the x-axis to log(count). We can see the distributions of the scatters are transformed from deep curves to nearly linear lines in Figure 1, 2, 3, 4.

From these figures, we can see all of them follow the standard rule of word frequency plots: there is often a nearly linear line in the medium-frequency part of the plot, and these words are often are the words we are interested in.



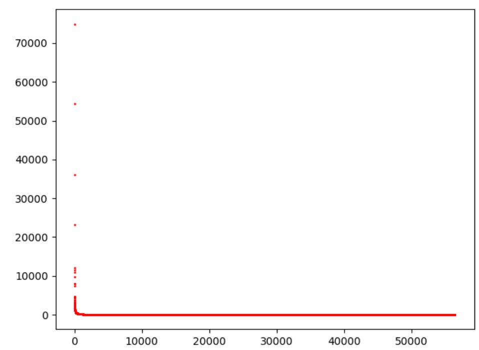
(a) The word frequency plot

(b) The log-log word frequency plot

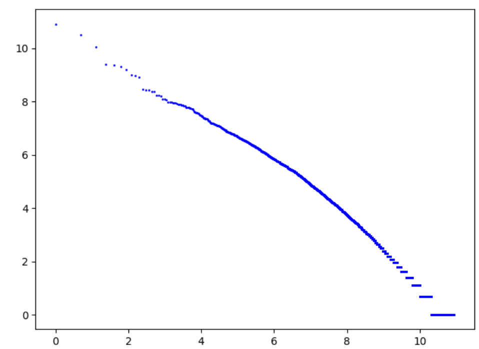
Figure 1: The word frequency plots of Reuters dataset

3.1.3. Word Count Ratio Analysis

It is easy to discover from the datasets and our daily life, that news articles and common languages tend to contain different types of words. News articles are likely to be more "formal",

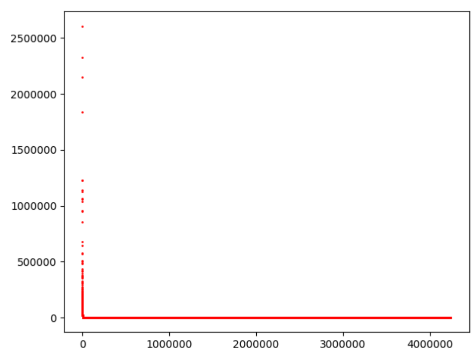


(a) The word frequency plot

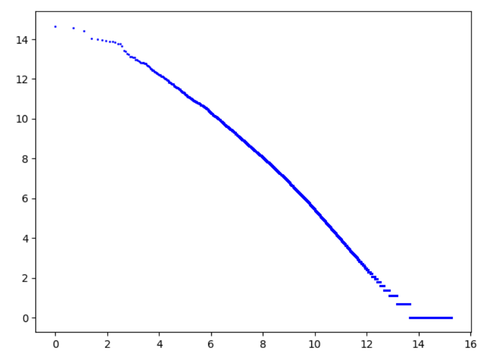


(b) The log-log word frequency plot

Figure 2: The word frequency plots of People Daily dataset

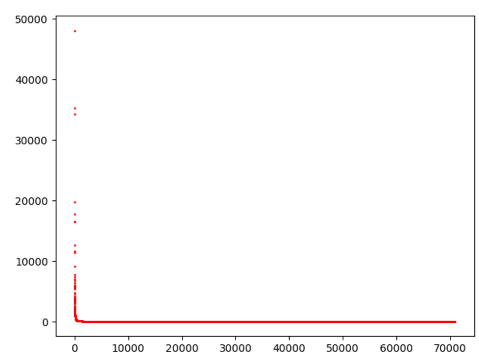


(a) The word frequency plot

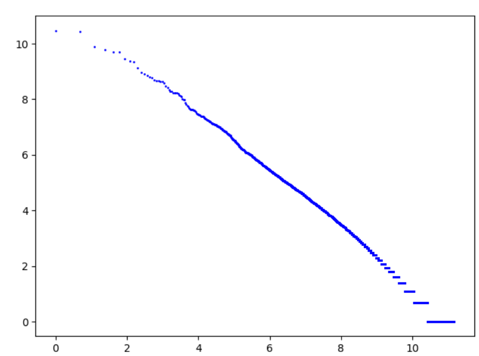


(b) The log-log word frequency plot

Figure 3: The word frequency plots of Twitter dataset



(a) The word frequency plot



(b) The log-log word frequency plot

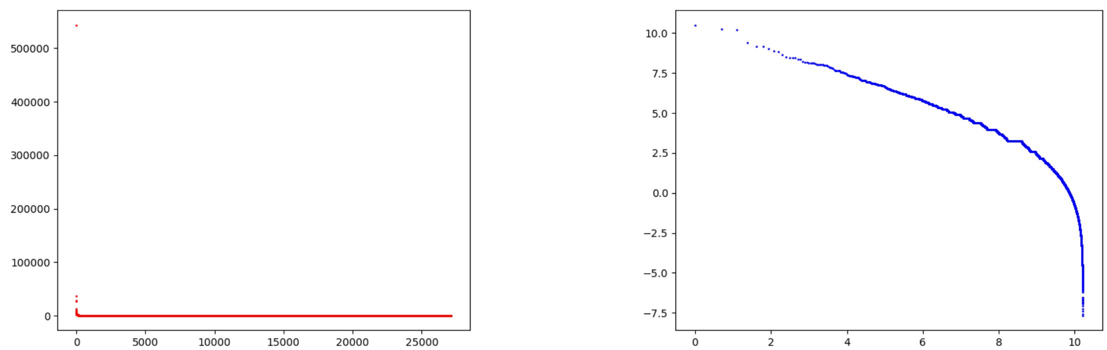
Figure 4: The word frequency plots of Weibo dataset

and contain terms in politics, science, economics and so on; while common languages are likely to contain a wide range of “informal” words. From the word count I have calculated above, we can rank them from “formal” to “informal” depending on the frequency ratio of their usage in news articles and common languages.

