Word Embedding Training and Analogical Reasoning for Finance Information System

Bin Yan

Intern AI Engineer, Kingdom AI, Beijing, China

Abstract

This project works as a technical foundation for the Information Analysis System for the Hong Kong Exchanges and Clearing Limited (HKEX) by Kingdom AI. In addition, it is part of the cooperative NLP Research for Chinese linguistic analysis and applications. It intends to create and train numerical linguistic models especially based on Chinese financial corpus and other resources. Through this model, we will prove its optimized potential to provide better results in embedding, categorization, word clustering and visualization for financial analyses, and propose the architecture to deal with the vast information for stock fluctuation prediction and performance analyses.

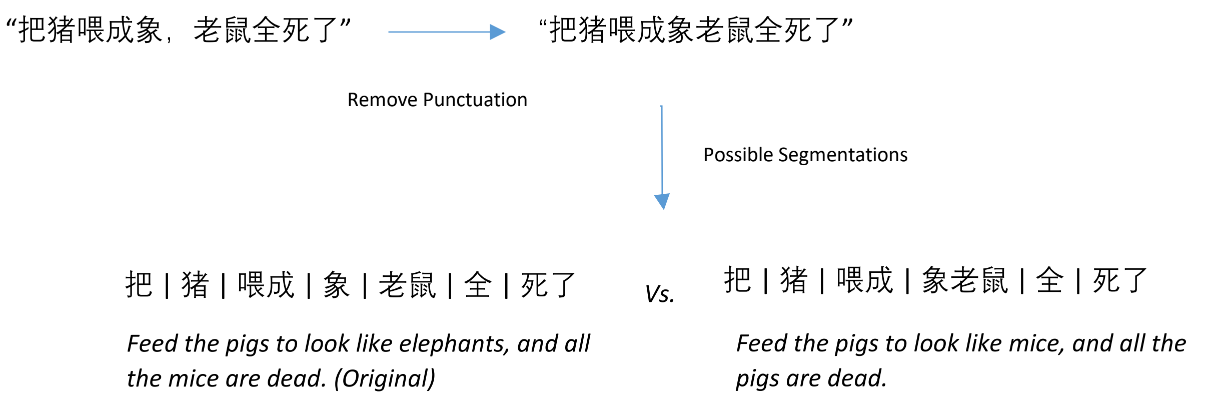
1. Corpus Extraction and Data Cleaning

We choose two different Chinese Corpus to compare their generality vs. convergence. The Sogou News General contains general news of various areas from 2000-2015; The Eastern Global Financial News, by comparison, includes financial news, reports and institutional analyses of the domestic and international market from 2000-2015 as well. The detail information is listed in Table 1.

Corpus	File Size	Tokens	Vocabulary	Link
Eastern Global Financial News:	2.8GB	506M	968K	http://app.jg.eastmoney.com/html_News
Sogou News General	3.7GB	649M	1226K	http://www.sogou.com/labs/resource/cs.php

Table 1. Detail information of the Corpora

The original corpora have small unnecessary data including links, file property, etc. This unnecessary information needs to be removed before the embedding training. Unlike English or other alphabet languages, Chinese does not have “ ” (space) in between words, thus data washing might change the original meaning of the sentence or connect unrelated words together. Therefore, a special sequence needs to be taken for the Chinese corpora to maintain the original meaning and word sequence. Examples of ambiguous words of Chinese can be seen as follows:



“把猪喂成象，老鼠全死了”

Remove Punctuation

“把猪喂成象老鼠全死了”

Possible Segmentations

把 | 猪 | 喂成 | 象 | 老鼠 | 全 | 死了

Feed the pigs to look like elephants, and all the mice are dead. (Original)

Vs.

把 | 猪 | 喂成 | 象老鼠 | 全 | 死了

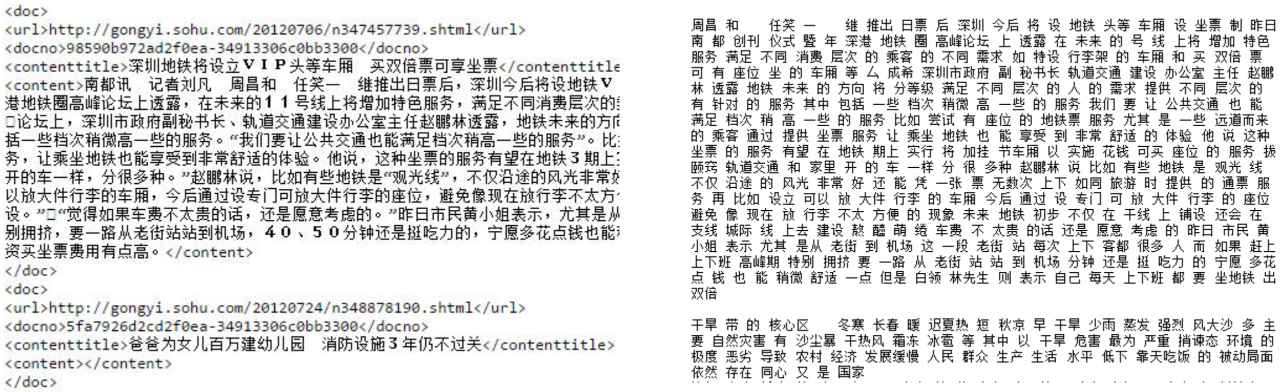
Feed the pigs to look like mice, and all the pigs are dead.

Figure 1. Ambiguous meaning from different word segmentation and context is common in Chinese

In order to minimize the influence of the manipulation on the original corpora, the sequence of the measures is significant. The optimized manipulation sequence is as follows:

1. Extract the news contents while removing HTTP links, unnecessary property information, etc.
2. Replace tabs, “/t”, “/n” with inserted spaces.
3. Replace all punctuations, single letters, and numbers with spaces.
4. Word segmentation using Jieba or HanLP.[2]
5. Remove words from the stop words list (1024 words)[3]