Let $f_j(i)$ denote the frequency of word i in type j, then we can define the “frequency ratio” for a word i as follows:

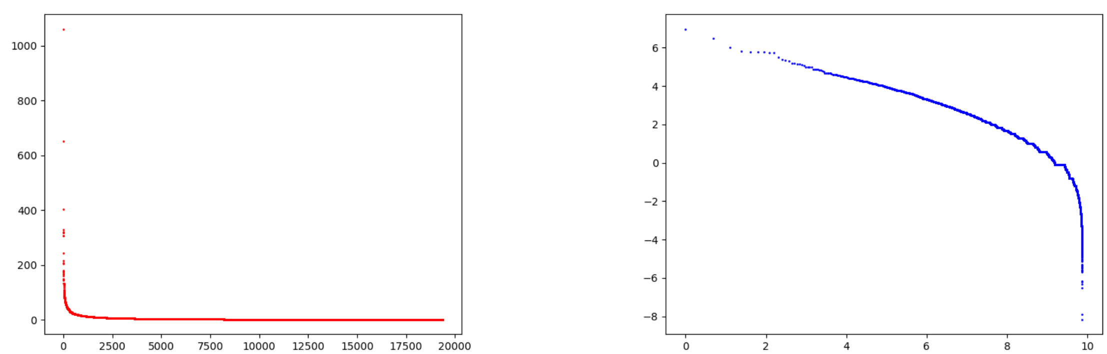
$$FR(i) = \frac{f_1(i)}{\sum_{i'} f_1(i')} \Bigg/ \frac{f_2(i)}{\sum_{i'} f_2(i')}$$

The word frequency ratio plot on English datasets is shown in Figure 5 and Figure 6.



(a) The word frequency ratio plot (b) The log-log word frequency ratio plot

Figure 5: The frequency ratio plots for word count ratio (English: Reuters v.s. Twitter)



(a) The word frequency ratio plot (b) The log-log word frequency ratio plot

Figure 6: The frequency ratio plots for word count ratio (Chinese: People Daily v.s. Weibo)

3.1.4. Word2Vec Analysis for English Datasets

Previous experiments have shown the properties of news and social network corpora in both English and Chinese. Now it is time to determine whether there are differences between them. I use the word2vec model described in Section 2.3. In order to plot the results on a two-dimensional figure, I also use the t-SNE method described in Maaten and Hinton

(2008) to visualizes high-dimensional data by giving each data point a location in a two-dimensional map. Let red points denote the words with the top word frequency ratios from Reuters (news), blue points denote the words with the top word frequency ratios from Twitter (social network). The result is shown in Figure 7.

It is exciting to see the two kinds of points lie in the different halves of the figure, with a clear boundary that can separate the two classes quite well. This phenomenon means that there are apparent differences in topics and meanings for the two different classes.

To further observe the properties of the words from the two classes, I pick several paragraphs from Google News and plot the words' corresponding vectors to the same vector space. We can see that words related to politics and economy are more likely to show in the red half, or "the news half", like "commerce", "currency", "deposits" and so on; and words related to real life are more likely to show in the blue half, or "the social network half", like "good", "him", "you" and so on. What we can conclude from this plot is: news are more likely to talk about serious topics and use more formal words; while people will use informal words more often on social networks.

Another phenomenon we can see is: most words from arbitrary articles on Google News are more likely to appear on the boundary of the two classes, many of them even lie in "the social network" part. It means that most words in news are just "common words" that are not so "formal" or "informal".

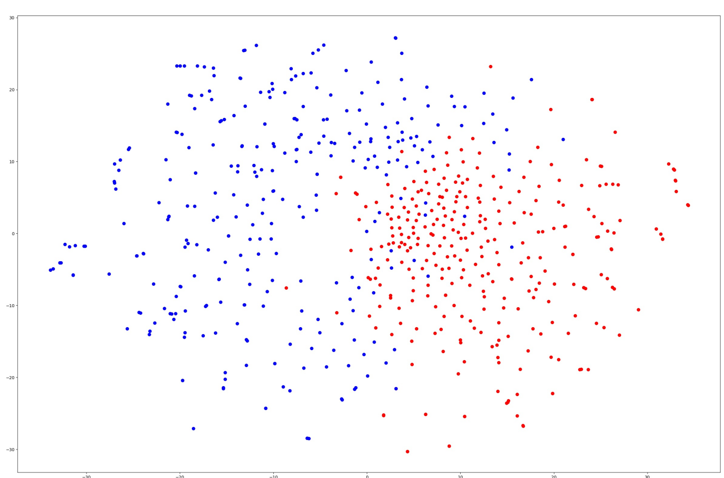


Figure 7: Two kinds of words in the 2-dimensional vector space
(red: news, blue: social network)