It is important to insert spaces to maintain the original separation between characters. Punctuations, symbols and other conditionally unnecessary characters could be removed to shrink the file size. Stop words should be deleted after the word segmentation, however, because it might cause unnecessary ambiguity for the separation algorithms. After the corpora manipulation and word segmentation, they are suitable for the word vector training with optimization of size and purity of contents. The samples of the results are illustrated below in Figure 2.



```
<doc>
<url>http://gongyi.sohu.com/20120706/n347457739.shtml</url>
<docno>98590b972ad2f0ea-34913306c0bb3300</docno>
<contenttitle>深圳地铁将设立ＶＩＰ头等车厢　买双倍票可享坐票</contenttitle
<content>南都讯　记者刘凡　周昌和　任笑一　继推出日票后，深圳今后将设地铁Ｖ
港地铁圈高峰论坛上透露，在未来的１１号线上将增加特色服务，满足不同消费层次的
□论坛上，深圳市政府副秘书长、轨道交通建设办公室主任赵鹏林透露，地铁未来的方
括一些档次稍微高一些的服务。“我们要让公共交通也能满足档次稍高一些的服务”。比
务，让乘坐地铁也能享受到非常舒适的体验。他说，这种坐票的服务有望在地铁３期上
开的车一样，分很多种。”赵鹏林说，比如有些地铁是“观光线”，不仅沿途的风光非常
以放大件行李的车厢，今后通过设专门可放大件行李的座位，避免像现在放行李不太方
设。”□“觉得如果车费不太贵的话，还是愿意考虑的。”昨日市民黄小姐表示，尤其是从
别拥挤，要一路从老街站站到机场，４０、５０分钟还是挺吃力的，宁愿多花点钱也能
资买坐票费用有点高。</content>
</doc>
<doc>
<url>http://gongyi.sohu.com/20120724/n348878190.shtml</url>
<docno>5fa7926d2cd2f0ea-34913306c0bb3300</docno>
<contenttitle>爸爸为女儿百万建幼儿园　消防设施３年仍不过关</contenttitle>
<content></content>
</doc>
```

```
周昌 和     任笑 一     继 推出 日票 后 深圳 今后 将 设 地铁 头等 车厢 设 坐票 制 昨日
南 都 创刊 仪式 暨 年 深港 地铁 圈 高峰论坛 上 透露 在 未来 的 号 线 上将 增加 特色
服务 满足 不同 消费 层次 的 乘客 的 不同 需求 如 特设 行李架 的 车厢 和 买 双倍 票
可 有 座位 坐 的 车厢 等 ム 成希 深圳市政府 副 秘书长 轨道交通 建设 办公室 主任 赵鹏
林 透露 地铁 未来 的 方向 将 分等级 满足 不同 层次 的 人 的 需求 提供 不同 层次 的
有 针对 的 服务 其中 包括 一些 档次 稍微 高 一些 的 服务 我们 要 让 公共交通 也 能
满足 档次 稍 高 一些 的 服务 比如 尝试 有 座位 的 地铁票 服务 尤其 是 一些 远道而来
的 乘客 通过 提供 坐票 服务 让 乘坐 地铁 也 能 享受 到 非常 舒适 的 体验 他 说 这种
坐票 的 服务 有望 在 地铁 期上 实行 将 加挂 节车厢 以 实施 花钱 可买 座位 的 服务 拔
颐窍 轨道交通 和 家里 开 的 车 一样 分 很 多种 赵鹏林 说 比如 有些 地铁 是 观光 线
不仅 沿途 的 风光 非常 好 还 能 凭 一张 票 无数次 上下 如同 旅游 时 提供 的 通票 服
务 再 比如 设立 可以 放 大件 行李 的 车厢 今后 通过 设 专门 可 放 大件 行李 的 座位
避免 像 现在 放 行李 不太 方便 的 现象 未来 地铁 初步 不仅 在 干线 上 铺设 还会 在
支线 城际 线 上去 建设 熬 醯 萌 绻 车费 不 太贵 的话 还是 愿意 考虑 的 昨日 市民 黄
小姐 表示 尤其 是从 老街 到 机场 这 一段 老街 站 每次 上下 客都 很多 人 而 如果 赶上
上下班 高峰期 特别 拥挤 要 一路 从 老街 站 站 到 机场 分钟 还是 挺 吃力 的 宁愿 多花
点 钱 也 能 稍微 舒适 一点 但是 白领 林先生 则 表示 自己 每天 上下班 都 要 坐地铁 出
双倍

干旱 带 的 核心区     冬寒 长春 暖 迟夏热 短 秋凉 早 干旱 少雨 蒸发 强烈 风大沙 多 主
要 自然灾害 有 沙尘暴 干热风 霜冻 冰雹 等 其中 以 干旱 危害 最为 严重 梢谏态 环境 的
极度 恶劣 导致 农村 经济 发展缓慢 人民 群众 生产 生活 水平 低下 靠天吃饭 的 被动局面
依然 存在 同心 又 是 国家
```

Figure 2. Original corpus (left) vs cleaned corpus for embedding training (right)

2. Word Embedding Training

One of the tasks of the project is to create word clustering. Take the Finance Information System as an example, from the vast amount of daily information; we would like to extract all information relating to a company as much and accurate as possible, to evaluate all the factors, which might influence the corporate performance and stock price fluctuation. Therefore, we research and experiment different word embedding methods including N-grams, CBOW, Skip-grams to test the contextual relativity each of them could obtain from the corpora. Both theoretical analyses and experiments indicate that the Skip-gram model maintains a higher level of contextual information under similar perimeter settings and training expense.