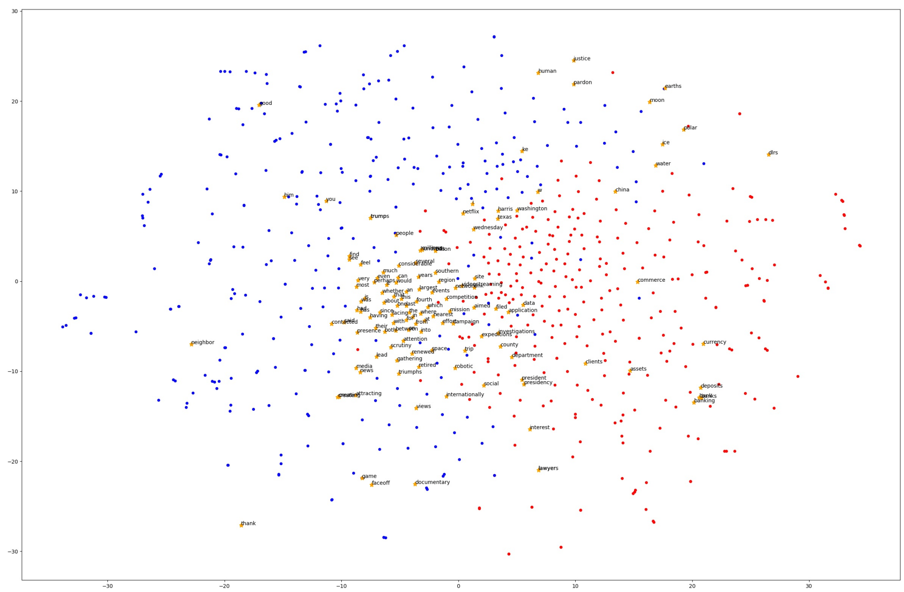


Figure 8: Two kinds of words and common words picked from news in the 2-dimensional vector space (red: news, blue: social network, star: common words picked from Google News)

3.2. Differences between Two Major English News Sources: YouTube and CGTN

The experiment results from the last subsection show that word2vec is a useful tool to discover the topic distribution and properties of several corpora. To be more specific in the several topics we are interested in (AlphaGo, Florida shooting, lunar rover, Thailand cave rescue), I will use similar experimental methods but focus on corpora from these topics in the following experiments.

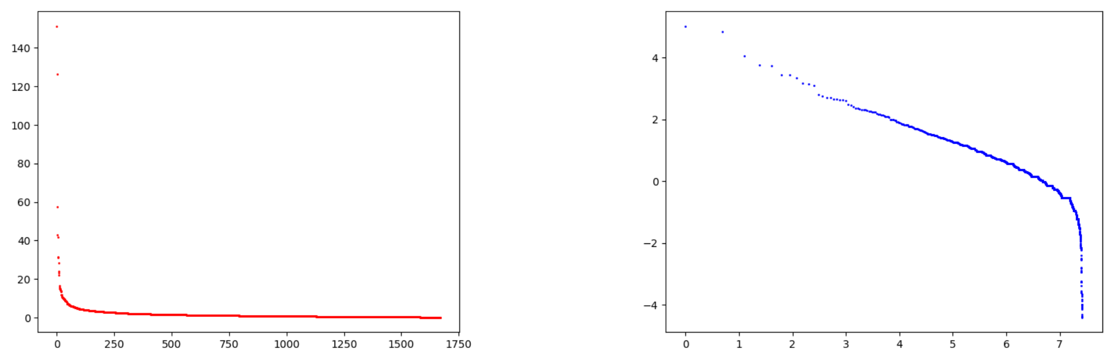
3.2.1. Corpora

In the following experiments, I will use the video descriptions collected by Andy last semester. These descriptions are collected from "CGTN" and "YouTube" using several keywords. "CGTN" stands for "China Global Television Network", which is a Chinese international English-language news channel that might have videos being our targets. There are 1600 video descriptions in total (200 for each topic and each source). The feature that both videos are in English is a great help to our task: comparing the cultural differences between news from Chinese sources and US sources.

3.2.2. Word Count Ratio Analysis

From Andy's experiments in last semester, we have got the word count analysis for all topics in all sources. All of them follow the rule of word frequency plots: there is often a nearly linear line in the medium-frequency part of the plot, and these words are often the words we are interested in.

Therefore, in this subsection, we can do the word count ratio analysis directly. Since the amount of texts is limited, and the average length of each description is not long, I decided to carry out this experiment using all descriptions in 4 topics. Although topics are mixed up, general patterns are expected to be found. I use the method described in Section 3.1.3 and get the following result shown in Figure 9. With much fewer words, it still has similar patterns on both graphs.



(a) The word frequency ratio plot

(b) The log-log word frequency ratio plot

Figure 9: The frequency ratio plots for word count ratio

3.2.3. Word2Vec Analysis

It will also be an interesting topic to think about "Will the topic distributions for video descriptions from YouTube and CGTN different?" and carry out the similar experiment described in Section 3.1.4. The result is shown in Figure 10. This time the pattern in the figure is not so clear since all the words belong to "video description words", and the difference between YouTube (blue) and CGTN (red) is not as big as the difference between news and social networks. We need to find a better way to discover the differences.

3.2.4. Word2Vec Analysis after Tagging

The "better way" we found is to do the word2vec analysis after tagging. As described in Section 2.4, I use the NLTK POS tagging tool to get the part-of-sentence features for each word. After that, I can carry out the "word2vec" analysis separately to each type of POS

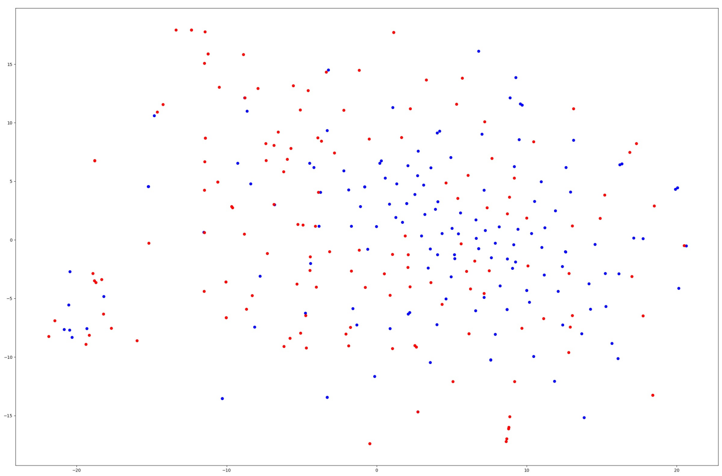


Figure 10: Two sources of words in the 2-dimensional vector space
(red: video description on CGTN, blue: video description on YouTube)

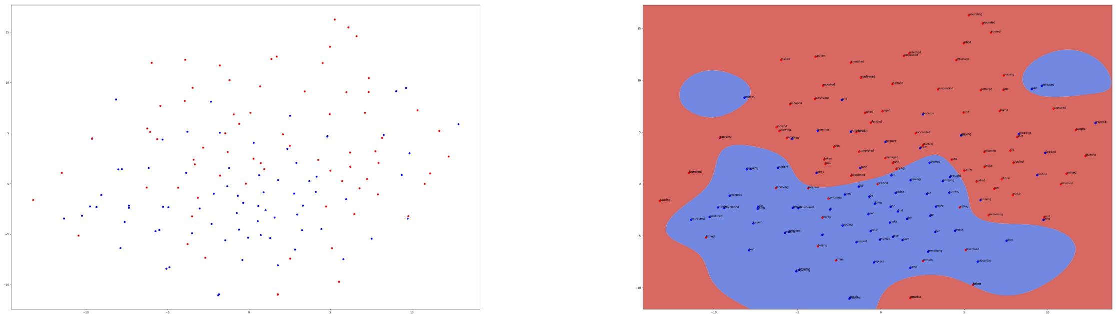
and find more detailed differences between two sources. Since Andy has already carried out the analysis in name entities based on word frequency last semester and found several differences in the use of name entities (mainly nouns), my work will focus more on verbs, adjectives, and adverbs, which are more related to article styles instead of a particular topic. After picking out verbs, adjectives and adverbs, I first use word2vec and t-SNE as before to plot the words on a 2-d vector space. After that, I use SVM (support vector machine) to do the "classification" step. This step does not mean to "train a classifier" for future words. It just serves as a method to find the boundary between the two classes. If the boundary is clear enough, we can conclude that there are considerable differences between the two classes.

Figure 11, 12, 13 show the results for this experiment. The left part of each figure are the words on a 2-d vector space, while the right part of each figure finds out the boundary for the two classes.

It is good news that there are somehow clear boundaries for verbs and adverbs, while there is an existing (although not so clear) boundary for adjectives. This also brings another surprising news for us: the average word length and complexity for CGTN is longer and higher than YouTube. So it is also essential to discover the reasons behind this.

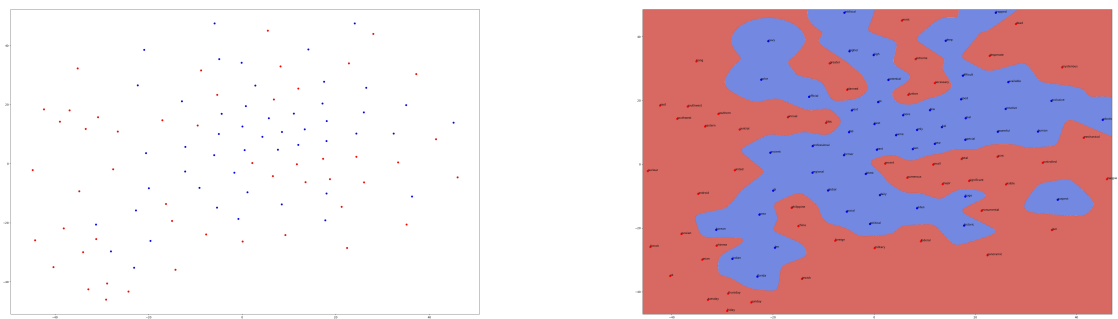
3.2.5. Analysis in Original Description

After observing the results shown above, we decide to go back to original description texts and find the reasons behind the differences. Some of the reasons can potentially be discovered