The skip-gram model matches the center word best with its contextual words, and thus through the optimization process, e.g. gradient descent, words with similar word vectors share a similar context naturally. We build separate word embedding model for each of the two corpora, and the shared parameter details are as follows:

Model	Iteration	Vector size	Window size	Min_count	Negative Sampling	Context distribution smoothing
Skip-gram	5	300	5	10	1e-3	0.75

Table 2. Detail Parameter for Embedding Training

3. Analogical Reasoning on Semantic Relations.

The word embedding transfers linguistic information to a numerical one. Thus, it is able to compare and calculate linguistic relations. Since a 300-dimension vector illustrates each of the words, their mutual relativity can be calculated by the vectors’ included angles. We test both of the embedding models from different corpora and compare their generality vs. convergence performance in representing the information of the corpora. Part of the test and semantic questions are listed below.

Semantic Relativity Analysis

Note: [EA] = Eastern Global Financial News; [SO] = Sogou News General

Search for most similar words: ('百度',topn=10)

[EA]: ('阿里巴巴', 0.6400333642959595), ('李彦宏', 0.6286805868148804), ('搜狗',0.6286042928695679), ('新浪', 0.5966048836708069), ('腾讯', 0.5804732441902161), ('搜狐', 0.5635340809822083), ('开放平台', 0.5559950470924377), ('谷歌', 0.5477848649024963), ('搜索’, 0.5392541289329529), ('网易', 0.5387555956840515)

[SO]: ('搜索引擎', 0.7155775427818298), ('百科', 0.6696987152099609), ('网易', 0.6601287126541138), ('文库', 0.656933069229126), ('腾讯', 0.648666501045227), ('浏览器', 0.6448997259140015), ('客户端', 0.6142610311508179), ('谷歌', 0.5935513973236084), ('微软', 0.5827982425689697), ('文档', 0.5815396308898926)

Evaluate the similarity ("百度", "纳斯达克")

[EA]: 0.3365986

[SO]: 0.2344546

Analogical Reasoning (['百度', '离职'], ['股票'，？], topn=5)

[EA]: ('下跌', 0.4867422580718994), ('卖出', 0.47246670722961426), ('回调', 0.47186100482940674), ('大盘', 0.4540167450904846), ('裁减', 0.4509059190750122)

[SO]: ('净值', 0.4945463538169861), ('雅虎', 0.4740077555179596), ('涨跌', 0.46732455492019653), ('跌幅', 0.4645947217941284), ('债券', 0.4555172920227051)

Different Corpora contains diverse information that can be extracted from. The corpus from Sogou General News(SO) often conveys the Breath of information, while the Eastern Global Financial News (EA), by comparison, provides information more relating to the business and cooperate performance. Based on their different features, we propose that the SO model might be more suitable for relativity analysis or search advising, while the EA model works better for information extraction and decision evaluation.

4. Data Visualization and Clustering

Multidimensional vectors can be compressed to be able to visualize in 2-D or 3-D coordinate system. Though some part of information might loss during the process, it is still a fast and straightforward way to examine the quality and features of data. We use both PCA and T-SNE methods to compress the 300-dimension vector to visualize sample of the vocabulary lists extracted from the aforementioned corpora.