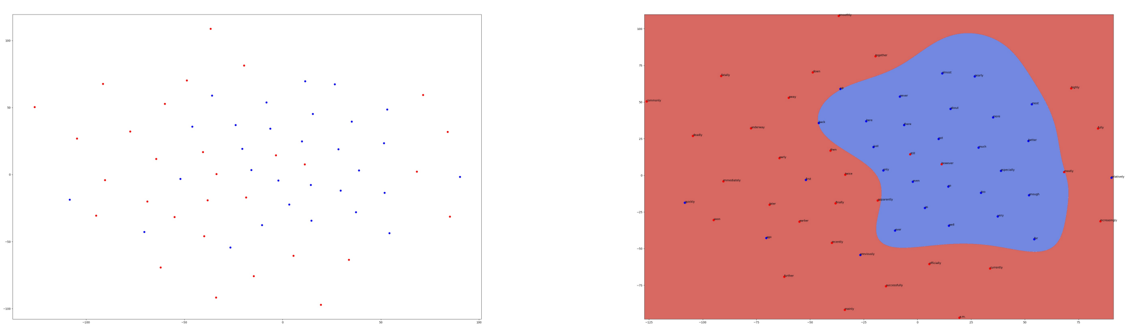
(a) The word2vec plot (b) The heat plot showing boundaries

Figure 11: The word2vec points and classification on verbs



(a) The word2vec plot (b) The heat plot showing boundaries

Figure 12: The word2vec points and classification on adjectives



(a) The word2vec plot (b) The heat plot showing boundaries

Figure 13: The word2vec points and classification on adverbs

as “cultural differences”, which are useful to the final goal of our project. After sorting and reading from the original texts, several words deserve our attention.

Foreign

This adjective "foreign" appears much more times in CGTN than in YouTube. When referring to the original videos, it turns out that the "Ministry of Foreign Affairs" is the primary source of this word in CGTN. MOFA is the ministry releasing information about important diplomatic activities in China[5]. Many videos from CGTN are about the regular press conference of this ministry. This can be treated as the main reason for the difference.

Injured

The verb "injured" appears much more times in CGTN than in YouTube. After discussion, we treat it as a typical cultural difference in news reports between China and foreign countries. The Chinese language has a relatively larger entropy than English according to Chang and Lin (1994), Brown et al. (1992). Therefore, the Chinese language can use only two syllables "死伤" (pronounced as "si shang") to express the meaning of the phrase "death and injury". It is also able to use only four syllables "三死九伤" (pronounced as "san si jiu shang") to express the meaning of "three people die, and nine people are injured." So after an accident happens, media tend to report on the numbers of dead people along with injured people (sometimes also missing people), which becomes a traditional template in Chinese news reports, even in English reports written by Chinese journalists. However on YouTube, "at least 14 victims were taken to the hospital" is a typical expression to report on the number of injured people.

Advertising and Promotion Words

There is a wide variety of sentences in video descriptions that are used to promote their media account on YouTube or other social network accounts like Facebook or Twitter. This is not a significant problem for CGTN since all of the CGTN descriptions have the same promotion sentences. However, as for other media on YouTube, the promotion sentences are in very different formats, or they use different types of words. Since the same media tends to use the same promotion sentences in all of their videos, these words will have a significant impact on the top words in both sources.

Immediately

The adverb "immediately" appears much more times in CGTN than in YouTube. It is probably because of the different article styles that need to be further discovered.

[5] https://www.fmprc.gov.cn/mfa_eng/

4. Data Preparation for Future Work

From the analysis above, we can see a lot of advertising and promotion words appear in video descriptions. So it is necessary for us to find a way to get rid of these words. Although machine learning methods can play essential roles in advertisement detecting, we still need a method or dataset that is more fit to the video descriptions. Inevitably, lot's of manual work will be utilized.

In this section, I first analyze the cultural differences between English and Chinese video descriptions. Then I make a naive and manual selection for advertising and promotion sentences from video descriptions, and then count the word frequencies to get "blacklists" and "whitelists". These sentences and these lists can be both used as dictionaries or datasets in future work.

4.1. Cultural Differences in Descriptions

We mainly dealt with video descriptions from English sources in Section 3. And in this part, we will deal with video descriptions in Chinese, which mainly come from "bilibili", "Tencent Video", "iQiyi" and "Youku". Despite the differences in the usages of words, we also find some unexpected but interesting differences during our research.

The first significant difference is: there are many official media accounts on YouTube, but there are not as many official media accounts on these Chinese video websites. Instead, most "AlphaGo" videos on these websites are uploaded by unofficial organizations or people. As a result, these websites contain a more extensive variety of videos, which also makes it more challenging to pick up the useful information we need.

The other significant difference is: "Description" part is always viewed as an important part on YouTube since it is a good place to show the abstract of the video and promote their social media accounts. So video descriptions on YouTube have longer paragraphs and are typically formal. However, most Chinese people tend to ignore the video descriptions, and some video websites even allow people uploading their videos without filling up the "Description" part. So the descriptions on Chinese video websites are mainly short and casual, sometimes empty. Besides the advertisements and something relevant to video content, there are also many messy sentences that appear in Chinese video descriptions, making it more challenging to analyze Chinese video descriptions.

To deal with this problem, I pick up blacklist words for video descriptions in English, along with blacklist and whitelist words for video descriptions in Chinese in the following sections. The blacklists contain the words from advertisement and promotion sentences, while the whitelists only care about the sentences that are relevant to our topic. Stanley's experiments also prove the effectiveness of these lists.

4.2. Video Descriptions on AlphaGo

4.2.1. Blacklist Words for Video Descriptions in English

Sentences containing advertisement and promotion are treated as “blacklist sentences”. The typical “blacklist sentence” is like this one[6]:

> *Subscribe to VICE News here: http://bit.ly/Subscribe-to-VICE-News*
> *Check out VICE News for more: http://vicenews.com*
> *Follow VICE News here:*
> *Facebook: https://www.facebook.com/vicenews*
> *Twitter: https://twitter.com/vicenews*
> *Tumblr: http://vicenews.tumblr.com/*
> *Instagram: http://instagram.com/vicenews*
> *More videos from the VICE network: https://www.fb.com/vicevideo*

The top 60 words in the blacklist after eliminating punctuation are shown in Table 2.

Word	Frequency	Word	Frequency
news	110	with	22
to	98	is	22
the	90	out	22
on	88	in	21
	77	a	21
subscribe	70	channel	20
vice	60	your	19
us	53	this	17
official	53	cnbc	17
for	52	website	16
and	50	arirang	16
here	46	by	16
httpwwwfacebookcomarirangtvtwitter	39	my	16
httptwittercomarirangworldinstagram	39	httpbitlysubscribetovicenewscheck	15
visit	36	httpvicenewscomfollow	15
of	35	herefacebook	15
‘arirang	34	httpswwwfacebookcomvicenewstwitter	15
news’	34	httpstwittercomvicenewstumblr	15
pagesfacebooknews	34	httpvicenewstumblrcominstagram	15
httpwwwfacebookcomnewsariranghomepage	34	httpinstagramcomvicenewsmore	15
httpwwwarirangcomfacebook	34	network	15
httpinstagramcomarirangworld	34	httpswwwfbcomvicevideo	15
more	31	like	14
facebook	29	tv	14
from	27	intel	14
youtube	26	please	13
you	26	at	13
videos	25	cbs	13
our	25	app	12
twitter	24	software	12

Table 2: Top 60 Words in English Blacklist on AlphaGo

From this blacklist, we can see several typical categories of words:

(a) Media promotion and subscription request. Such as: news (110), videos (25),subscribe (70), visit (36), please (13).

[6] https://www.youtube.com/watch?v=8dMFJpEGNLQ

(b) URL links and social network account. Such as: httpwwwfacebookcomarirangtvtwitter (39), facebook (29).

(c) Media name. Such as: vice (60), ‘arirang (34). In fact, the most “vice”s appear in the descriptions are not acting as “vice president”, but “VICE News” instead.

This list will help us a lot in future work dealing with other video descriptions and transcripts since there are many similar patterns in videos from other topics.

4.2.2. Blacklist and Whitelist Words for Video Descriptions in Chinese

Sentences containing advertisement and promotion are treated as “blacklist sentences”. The typical “blacklist sentence” in Chinese is like this one[7]:

> 螃蟹科技微信公众号：螃蟹科技（*pangxiekeji*）螃蟹科技 *QQ* 群 *419859745*
> 如果对我们的栏目有什么建议或者对智能数码有什么需要了解的，在公众号中回复你想了解的，我们来帮你解答。
> ***Translation:*** *“Crab Technologies” Wechat Official Account: Crab Technologies (pangxiekeji) “Crab Technologies” QQ Group 419859745. If you have any suggestions for our column or need to know about smart digital, reply what you want to know in the Wechat official account, we will answer.*

And the typical “whitelist sentence” in Chinese is like this one[8]:

> 柯洁将在下月迎战谷歌旗下的著名人工智能围棋软件 *AlphaGo*。
> ***Translation:*** *Ke Jie will be battling with Google’s famous AI Go software AlphaGo next month.*

As described in Section 4.1, the cultural differences on descriptions make it more difficult for analysis in Chinese. In this particular topic on AlphaGo, as we can expect, besides the sentences most relevant to AlphaGo and Ke Jie (blacklist sentences) and the promotion and advertising sentences (whitelist sentences), there are also other sentences that don’t belong to any of these two lists, which are referred as “irrelevant sentences”. Although time-consuming, it is quite interesting to read through all of these Chinese descriptions. Some typical irrelevant sentences are shown below.

The following one[9] is collected from a “technology news weekly digest” video. AlphaGo only serves as a small part in this video. So most contents in this video are irrelevant with AlphaGo and Ke Jie.