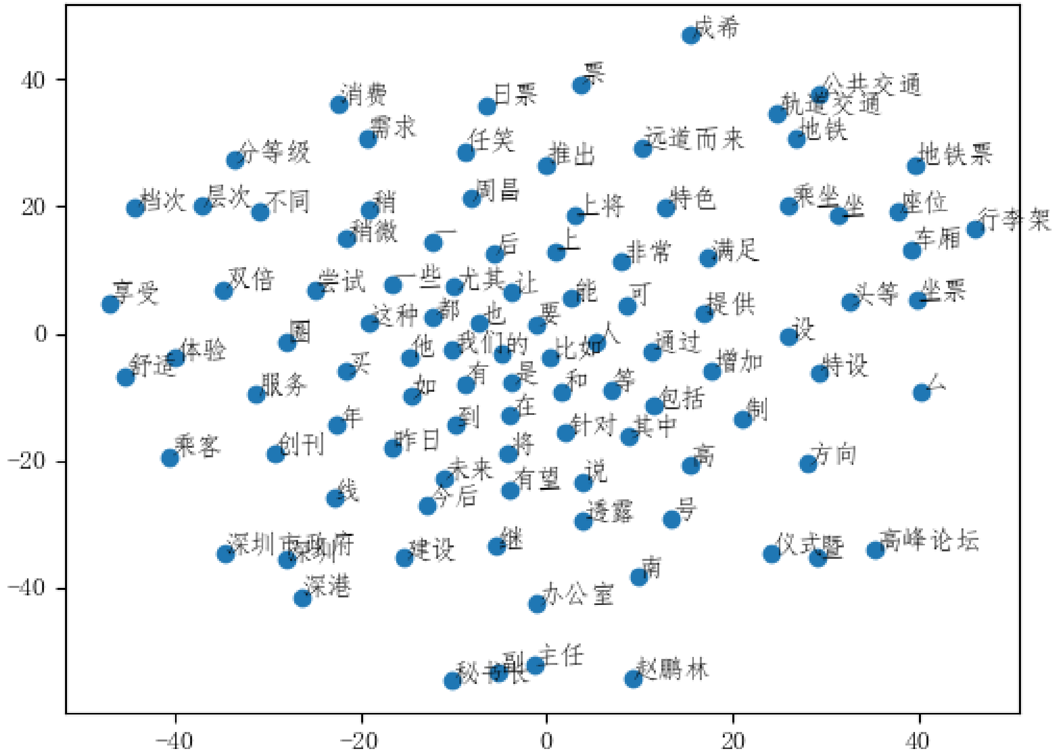


Figure3. Distribution of sample vocabulary from the SO model

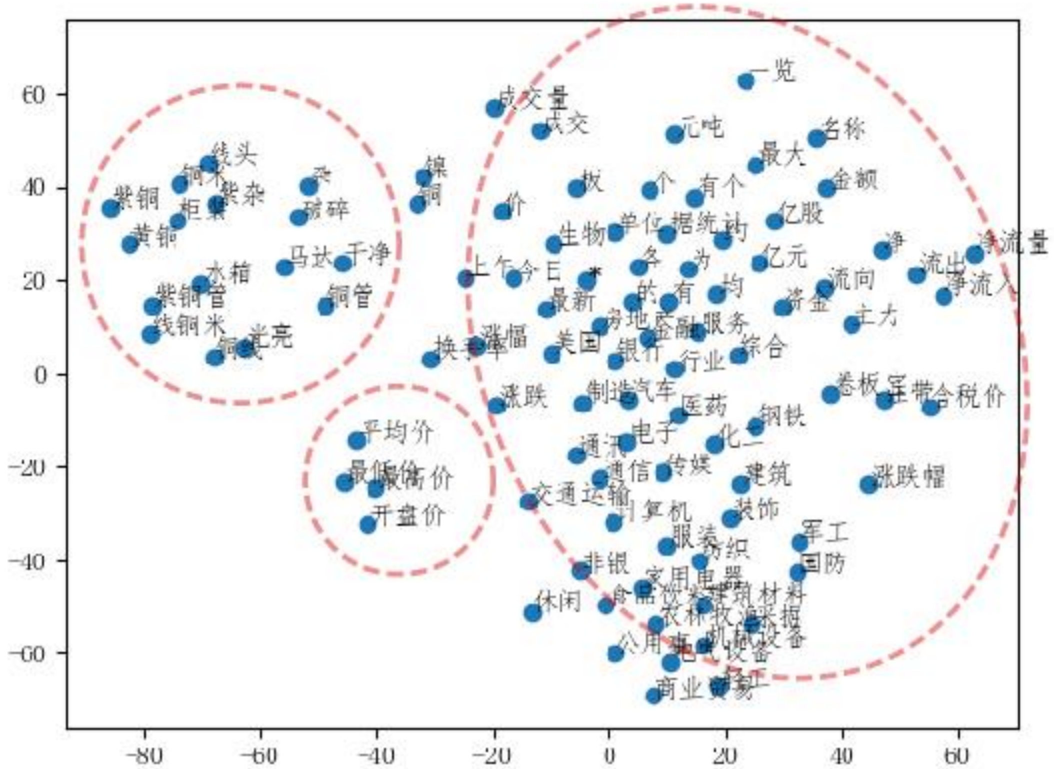
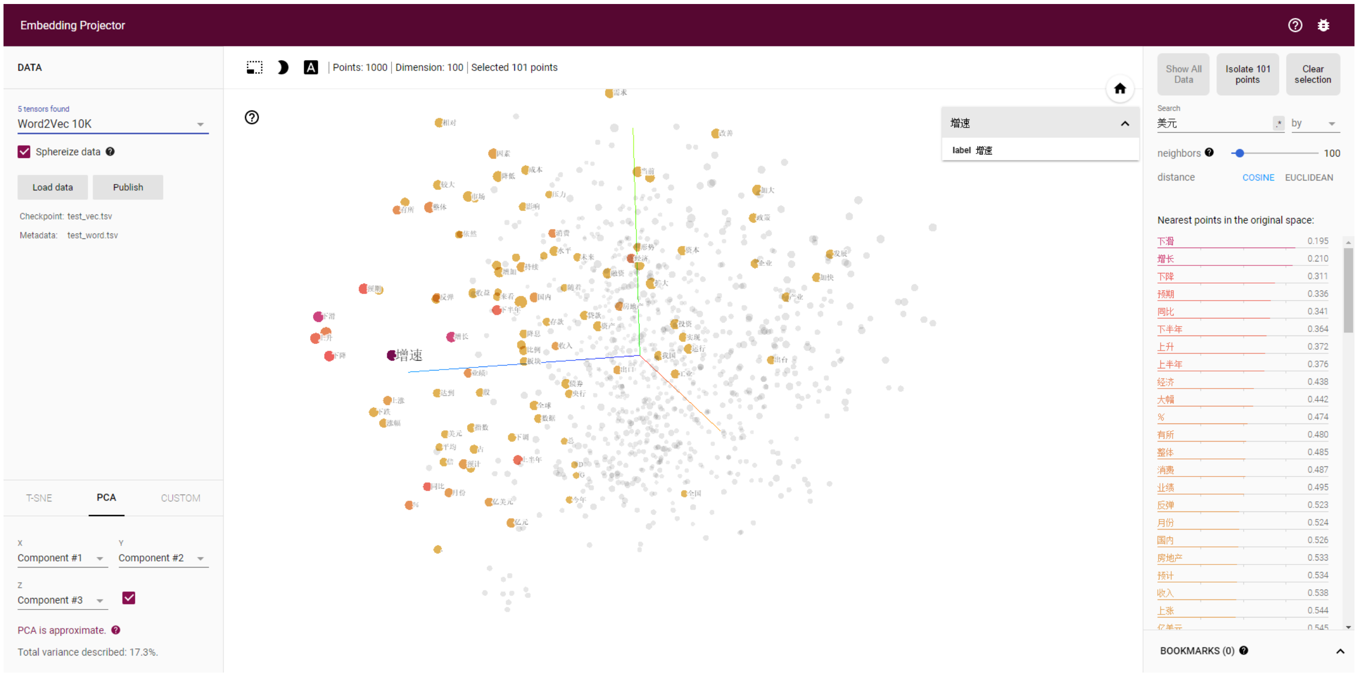


Figure4. Distribution of sample vocabulary from the EA model

The distribution from the Sogou General News is relatively even without clear categories of the vocabulary; the distribution of the EA model, by comparison, because of its contents more focusing on similar topics, illustrates more obvious clustering effects in its vocabulary. The Clustering effect in the EA model indicates a better performance in information classification and analytical reasoning.