> 三星 *Note6/7* 工程图曝光　联想 *Moto Z* 真机图泄露　柯洁 *AlphaGo* 即将开战
> *10* 万块军工级手机发布　全球首款带夜视仪的手机发布
> ***Translation:*** *Exposure of Samsung Note6/7 Engineering Drawings, Lenovo Moto Z Real Machine Map Leakage, Ke Jie and AlphaGo are about to battle, 100,000 military grade mobile phones released, first mobile phone with night vision in the world*

[7] https://www.bilibili.com/video/av4116312
[8] https://www.iqiyi.com/v_19rrbbtzbw.html
[9] http://v.qq.com/page/u/m/2/u0305zm4lm2.html

The following one[10] is collected from a funny video imagining AlphaGo playing League of Legends game. These kinds of videos are not from the news, but there are several such kinds of videos on these Chinese websites.

153. 如果 AlphaGo 来玩英雄联盟
Translation: *If AlphaGo plays LOL*

The following one[11] is collected from an industry introduction sentence. It used "AlphaGo" to express that they are using the modern techniques and they are among the first tier.

英飞凌德累斯顿智能工厂，工业 4.0 的 "AlphaGo"
Translation: *Infineon Dresden Intelligent Factory, the "AlphaGo" of Industrial 4.0*

The following one[12] is a bit special. This is a self-edited video with no informative content, and there are several similar videos like this on Chinese video websites. The uploaders of these videos want to express their fondness for somebody or something, so they made these videos using the existing video footage. In this video, the content is mainly collected and edited from news video clips, so most scenes are relevant to the AlphaGo topic. Also, there are many keywords on this topic in the description. Therefore, it will be easily recognized as "related news" if using blacklists and whitelists only.

这个视频的构思想了一年多（是的没写错）从去年小李人机的时候开始想，直到今年才在小十一的古力……
啊不是，鼓励之下开始动手
一个 AI 爱上了人类，最终他们在一起了故事 \# 严肃
第一次做剧情向，剧情比较凌乱，希望能看懂
送给小十一！希望喜欢！！注 1：AlphaGo 来自于 Ex Machina-Domhnall Gleeson
注 2：主 CP 为 AlphaGo/柯洁，副 CP 为木谷实/吴清源，古力/李世石
注 3：2017 年 6 月 2 日更新微调版本。具体剧情见回复
Translation: *This video has been conceived for more than a year (yes, correctly written) since Lee Sedol's battle last year, and it was not until this year that Gu Li was in eleventh ranking.*
This is a story. An AI falls in love with a human being and eventually they get together \# seriously
This is the first time that I make a story video, and the plot is messy. I hope you can understand it.
It's a present for the Eleventh! Hope you like it! Note 1: AlphaGo comes from Ex Machina-Domhnall Gleeson
Note 2: The main couple is AlphaGo / Ke Jie, secondary couples are Minoru Kitani / Wu Qingyuan, Gu Li / Lee Sedol.
Note 3: Updated fine-tuned version on June 2, 2017. See the reply for the specific plot.

In conclusion, Chinese descriptions are much more complicated, so it is challenging to carry out a two-class classification for Chinese descriptions.
After eliminating these irrelevant sentences, the top 60 words in the blacklist are shown in

[10] https://www.iqiyi.com/w_19rub12smp.html
[11] https://v.qq.com/x/page/i0188drze8u.html
[12] https://www.bilibili.com/video/av10975529/

Table 3, the top 60 words in the whitelist are shown in Table 4.

Word	Translation	Frequency	Word	Translation	Frequency
，		87	科技	science and technology	10
的	’s	64	更多	more	9
：		61	！		9
。		25	我	I	9
碧蓝	Azur	23	玩家	player	9
在	at	21	加入	enter	9
航线	Lane	20	了	have done ...	9
中途岛	Midway	16	qq		8
群	group	15	《		8
游戏	game	14	》		8
如果	if	14	com		8
微信	Wechat	14	…		8
集	episode	13	您	you	8
主	main	13	可以	can	8
alphago		13	喂	Hello	8
公众	public	12	id		8
号	account	12	服务器	server	8
交流	communicate	12	服	server	8
欢迎	welcome	12	644132397		8
都	all	11	请	please	7
allen		11	粉丝	fans	7
关注	follow	11	也	also	7
更	more	11	详细	detailed	7
up		11	攻略	strategy	7
有	have	10	尽	use all	7
围棋	Go	10	wiki		7
是	is	10	你	you	6
视频	video	10	对	to	6
、		10	和	and	6
大家	everyone	10	【		6

Table 3: Top 60 Words in Chinese Blacklist on AlphaGo

From the lists shown, we can see that the whitelist for Chinese is much more reliable than blacklist: For whitelist, there are about 10 words that appear more than 100 times, most of which are highly relevant to the topic. However, the top blacklist that has real meaning is “碧蓝 (Azur)”, which only has a frequency of 23. This phenomenon has shown that the blacklist in Chinese is much messy than whitelist, thus much less reliable.

[13] The word “dog” has the same pronunciation as the word “Go” in Chinese, so “Alpha Go” will sometimes be referred as “Alpha Dog” in Chinese news.