The *TensorBoard* platform by *Tensorflow* also provides efficient access to multi-dimension vector visualization. It also utilizes PCA and T-SNE methods to compress data into 3-D coordinate system, and provide quick access to similarity search, semantic reasoning and other NLP tools. Preview of the visualization of the corpora can be seen in Figure 5.



Embedding Projector

DATA

5 tensors found
Word2Vec 10K

☑ Sphereize data

Load data | Publish

Checkpoint: test_vec.tsv
Metadata: test_word.tsv

T-SNE | PCA | CUSTOM

X: Component #1 | Y: Component #2
Z: Component #3 ☑

PCA is approximate.

Total variance described: 17.3%.

Points: 1000 | Dimension: 100 | Selected 101 points

增速
label 增速

Show All Data | Isolate 101 points | Clear selection

Search: 美元 by

neighbors: 100

distance: COSINE EUCLIDEAN

Nearest points in the original space:

下滑	0.195
增长	0.210
下降	0.311
预期	0.336
同比	0.341
下半年	0.364
上升	0.372
上半年	0.376
经济	0.438
大幅	0.442
%	0.474
有所	0.480
整体	0.485
消费	0.487
业绩	0.495
反弹	0.523
月份	0.524
国内	0.526
房地产	0.533
预计	0.534
收入	0.538
上涨	0.544
亿美元	0.545

BOOKMARKS (0)

Figure5. Distribution of sample vocabulary from the EA model by TensorBoard.

5. Future Applications for HKEX Finance Information System and Conclusion

The HKEX Finance Information System requires extracting information that might influence the performance of certain companies from daily news, reports and analyses. Based on the classified information, the system will evaluate the possible positive or negative stock price fluctuation for the near future. Basic framework of the system is as follows:

1. Categorize information based on the contents and match to the firm list.
2. Evaluate the corpora and extract key words as indication of the firm's future performance.
3. Infill these extracted words to a pre-trained neural networks.
4. Evaluate the positive or negative movement possibility based on the results of the calculation.

This project provides necessary foundation for the first and second part of the Information System. Word Embedding trained from specific corpora contains relatively accurate information for categorization. Meanwhile, it is also feasible to extract key words by relativity search or semantic analysis. Specialized corpus such as the Eastern Global Financial News in the previous experiments contains vocabulary more clustered than the general corpus and thus contributes to results that are more accurate.

Acknowledgment

This project was finished during my internship as an AI engineer in Kingdom AI, Beijing, China. I really appreciate Mr. Yi Zhang, founder of Kingdom AI, for guiding me into the world of Nature Language Processing. Also, this work could not be done without the help of Mr. Duan Qinghua and Mr. Ma Xinlei, both of them provide important technical suggestions to build up the main algorithms.

Reference

Yuehua Liu, Wenyu Pan, and Wei Gu. 2001. *Practical grammar of modern Chinese*. The Commercial Press.

Liner Yang and Maosong Sun. 2015. Improved learning of chinese word embeddings with semantic knowledge. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, Springer, pages 15–25

.Zhe Zhao, Tao Liu, Shen Li, Bofang Li, and Xiaoyong Du. 2017. Ngram2vec: Learning improved word representations from ngram co-occurrence statistics. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Pages 244–253

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606* .

Xinxiong Chen, Lei Xu, Zhiyuan Liu, Maosong Sun, and Huan-Bo Luan. 2015. Joint learning of character and word embeddings. In *IJCAI*. pages 1236–1242.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn’t. In *Proceedings of the NAACL Student Research Workshop.*