Word	Translation	Frequency	Word	Translation	Frequency
alphago		418	战胜	defeat	29
，		366	将	will do	29
的	's	302	对弈	play chess with	29
柯洁	Ke Jie	177	狗	dog[13]	28
。		170	0		25
围棋	Go	131	棋手	chess player	25
大战	battle	117	deepmind		24
人机	human and computer	113	上	up	23
了	have done ...	106	4		23
“		89	比赛	game, competition	23
李世石	Lee Sedol	85	中	middle	23
”		84	谷歌	Google	22
人类	human beings	83	vs		22
在	at	82	第	-th	22
是	be	70	master		21
：		63	阿尔法	Alpha	21
人工智能	artificial intelligence	59	《		21
月	month	57	》		21
5		51	1		21
日	date	51	被	be done	20
3		44	用	use	20
与	and	43	不	no	20
战	battle	42	人	human	20
和	and	38	手	hand	19
中国	China	37	乌镇	Wuzhen (a place in China)	19
？		36	进行	be in progress	19
对	to	35	团队	team	19
、		31	我们	we	19
ai		30	你	you	18
！		30	马云	Jack Ma	18

Table 4: Top 60 Words in Chinese Whitelist on AlphaGo

4.3. Video Descriptions on Florida Shooting

After a discussion, we found that the AlphaGo event is a little general, which means several irrelevant events appeared under the search result of the "AlphaGo" keyword. In order to sort out better "blacklist" and "whitelist", we turn our eyes to another event, the Florida Shooting tragedy.

4.3.1. Blacklist Words for Video Descriptions in English

The top 60 words in the blacklist after eliminating punctuation are shown in Table 5.

From the original data, we can observe that most of the videos come from different sources comparing to AlphaGo videos. However, they share many common blacklist words. For example, the English Blacklist on Florida Shooting shown in Table 5 and the English Blacklist on AlphaGo shown in Table 2 share 5 words in top 10, 8 words in top 20, 22 words in top 40. This means different media also use similar words in advertising and promotion. This observation makes our future work much easier since we can use the blacklists above to discover most of the targets in new topics.

Word	Frequency	Word	Frequency
news	1074	full	83
on	657	is	83
the	619	fox	78
and	511	local	77
cbs	501	episodes	74
to	400	google	74
here	362	all	74
of	324	cbc	72
nbc	234	our	69
subscribe	196	it	68
evening	182	as	67
you	164	broadcast	66
with	156	access	63
a	132	devices	61
morning	126	day	59
watch	121	stories	59
twitter	118	business	59
your	113	original	58
today	110	coverage	56
facebook	109	guardian	54
instagram	108	entertainment	53
	106	new	52
channel	105	video	52
for	105	digital	52
in	105	source	50
this	104	mobile	48
more	98	shows	47
live	97	breaking	46
latest	95	apps	45
from	91	across	45

Table 5: Top 60 Words in English Blacklist on Florida Shooting

4.3.2. Blacklist and Whitelist Words for Video Descriptions in Chinese

(to be added after Stanley provides the original description data)

5. Conclusion and Future Work

From the work described above, we can make several conclusions.

(a) Corpora with considerable amount of words tend to follow the standard rule of word frequency plots: there is often a nearly linear line in the medium-frequency part of the plot, and these words are often are the words we are interested in. This is a great help for recognizing the important words in the corpus.

(b) Words used in news tend to be more formal than those in social media. This phenomenon can be proved by the results of Word2vec.

(c) Using Word2vec and POS tagging, we can observe many differences between words used in two major English news sources: YouTube and CGTN. Some are because of cultural and historical differences, while others involve the words for advertisement and promotion.

(d) We can manually collect “blacklist” and “whitelist” sentences and words from video descriptions. Due to the cultural differences, people treat descriptions differently on English and Chinese platforms. Also, video descriptions collected from Chinese video websites has more variety.

In next few months, I plan to do more experiments to get further and deeper observations based on our research.

(a) Read more relevant papers on natural language processing to find out more methods to discover the cultural differences based on words and sentences.

(b) Carry on Stanley and Kathleen’s previous work to get important video frames from transcript. This will be useful for further experiments, such as Stanley’s network based on both texts and graphics.

(c) Build a larger “blacklist” and “whitelist” dataset. Try to develop a classifier to recognize these kind of words for future topics.

References

Z. Cheng, J. Caverlee, K. Lee, You are where you tweet: a content-based approach to geo-locating twitter users, in: Proceedings of the 19th ACM international conference on Information and knowledge management, ACM, pp. 759–768.

D. E. Rumelhart, G. E. Hinton, R. J. Williams, et al., Learning representations by back-propagating errors, Cognitive modeling 5 (1988) 1.

T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2013a).

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in neural information processing systems, pp. 3111–3119.

R. Řehůřek, P. Sojka, Software Framework for Topic Modelling with Large Corpora, in: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, ELRA, Valletta, Malta, 2010, pp. 45–50. http://is.muni.cz/publication/884893/en.

E. Loper, S. Bird, Nltk: the natural language toolkit, arXiv preprint cs/0205028 (2002).

J. Sun, 'jieba' chinese word segmentation tool, 2012.

L. v. d. Maaten, G. Hinton, Visualizing data using t-sne, Journal of machine learning research 9 (2008) 2579–2605.

J.-s. Chang, Y.-J. Lin, An estimation of the entropy of chinese-a new approach to constructing class-based n-gram models, in: Proceedings of Rocling VII Computational Linguistics Conference VII, pp. 149–169.

P. F. Brown, V. J. D. Pietra, R. L. Mercer, S. A. D. Pietra, J. C. Lai, An estimate of an upper bound for the entropy of english, Computational Linguistics 18 (1992) 31–40.